Rethinking Cyber Safety and Cybersecurity in the Age of AI: A Position Paper

Christina Kolb
IEBIS
The University of Twente
The Netherlands
c.kolb@utwente.nl

Megha Quamara
Department of Informatics
King's College London
London, UK
megha.quamara@kcl.ac.uk

Daniel Braun
Department of Mathematics and Computer Science
Marburg University
Marburg, Germany
daniel.braun@uni-marburg.de

***Abstract*—Distinguishing between cyber safety and cybersecurity is crucial, as understanding and addressing both are essential for developing comprehensive strategies to protect systems, data, users, and their environments. In this article, we argue that the terms cyber safety and cybersecurity lack a clear distinction in research and practice. We also explore how these concepts interact and potentially influence each other, particularly in the context of generative AI.**

***Index Terms*—cyber safety, cybersecurity, interplay, generative AI, large language models**

I. Introduction

Safety (protection from unintended, accidental harm to a valuable asset) and security (protection from intentional, malicious harm to a valuable asset) address different concerns and are often examined separately during the development of software-based systems [1]. Although their causes, consequences, and mitigation strategies typically differ, safety and security often influence each other in practice. Sometimes they *conflict*; for example, in autonomous vehicle communication, encryption to protect message confidentiality may introduce delays, potentially interfering with timeouts designed to guarantee timely delivery of safety-critical updates. In other cases, they can mutually *reinforce*, such as when secure sensor data validation helps prevent both malfunctions and attacks.

Recognizing such interactions between safety and security is important for identifying vulnerabilities and developing effective countermeasures (e.g., access controls, integrity checks, versioning, and backups) in an integrated fashion. It is even more important to distinguish between these domains, their underlying principles, and interplay from a cyber perspective, particularly in systems connected to the Internet, where both malicious threats and unintended failures can propagate rapidly and have real-world consequences. This is also relevant in modern Artificial Intelligence (AI)-driven systems [2]. Consider a smart, connected car using a Large Language Model (LLM)-based assistant for navigation and decision support; if the system lacks proper security, an attacker could manipulate it to suggest dangerous routes to the driver or disable safety warnings. Even without malicious interference, a poorly aligned or inadequately tested LLM might offer unsafe advice, such as recommending illegal U-turns or ignoring school zones, potentially leading to accidents.

Cybersecurity has long been a priority for governments worldwide. For example, the European Union established the European Union Agency for Cybersecurity (ENISA) in 2004, which publishes an annual Threat Landscape Report. A recent report focuses on cybersecurity threats, but also highlights concerns that current AI tools and their decision-making pose, for example, those related to safety [3]. In this paper, we advocate for a clear distinction between cyber safety and cybersecurity to better identify concerns specific to AI-driven systems. As we discuss in Sec. II, the literature often conflates these concepts, hindering the development of robust, integrated safeguards essential for trustworthy AI. The growing complexity of AI systems, particularly LLMs, further reveals how closely cyber safety and cybersecurity are intertwined. Evolving challenges, such as unintended human-AI interactions, misinformation, and misuse, span both domains [4] and underscore the need for clear conceptual boundaries (Sec. III). We conclude by posing key questions to the research and development community to guide the development of solutions that integrate cyber safety and cybersecurity in these systems (Sec. IV).

II. Cyber Safety and Cybersecurity

With converging technologies and capabilities, such as Internet connectivity, automation, cloud infrastructure, and robotics, considerations of cyber safety and cybersecurity are shaping advancements across many research fields. For instance, in software-based systems engineering, existing research has explored keyword-based modeling approaches to cybersecurity [5]. However, these keyword sets often contain terms related to both cyber safety and cybersecurity (e.g., unintentional, intentional, security safeguard, risk, threat) without a clear distinction between the two. Besides clarifying the terms cyber safety and cybersecurity, understanding their interactions is crucial for implementing effective countermeasures. In robotics, cyber safety has been defined as the prevention of vulnerabilities and scenarios (e.g., unintended behavior) in which robotic systems (e.g., vacuum cleaners, drones) may harm humans [6]. For example, in the case of last-mile robots [7], a cyberattack could lead to a safety-related situation where a robot collides with a pedestrian.

In mathematical modeling, safety and security interactions have been examined through various formalisms [8]. Commonly referenced case studies include an industrial pipeline system and an electrical locked-door. However, these examples primarily address traditional safety and security concerns, without explicitly considering cyber safety or cybersecurity. While the importance of defining these concepts is acknowledged, precise specifications are often left for future work.

An example of the interaction between cyber safety and cybersecurity is the antagonistic relationship observed in autonomous driving vehicles [9]. This antagonism is modeled using a Petri net, which can be flexibly embedded at various points within the system model. For instance, a sensor can be switched on to enhance safety in distance recognition or switched off to prevent a potential cyberattack on the sensor. Turning the sensor off reduces safety and is, therefore, antagonistic to security. Another study [10] investigates the mutual reinforcement between safety and security in the context of autonomous driving. An autonomous vehicle, thus, must be cybersecure to ensure the driver's safety. However, to the best of our knowledge, it remains an open question how mathematical modeling can effectively analyze the interaction between cyber safety and cybersecurity in such systems or the others, specifically, identifying the exact points during a cyber attack when human intervention is required to maintain safety.

