Interdisciplinary Research Project 'AI Shield'

Thijs van Ede
Semantics, Cybersecurity and Services (SCS) group
the University of Twente
Enschede, the Netherlands
t.s.vanede@utwente.nl

Trix Mulder
Legal Aspects of Entrepreneurship
the Hanze University of Applied Sciences
Groningen, the Netherlands
tr.mulder@pl.hanze.nl

Evgeni Moyakine
Transboundary Legal Studies (TLS)
the University of Groningen
Groningen, the Netherlands
e.v.moyakine@rug.nl

***Abstract*—In the modern day and age, cybersecurity faces numerous challenges. Computer systems and networks become more and more sophisticated and interconnected, and the attack surface constantly increases. In addition, cyber-attacks keep growing in complexity and scale. In order to address these challenges, security professionals started to employ generative AI (GenAI) to quickly respond to attacks. However, this introduces challenges in terms of how GenAI can be adapted to the security environment and where the legal and ethical responsibilities lie. The Universities of Twente and Groningen and the Hanze University of Applied Sciences have initiated an interdisciplinary research project to investigate the legal and technical aspects of these LLMs in the cybersecurity domain and develop an advanced AI-powered tool. This project is currently being developed and will form the basis of the grant application to be submitted in the near future at the Dutch Research Council ('Nederlandse Organisatie voor Wetenschappelijk Onderzoek' or NWO).**

Keywords—generative AI, large language models, cybersecurity, ethical and legal compliance

I. Introduction

Currently, organisations heavily rely on digital infrastructures and technologies to conduct their (business) operations. Systems used for these purposes often serve as the backbone of organisations' digital presence and are essential for their functioning. Importantly, they contain significant amounts of sensitive information that needs to be collected, stored and processed with utmost care and in accordance with a wide spectrum of legal requirements. Some examples of such systems include those dealing with enterprise resource planning, customer relationship and financial data. Various types of cyber-attacks pose significant risks to the operation of these computer systems and often lead to major legal, financial and reputational consequences for the affected organisations. [1] These attacks regularly originate from advanced persistent threats (APTs), involving cybercriminals and state actors or state-affiliated groups, whose offensive cyber operations cannot be easily attributed to specific states. [2] In this context, the rise of generative AI (GenAI) constitutes a promising opportunity for strengthening organisations' cybersecurity resilience and introducing measures addressing different cyber threats. This can be done by a GenAI instrument capable of identifying exploits and vulnerabilities of a given system and proposing ways of dealing with them in an effective and efficient manner.

The innovative research project 'AI Shield' developed and advanced by the Universities of Twente and Groningen and the Hanze University of Applied Sciences seeks to create such an instrument by combining technical, legal and operational expertise of the project partners. The deployment of this GenAI tool has an aim of analysing organisations' infrastructure, determining possible attack paths, which would allow these organisations to generate advice, response plans and prevention steps for tackling the identified threats and improving their cybersecurity posture. It should be noted at the outset that this paper does not address the final outcomes of the project and does not present the finalised GenAI system called 'AI Shield' (the name is subject to change). It rather focuses on the research project that is still in its initial stages and will be developed and completed following the acquisition of the necessary funding of the Dutch Research Council and discusses the main aspects of both the project and the GenAI-powered tool.

What is important to observe at this point is that the system to be created within the scope of the project has potentially a number of legal implications that the partners with extensive legal expertise will delve into and comprehensively examine in conjunction with the applicable legal frameworks in the fields of privacy/data protection and cybersecurity. Also, practical implications of deploying the tool will be explored on the basis of extensive testing that will be performed by the technical project partner. This will allow the legal experts involved in the project to evaluate the effectiveness of the GenAI system built in conformity with the identified legal requirements and to make conclusions regarding its potential to bolster cybersecurity of organisations, which in turn will pave the way for responsible and sustainable innovation in the sphere of digital security. The technical dimension of the project is presented below. In addition, pertinent ethical and legal issues as brought to light so far and the legal landscape surrounding them are briefly analysed and outlined below.

II. Technical aspects of AI Shield

A. GenAI for Security Operations

Cybersecurity professionals are involved in an arms race to keep ahead of adversaries and design appropriate defence mechanisms. To this end, they share vast amounts of threat knowledge, consisting of both Indicators of Compromise (IoCs) and human readable Cyber Threat Intelligence (CTI). To enable fast and effective processing of this vast, often unstructured information, security specialists started to adopt Large Language Models (LLMs) as form of GenAI to identify threat scenarios and mitigation techniques for IT infrastructures they analyse and suggest responses for mitigation. However, this leads to various problems as it is unclear from a technical perspective how to adapt these LLMs for the highly-specialised cybersecurity language or the IT environment in which they operate. Moreover, the use of such models will have a vast impact in case they hallucinate or overlook vital security information as real-world security threats can materialise. Therefore, in this project, we aim to evaluate and improve the training and application of these LLMs for cybersecurity purposes. We will evaluate the performance for the tasks of automated information extraction from CTI and applying this information to find potential attack

paths within an organisation that can be used to create improved defences.

