Report from Dagstuhl Seminar 23431

# Network Attack Detection and Defense – AI-Powered Threats and Responses

**Sven Dietrich*[1], Frank Kargl*[2], Hartmut König*[3], Pavel Laskov*[4], and Artur Hermann†[5]**

1 **City University of New York, US.** spock@ieee.org
2 **Universität Ulm, DE.** frank.kargl@uni-ulm.de
3 **ZITiS München, DE.** hartmut.koenig@b-tu.de
4 **Universität Liechtenstein, LI.** pavel.laskov@uni.li
5 **Universität Ulm, DE.** artur.hermann@uni-ulm.de

## Abstract

This report documents the program and the findings of Dagstuhl Seminar 23431 "Network Attack Detection and Defense – AI-Powered Threats and Responses". With the emergence of artificial intelligence (AI), attack detection and defense are taking on a new level of quality. Artificial intelligence will promote further automation of attacks. There are already examples of this, such as the Deep Locker malware. It is expected that we will soon face a situation in which malware and attacks will become more and more automated, intelligent, and AI-powered. Consequently, today's threat response systems will become more and more inadequate, especially when they rely on manual intervention of security experts and analysts. The main objective of the seminar was to assess the state of the art and potentials that AI advances create for both attackers and defenders. The seminar continued the series of Dagstuhl events "Network Attack Detection and Defense" held in 2008, 2012, 2014, and 2016. The objectives of the seminar were threefold, namely (1) to investigate various scenarios of AI-based malware and attacks, (2) to debate trust in AI and modeling of threats against AI, and (3) to propose methods and strategies for AI-powered network defenses. At the seminar, which brought together participants from academia and industry, we stated that recent advances in artificial intelligence have opened up new possibilities for each of these directions. In general, more and more researchers in networking and security look at AI-based methods which made this a timely event to assess and categorize the state of the art as well as work towards a roadmap for future research. The outcome of the discussions and the proposed research directions are presented in this report.

**Seminar** October 22–27, 2023 – https://www.dagstuhl.de/23431

**2012 ACM Subject Classification** Networks → Network security; Security and privacy → Intrusion/anomaly detection and malware mitigation; Security and privacy → Network security; Security and privacy → Systems security

**Keywords and phrases** artificial intelligence, cybersecurity, intrusion detection, machine learning

**Digital Object Identifier** 10.4230/DagRep.13.10.90

* Editor / Organizer

† Editorial Assistant / Collector

Network Attack Detection and Defense – AI-Powered Threats and Responses, *Dagstuhl Reports*, Vol. 13, Issue 10, pp. 90–129
Editors: Sven Dietrich, Artur Hermann, Frank Kargl, Hartmut König, and Pavel Laskov
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Seminar Motivation and Summary

*Sven Dietrich*
*Artur Hermann*
*Frank Kargl*
*Hartmut König*
*Pavel Laskov*

Computer networks and the services they provide have become indispensable tools these days. Consequently, they are also a popular target for attacks that are constantly increasing in complexity and sophistication. Although there are a variety of effective systems to counter such attacks, like firewalls or intrusion detection systems (IDSs), the immense diversity and number of threats make it difficult for system administrators to keep pace with the alerts triggered and respond within adequate time limits.

This problem will intensify in the future. There are signs that attacks will become more and more automated, as, for instance, indicated by the 2016 DARPA Cyber Grand Challenge in which automation of attacks was a main focus and the basic feasibility was demonstrated. Another indication of a higher degree of automation is advanced malware where Large-Language-Models (LLMs) start to get applied to craft highly sophisticated phishing emails. Experts already foresee that more and more AI mechanisms will find their way into such malware. This leads to the conclusion that we will soon face a situation in which malware and attacks will become more and more automated, intelligent, and AI-powered.

As a consequence, today's threat response systems will become more and more inadequate, especially where they rely on manual intervention of security experts and analysts. Hence, the deployment of automation and AI is the only way to attain and retain a strategic advantage in the arm's race between the attack and the defense. Usage of AI mechanisms is already the case in some security mechanisms like anomaly-detecting IDSs or virus scanners. But one could easily imagine substantially higher degrees of AI-based automation in system defense. However, automated defense may also be a double edged sword as it could be misused by attackers to trigger counterproductive responses.

In this Dagstuhl Seminar, we together with all the participants therefore tried to assess the state of the art and potentials that AI advances create for both attackers and defenders because we believe it is crucial to consider both sides when discussing the relation between AI and security.