III. Key Considerations for Generative AI

The wide deployment and use of Generative AI, such as LLMs, has sparked a debate about the associated risks. Under the term "AI Safety", the focus is often on existential risks to humanity, overlooking more concrete and immediate threats current models pose. When concrete problems are discussed, the emphasis tends to be on security ones like robustness against adversarial attacks, despite being labelled as "AI Safety" [11]. While previous generations of generative AI were limited to the information they acquired during the training process, modern models connected to the Internet have access to more recent information, introducing new safety and security risks. Attackers can exploit this connectivity to instruct the LLM to perform actions without the user's knowledge, a technique known as *prompt injection*. Through prompt injection, a model can, for example, extract private information from the user and leak it to an attacker [12]. Since these risks arise from the models' Internet connectivity, they are inherently linked to cyber safety and cybersecurity. Integrating LLMs into other systems, including Cyber-Physical Systems (CPSs) such as those in the automotive industry, creates new threats and amplifies the potential severity of harm. Even with strict software-level separation, cyberattacks via LLMs can pose physical risks to drivers. For example, attackers could exploit user-generated Points of Interest (POIs) reviews on trusted platforms accessed by in-car entertainment systems to inject prompts that instruct the LLM to direct the driver to an unsafe location or leak location-related data.

Beyond security, enabling LLMs with Internet access can also introduce safety risks. These models might access and reproduce dangerous information without sufficient reflection, such as details about illegal street racing. Alternatively, an LLM might provide incorrect information; for example, by being unfamiliar with the rules of a specific jurisdiction or by making errors, such as miscalculating a vehicle's available range, which could potentially lead drivers into dangerous situations. Mitigating these issues requires analyzing their root causes, which depends on clearly distinguishing between cyber safety and cybersecurity concerns.

IV. Outlook

In the context of distinguishing and relating cyber safety and cybersecurity, especially in light of emerging technologies like LLMs, the following research questions arise:

1) How to clearly define cyber safety and cybersecurity?
2) In what ways do cyber safety and cybersecurity interact, and how do these interactions have real-world impacts?
3) Why might traditional approaches for classical systems fall short for autonomous CPSs or generative AI, and what strategies can improve cyber safety and cybersecurity, particularly given potential conflicts (e.g., between user autonomy and user/environmental safety)?

References

[1] G. Pedroza, "Towards safety and security co-engineering: challenging aspects for a consistent intertwining," in *International Workshop on Cyber Security for Intelligent Transportation Systems*. Springer, 2018, pp. 3–16.

[2] R. May, J. Krüger, and T. Leich, "Sok: How artificial-intelligence incidents can jeopardize safety and security," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, 2024, pp. 1–12.

[3] European Union Agency for Cybersecurity (ENISA), "Enisa threat landscape 2024," https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024, 2024, accessed: May 2025.

[4] D. Hendrycks, M. Mazeika, and T. Woodside, "An overview of catastrophic ai risks," *arXiv preprint arXiv:2306.12001*, 2023.

[5] C. Brookson, S. Cadzow, R. Eckmaier, J. Eschweiler, B. Gerber, A. Guarino, K. Rannenberg, J. Shamah, and S. Górniak, "Definition of cybersecurity-gaps and overlaps in standardisation," *Heraklion, ENISA*, 2015.

[6] S. Morante, J. G. Victores, and C. Balaguer, "Cryptobotics: Why robots need cyber safety," *Frontiers in Robotics and AI*, vol. 2, p. 23, 2015.

[7] C. Kolb and L. Xie, "Security and safety in urban environments: Evaluating threats and risks of autonomous last-mile delivery robots," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2024, pp. 34–46.

[8] S. M. Nicoletti, M. Peppelman, C. Kolb, and M. Stoelinga, "Model-based joint analysis of safety and security: Survey and identification of gaps," *Computer science review*, vol. 50, p. 100597, 2023.

[9] M. Quamara, C. Kolb, and B. Hamid, "Analyzing origins of safety and security interactions using feared events trees and multi-level model," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2023, pp. 176–187.

[10] M. Quamara, C. Kolb, and A. Lohachab, "Where do safety and security mutually reinforce? a multi-level model-based approach for a consistent interplay," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2024, pp. 316–328.

[11] R. Ren, S. Basart, A. Khoja, A. Gatti, L. Phan, X. Yin, M. Mazeika, A. Pan, G. Mukobi, R. Kim *et al.*, "Safetywashing: Do ai safety benchmarks actually measure safety progress?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 68 559–68 594, 2024.

[12] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023, pp. 79–90.