B. Attack Path analysis

Traditionally, security teams are tasked with defending digital environments. These teams typically use vulnerability and misconfiguration scanners to assess the state of their systems and need to react to a range of problems. Unfortunately, this leads to a barrage of issues: (1) complex systems can contain myriad issues that are not exploitable by attackers, leading to alert fatigue, and (2) there is not enough security personnel to address all these issues in the first place. To address (1) and (2), the security community introduced 'attack paths', likely steps attackers may take to cause damage, to help security teams prioritize their mitigation strategies – instead of focusing on individual vulnerabilities and misconfigurations. Identifying attack paths requires in-depth knowledge about threats and specific digital environments, making it both laborious and time-consuming. Especially as both the environment and attacker strategies constantly evolve. Therefore, if we do not change the current workflow of defenders, we will see more attacks with dire consequences for the digital economy, our privacy and even our safety.

Moreover, the research conducted by us in the past has demonstrated that automation can assist security operators in not only monitoring [3] but also triaging of security events [4]. To increase the effectiveness of defenders, we focus on improving attack-path analysis by leveraging GenAI to assess the attack surface of digital environments. As opposed to current algorithmic approaches, generative AI is better able to take into account the context of the environment, likely leading to more efficient red teaming. The idea is to train an LLM or transformer model on a wide variety of attack knowledge (CVEs, CWEs, CTIs, CAPEC, blogs, news articles, MITRE ATT&CK and other material) against both real-world and synthesized attacks and attack paths. Previous research has revealed that analysing cyber threat intelligence provides valuable information about the threat landscape. [5][6] This information is, however, merely used to describe past attacks rather than to predict future attack paths for other IT infrastructures. In addition, vulnerabilities are generally identified through penetration testing. While some efforts have been made to automate this process [7], targeting the attack surface remains a major challenge. Our proposed method aims to improve this by analysing the full context of attacks leveraging likely attack paths in order to be able to better direct the search for vulnerabilities. For this specific purpose, we focus on attack paths that use steps from the widely used MITRE ATT&CK framework containing different adversary tactics and techniques derived from and built on real-world observations and findings.[8] Those are the attack paths from – among others – the categories 'credential access', 'discovery', 'exfiltration', 'initial access' and 'lateral movement'. In this respect, one can think of such adversary techniques as 'Adversary-in-the-Middle', 'Browser Information Discovery', 'Automated Exfiltration', 'Drive-by Compromise' and 'Software Deployment Tools'.

The next step after training an LLM or transformer model is to use this trained model as a foundation to assess if it (1) can be used to replace existing attack-path analysis; (2) can find new attack paths in applications where no algorithmic attack paths could be found before; (3) assess the extent to which new attack information can be added to the model and be useful in attack-path analysis; (4) can replace previously hand-calculated attack-path probabilities with inference probabilities; and (5) augment existing red teams capabilities thereby dramatically increase the effectiveness of red teams.

This novel approach should (1) find attack paths already identified by proven methods; (2) correctly detect new attack paths not found by existing work; (3) automatically adapt its analysis when presented with new vulnerability knowledge; (4) measure to what extent the foundational model itself is vulnerable to attacker inference; and (5) increase the effectiveness of red teams by shifting their responsibilities from inventing attack paths to assessing and verifying automatically generated paths.

To this end, we employ Retrieval Augmented Generation (RAG) based on the MITRE ATT&CK framework to instruct the model about individual attack techniques that an attacker may take. The advantage of a RAG is that it can be updated with novel information, thereby letting the system automatically discover additional ways for attacks to proceed. Moreover, we include a description of the IT infrastructure in which the attack took place as input to the model. Hereby, we let the GenAI model predict potential attack paths that an attacker could take to infiltrate the IT infrastructure. We can then take these suggested attacks, simulate to what extent they work and feed that back into the GenAI model such that it can learn which combination of attack techniques lead to actual breaches. The novelty of this specific approach and the actual innovation of the project lie in the combination of including additional information to the system through existing methods (i.e., RAG) and the domain-specific feedback system to determine whether an attack will succeed or not. Importantly, the LLM-based models used in the tool to be created will be able to identify new attacks, such as those that have not been discovered at the training stage. These new attack paths certainly constitute a major challenge but can and will be identified by adding new techniques to the MITRE ATT&CK framework, which can subsequently be included in the RAG. Moreover, the system can learn from varying IT infrastructures using the domain-specific feedback loop that we use for fine tuning.