In particular, the seminar pursued the following objectives:

1. Investigate various attack scenarios and attacker models of AI-based malware and attacks,
2. Map the space of AI-based security countermeasures going beyond the usual anomaly-based intrusion detection systems,
3. Discuss where else AI-based methods are or could be employed, and
4. Discuss how to estimate and predict the impact of countermeasures and possible side effects.

To provide initial material for such discussions, we had three keynotes by distinguished speakers. Pavel Laskov proposed "Three Faces of AI in Cybersecurity," providing a thorough account of how AI could be used in defense, for offensive purpose, and how AI itself can be an attack target. Konrad Rieck took a deep dive into the first aspect in his keynote "Bumpy Road of AI-based Attack Detection." Finally, Robin Sommer completed the picture by looking “Beyond Detection: Revisiting AI For Effective Network Security Monitoring.” Those presentations were complemented by a number of short lightning talks given by our participants to introduce the audience to various current research.

A significant share of the seminar’s time was spent in working groups, with participants discussing individual aspects of interest. The topics for those working groups were partly solicited before the seminar and then finally determined on the first day. Specifically, the topics were:

1. Assessment of AI-Based Attacks in Cybersecurity,
2. Security of Large Language Models,
3. Trust in AI and Modeling of Threats against AI in Network Defense, and
4. AI-Powered Network Defenses

The working groups report on their individual results below. In order to bring all these findings together and distill outcomes and an outlook into what could be next steps, we used the format of a World Café where in the afternoon of day 4, people split into small groups to provide their input on five pre-defined questions. As groups were shuffled randomly after every 20 minutes, everyone joined each World Café table and discussed each of the questions. The outcomes then formed the basis for our wrap-up session on Friday morning.

The seminar was originally proposed and prepared together with Marc C. Dacier from KAUST who couldn’t attend the seminar at the last minute. We owe him many ideas and contributions during the preparation phase. Pavel Laskov was so kind as to fill the empty slot on short notice.

## 2 Table of Contents

**Seminar Motivation and Summary**

**Overview of Keynotes**

**Overview of Lightning Talks**

**Working groups**

**World Café and Outlook**

# 3 Overview of Keynotes

## 3.1 Three Faces of AI in Cybersecurity

*Pavel Laskov (Universität Liechtenstein, LI)*

Artificial Intelligence (AI) has a stronger tradition in cybersecurity than one might think. Despite their seemingly contrasting scientific and methodical traits – AI is all about probabilistic events and assertions, whereas a (practical) security mindset is mostly concerned with apparent and deterministic evidence of systems being broken into – the task of detecting systems being broken into is inherently connected with observing and making sense of huge volumes of diverse bits and pieces of digital evidence commonly understood as "data." This is something that AI, albeit not originally born as an empirical science, has manifested itself as an omnipotent instrument for.

The relation between AI and cybersecurity is not just profound, but multi-faceted as well. Each of its "faces" has a different history and a different level of maturity. The oldest and most obvious role of AI in cybersecurity is to build models for detection of various kinds of attacks. The sole notion of "intrusion detection" has been implicitly defined by Denning as an AI problem, even though neither did she use this term in her seminal work [1], nor did this term have the same meaning at that time as we understand it today. The tremendous capability of AI to facilitate and speed up detection of various kinds of threats is now widely acknowledged, both in the academia and the industry. Despite a large number of still unresolved problems, e.g., development of benchmark datasets, concept drift, explainability, and many others, AI is perhaps the only reason why modern defenses are still able to keep up with the staggering increase in professionalization of the "attack industry." Furthermore, as recently pointed out by Apruzzese et al. [2], substantial merit can be gained by deployment of AI for a number of other cybersecurity tasks beyond threat detection, e.g., alert management, risk assessment and cyber threat intelligence.

As any other technology deployed for security, AI must be scrutinized for its insecurity. This axiom of security research has triggered research in security of AI, pioneered by at al. [3] and Biggio et al. [4]. This line of research can be seen as the second "face" of AI and security. Its importance clearly transcends the field of information security. While early attacks against AI systems were somewhat related to the conventional triad of security objectives – confidentiality, integrity, and availability, – the recent work has led to a discovery of various "AI endemic" attacks such as model stealing, model backdoors, attribute inference, as well as attacks against explainability. Despite the fact that several thousand papers have been hitherto published on security of AI, many important questions are still wide open, especially regarding to defenses as well as potential economic motivation behind the attacks.