III. Ethical and legal considerations

A. Ethical aspects

In order to meet the research objectives, it is imperative to tackle ethical issues which might and possibly will arise in the process of developing and testing this advanced AI-powered system. In this respect, a 'human-centric approach' respecting European values and principles must be adopted, following the EU guidelines on ethics in AI. The legal experts involved in the project have not only extensive knowledge of applicable laws and other regulations in this specific field, but one of them also serves at the research ethics committee of his faculty. They will reflect on the ways of ensuring compliance with the key requirements for developing and deploying the GenAI tool that adheres to the relevant ethical principles and values. Particularly, it is required to ensure human agency and human oversight and achieve a high level of technical robustness and safety. Moreover, transparency must be at the centre of the functioning of the system and accountability mechanisms should be designed and put in place. Also, unfair bias is to be prevented and the requirement of implementing diversity, non-discrimination and fairness needs to be met. Finally, it is indispensable to also examine possible negative

effects of the tool on society, democracy and environment and to avoid or reduce them as far as possible.

B. Legal aspects: the General Data Protection Regulation

Given that personal information is currently processed at various levels within organisations, 'AI Shield' is expected to rely on the processing of personal data which are stored within their systems that can and hopefully will integrate the GenAI system developed in the project. Once personal information is involved, its processing must comply with the requirements of Regulation (EU) 2016/679, known as the General Data Protection Regulation (GDPR). In Article 4(1) GDPR, personal data are defined as any information that relates to identified or identifiable natural persons who are called 'data subjects'. This may include regular personal data in the form of first and last names, home addresses, email addresses and identification numbers of employees and customers of organisations. In addition, sensitive personal data revealing racial or ethnic origins, political opinions, health conditions and biometric features of those data subjects can potentially be processed by the project's system. As a result, several legal questions and issues should be taken into account and explored in the development phase in order to ensure a high level of compliance with the GDPR and make the tool compatible with the standards outlined in this legal act dealing with data protection in the European Union.

In first instance, it must be determined what roles those who are involved in the functioning of the system have in practice. Under the GDPR, data controllers and data processors handling personal data have several obligations and it is crucial to establish who can be qualified as such in the processing operations taking place in the context of the functioning of 'AI Shield'. Data controllers are responsible for compliance with core principles relating to the processing of personal data mentioned in Article 5 Par. 1 GDPR. It is, among others, required that personal information is processed lawfully, fairly and transparently, is kept accurate and is up-to-date. Importantly, the second paragraph of this provision requires data controllers to be able to demonstrate compliance with the essential data processing principles. The question is, however, how to achieve these goals without negatively affecting the operation and performance of the GenAI tool. Also, attention should be devoted to data subjects' rights that the GDPR grants to individuals whose personal data are handled in the process. What are those rights in this specific context and how can they be fulfilled? What is, for example, the most adequate way of complying with the obligation to inform data subjects ex Articles 13 and 14 GDPR and what is the required degree of information provision regarding processing operations performed by 'AI Shield'? How should the developers perform a delicate balancing act between on the one hand data protection and transparency and on the other hand digital security and confidentiality aimed at protecting information from unauthorised access?

It could happen that during the process of mapping potential attack paths, it becomes apparent that a prior cyber incident has occurred, resulting in a data breach. If this is the case, data controllers will have to comply with their reporting obligations and appropriate actions will have to be taken. More specifically, when a data breach occurs, data controllers are required in accordance with Article 33 GDPR to notify national supervisory authorities, such as the Dutch data protection authority ('Autoriteit Persoonsgegevens'), if the breach in question is likely to result in a risk to the rights and freedoms of natural persons. Also, if there is a high risk to the rights and freedoms of natural freedoms, personal data breaches must be communicated to data subjects, as stipulated in Article 34 GDPR. It should be clarified how to comply with these obligations and who to communicate possible data breaches to. Additionally, it is essential to elaborate on the adoption of technical and organisational measures to promptly address the potential breach.

C. Legal aspects: other applicable regulations

In August 2024, the AI Act (Regulation (EU) 2024/1689) applicable in the entire European Union officially entered into force. The act comprehensively defines 'AI system' in Article 3(1) as 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'. Although this regulation will apply from 2 August 2026, it is crucial to thoroughly understand the requirements outlined in this legal document, given that 'AI Shield' might fall under its scope of application. Specifically, an assessment should determine whether the GenAI tool under development represents an AI system with minimal risk, high risk or unacceptable risk. Based on this determination, the corresponding requirements will have to be met.