A third dimension along which the relationship between AI and security is rapidly developing is "offensive AI," i.e., abuse of AI for nefarious goals. Examples of such abuse have been reported in practice as well as in the scientific literature for several years. The recent review of AI-powered threats in the organizational context has demonstrated that virtually all stages of the conventional attack "kill chain" can be facilitated by AI [5]. While it is still largely unclear to what extent AI is currently used to assist conventional security exploitation, it becomes increasingly apparent that many AI tools can be used in a dual way. Ethical discussions of these issues are likely to emerge.

**References**

1 An intrusion-detection model. *IEEE Transactions on software engineering*, (2):222–232, 1987.

2 Giovanni Apruzzese, Pavel Laskov, Edgardo Montes de Oca, Wissam Mallouli, Luis Brdalo Rapa, Athanasios Vasileios Grammatopoulos, and Fabio Di Franco. The role of machine learning in cybersecurity. *Digital Threats: Research and Practice*, 4(1):1–38, 2023.

3 Marco Barreno, Blaine Nelson, Anthony D Joseph, and J Doug Tygar. The security of machine learning. *Machine Learning*, 81:121–148, 2010.

4 Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, pages 1807–1814, 2012.

5 Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, et al. The threat of offensive ai to organizations. *Computers & Security*, 124:103006, 2023.

## 3.2 The Bumpy Road of AI-based Attack Detection

*Konrad Rieck (TU Berlin, DE)*

This talk examines the development of AI-based attack detection, both in its historical context and its future prospects. It provides an overview of the evolution of intrusion detection and pinpoints promising opportunities for recent AI techniques. The talk opens with a focus on classical learning-based detection approaches, which have developed over time but also repeatedly failed in practice due different pitfalls in their design and evaluations.

To remedy this situation, the talk then highlights the role of explainable AI (XAI), which makes the decision process of learning models transparent. While XAI represents a significant advance in building trust, it also encounters a number of challenges in the context of security, such as inconsistency and infidelity. These issues are discussed in detail and provide insights into how XAI can be employed without neglecting its limitations. In addition, the talk introduces toy examples to show how generative AI can be used to create detection signatures for attacks. These examples are used to demonstrate both the strengths of generative AI and its notable weaknesses such as hallucinations and lack of reasoning ability.

Overall, the talk aims to provide a balanced perspective on the current state and future direction of AI-assisted attack detection. By critically analyzing the hurdles on the road to success and exploring potential solutions, the talk hopefully points to possible paths for future research.

### 3.3 Beyond Detection: Revisiting AI For Effective Network Security Monitoring

*Robin Sommer (Corelight – Planegg, DE)*

While AI has been proposed for finding novel network attacks many times, such approaches have not found much traction in operational deployments. In this talk we first revisit some challenges with classic intrusion detection. We then look at "threat hunting," a modern twist on finding malicious activity that focuses on the human analyst driving the process, and we examine the potential of AI to support threat hunting workflows.

## 4 Overview of Lightning Talks

### 4.1 Robust, Explainable, and Privacy-Respecting Sybil Attack Defense

*Christian Bungartz (Universität Bonn, DE)*

**Joint work of** Christian Bungartz, Felix Boes, Michael Meier

Sybil attacks target decentralized networks and exploit trust relationships between peers. A network type of special interest are online social networks (OSN). Existing defense mechanisms often hinge on domain-specific metadata, potentially compromising user privacy and limiting applicability. A solution is a detection approach utilizing the global topological structure of the underlying network. However, the main assumption of a fast-mixing subgraph of honest peers often is not aligned with the reality of OSN structures. To tackle these issues, this work outlines five open problems and proposes an approach leveraging local, structural information. This privacy-friendly method demonstrates promising results in Sybil detection, offering a critical step towards safeguarding the trust and integrity of OSNs.

### 4.2 The SuperviZ project – towards enhanced Security Orchestration, Automation and Response

*Hervé Debar (Télécom SudParis, FR)*

**Joint work of** Hervé Debar, Ludovic Mé

The SuperviZ project is part of the "system security" axis of the PEPR cybersecurity program. It addresses the field of "system, software and network security." More precisely, it targets the detection, response and remediation to computer attacks, subjects grouped under the name of "security supervision."

The digitization of all infrastructures makes it almost impossible today to secure all systems a priori, as it is too complex and too expensive. Supervision seeks to reinforce preventive security mechanisms and to compensate for their inadequacies.

Supervision is fundamental in the general context of enterprise systems and networks, and is just as important for the security of cyber-physical systems. Indeed, with “objects” that should eventually be all, or almost all, connected, the attack surface increases significantly. This makes security even more difficult to implement. The increase in the number of components to be monitored, as well as the growing heterogeneity of the capacities of these objects in terms of communication, storage and calculation, makes security supervision more complex.