Furthermore, some organisations intending to use the tool of the project may be subject to the NIS2 (Network and Information Systems) Directive (Directive (EU) 2022/2555), especially in sectors such as energy, transportation, and healthcare. Although the NIS1 Directive (Directive (EU) 2016/1148) was repealed by the NIS2 Directive, only a few EU countries, such as Belgium and Hungary, have managed so far to meet the transposition deadline of 17 October 2024. Other EU Member States are currently in the process of implementing the provisions of the NIS2 Directive in their national legislation and will also have to make a distinction between two types of organisations: essential (Annex I of Directive (EU) 2022/2555) and important entities (Annex II of Directive (EU) 2022/2555). Under the new legal framework, not only transport undertakings and healthcare providers but also public administration entities are among others considered as belonging to the sectors of high criticality. (Annex I of Directive (EU) 2022/2555) Due to the fact that these and other organisations may consider deploying the GenAI tool in their systems, it is of crucial significance to establish a spectrum of obligations that they have in this regard. Both important and essential entities are required to adopt appropriate and proportionate technical, operational and organisational measures to manage risks that are posed to the security of their network and information systems, as specified in Article 21 NIS2 Directive. They also have reporting obligations laid down in Article 23 that must be carefully examined to determine how the respective national authorities should be notified about possible cybersecurity incidents.

Finally, one cannot disregard the European Cyber Resilience Act (COM/2022/454 final) that is expected to enter into force late 2024 and will then be applicable as of late 2026. When the final technical solution proposed in this project becomes accessible in the market, it can be regarded as a Class I critical product with a digital component. Therefore, sufficient attention should be devoted to ensuring that the final product is designed and developed with an appropriate

cybersecurity level in mind. Among other things, it must be free from known exploitable vulnerabilities, protect the confidentiality of the processed data and ensure protection from unauthorised access.

IV. Conclusion

The research project 'AI Shield' that is currently being designed by three researchers from the Netherlands working for different higher education institutions and prepared for the grant application at the Dutch Research Council seeks to develop an advanced system utilising generative AI. It must be observed here that it has not resulted in the development of the actual 'AI Shield' instrument meant for examining IT infrastructures of various organisations and aimed at significantly improving identification and analysis of attack paths in those digital environments so far. The primary objective of the project is to enable organisations to not only detect but also respond to cyber threats as efficiently and as effectively as possible.

By attending the CSNet 2024 conference, the involved researchers hope to be able to share their research insights and preliminary findings with the experts attending the event and receive their feedback that undoubtedly will prove highly valuable. Although the technical aspects of the project constitute its foundation and are central to its essence, compliance with both ethical guidelines and legal requirements will also receive focused attention. In order to enable broad deployment of the sophisticated GenAI tool to be designed and to achieve the ultimate goal of enhancing the security posture of European organisations using modern information systems, the project will adhere to strict ethical and legal standards and will comply with all relevant ethical guidelines, laws and regulations, as discussed above.

References

[1] E. Moyakine, "Bits and locks from a cybersecurity perspective: Understanding and preventing ransomware attacks with LockBit as an example of top-tier ransomware," Tijdschrift voor Internetrecht, vol. 17(1), pp. 17-21, 2024. Available: https://pure.rug.nl/ws/portalfiles/portal/962308649/Bits_en_sloten_vanuit_een_cybersecurityperspectief.pdf (original title (in Dutch): "Bits en sloten vanuit een cybersecurityperspectief: Begrijpen en voorkomen van ransomware-aanvallen met LockBit als voorbeeld van gijzelsoftware van de hoogste orde")

[2] E. Moyakine, "Pulling the strings in cyberspace: Legal attribution of cyber operations based on state control," in F. Delerue, A. Sukumar and D. Broeders, Responsible behaviour in cyberspace: Global narratives and practice, Luxembourg: Publications Office of the European Union, 2023, pp. 200-218. Available: https://pure.rug.nl/ws/portalfiles/portal/762940205/Responsible_behaviour_in_cyberspace_-_closing-the-gap.pdf

[3] T. van Ede et al., "FlowPrint: Semi-supervised mobile-app fingerprinting on encrypted network traffic," Network and distributed system security symposium (NDSS), vol. 27, 2020. Available: https://www.ndss-symposium.org/wp-content/uploads/2020/02/24412.pdf

[4] T. van Ede et al., "Deepcase: Semi-supervised contextual analysis of security events," 2022 IEEE Symposium on Security and Privacy (SP), IEEE, 2022. Available: https://ieeexplore.ieee.org/document/9833671

[5] P. Gao et al., "ThreatKG: A threat knowledge graph for automated open-source cyber threat intelligence gathering and management," 2022, arXiv:2212.10388. Available: https://arxiv.org/abs/2212.10388

[6] K. Satvat, R. Gjomemo and V. N. Venkatakrishnan, "Extractor: Extracting attack behavior from threat reports," 2021 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2021. Available: https://ieeexplore.ieee.org/document/9581182

[7] Hu, R. Beuran and Y. Tan, "Automated penetration testing using deep reinforcement learning," 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2020. Available: https://ieeexplore.ieee.org/document/922975

[8] MITRE, *ATT&CK.* Available: https://attack.mitre.org