In this context, we address challenges related to (1) the increase in the number and diversity of objects to be supervised (which requires the development and adaptation of new detection mechanisms for heterogeneous environments, with false positive and negative rates that have not been achieved to date), (2) the complexity of systems interconnected to form large critical infrastructures on a European scale (which requires new detection and supervision models that take into account the criticality and cyber-physical nature of these systems), (3) the existence of increasingly complex and silent targeted attacks (which requires an observation of the global threat landscape, a capability model of the attackers and a significant improvement of the detection and reaction time), and (4) the treatment of massive attacks which rapidly affect a significant number of victims (in order to limit the damage suffered by these victims).

Faced with these challenges, it is necessary to significantly improve the efficiency of the detection-reaction chain (response and remediation). The main objective of the project is therefore to provide new solutions and to advance the current scientific state of the art.

These contributions will come from almost all the national research forces in the field, which will be strengthened by this project and will see their links tightened, which is also an objective. Moreover, in coherence with the objectives of the PEPR, we also aim to prepare the transfer of our results to the national industrial community. To this end, the scientific work will lead to prototypes and demonstrators that will be deployed on platforms built within the project. These platforms will be accessible to industrial partners.

## 4.3 A Strategy to Evaluate Test Time Evasion Attack Feasibility

*Stephan Kleber (Universität Ulm, DE)*

**Joint work of** Stephan Kleber, Patrick Wachter
**Main reference** Stephan Kleber, Patrick Wachter: “A Strategy to Evaluate Test Time Evasion Attack Feasibility”, Datenschutz und Datensicherheit, Vol. 47(8), pp. 478–482, 2023.
**URL** http://dx.doi.org/10.1007/S11623-023-1802-0

New attacks against Computer Vision systems and other perceptive machine learning approaches are currently published in high frequency. Often the assumptions or limitations of these works are so strict that the attacks seem to have no practical relevance. On the other hand, recent reports show the effectiveness of attacks against cyber-physical systems (CPS). In particular, attacks on automotive systems demonstrate the safety-impact in real-world scenarios. We discuss the practical relevance of security threats for machine learning approaches in automotive use cases and we propose a strategy to evaluate the feasibility of such threats. This includes a method to potentially discover existing vulnerabilities and rate their exploitability in the use case.

## 4.4 Privacy-preserving Artificial Intelligence for Telecommunications

*Nicolas Kourtellis (Telefónica Research – Barcelona, ES)*

**Main reference** Diego Perino, Kleomenis Katevas, Andra Lutu, Eduard Marin, Nicolas Kourtellis: "Privacy-preserving AI for future networks", Commun. ACM, Vol. 65(4), pp. 52–53, 2022.
**URL** http://dx.doi.org/10.1145/3512343
**Main reference** Kleomenis Katevas, Diego Perino, Nicolas Kourtellis: "FLaaS – enabling practical federated learning on mobile environments", in Proc. of the 20th Annual International Conference on Mobile Systems, Applications and Services, MobiSys '22, p. 605–606, Association for Computing Machinery, 2022.
**URL** http://dx.doi.org/10.1145/3498361.3539693
**Main reference** Varun Chandrasekaran, Suman Banerjee, Diego Perino, Nicolas Kourtellis: "Hierarchical Federated Learning with Privacy", CoRR, abs/2206.05209 2022.
**URL** https://doi.org/10.48550/arXiv.2206.05209
**Main reference** Fan Mo, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, Nicolas Kourtellis: "PPFL: privacy-preserving federated learning with trusted execution environments", in Proc. of the 19th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '21, p. 94–108, Association for Computing Machinery, 2021.
**URL** http://dx.doi.org/10.1145/3458864.3466628
**Main reference** Nicolas Kourtellis, Kleomenis Katevas, Diego Perino: "FLaaS: Federated Learning as a Service", in Proc. of the 1st Workshop on Distributed Machine Learning, DistributedML'20, p. 7–13, Association for Computing Machinery, 2020.
**URL** http://dx.doi.org/10.1145/3426745.3431337

Telco networks and systems are highly complex, distributed ecosystems composed of very diverse sub-environments. Traditional solutions for network management (e.g., for provisioning, data management, etc.) are reaching their limits within such complex ecosystems, and with the arrival of faster, more demanding 5/6G networks. We require novel solutions that provide 1) effective resource management, while guaranteeing 2) strict service requirement completion and 3) absolute preservation of user privacy. Towards that goal, within Telefonica, we investigate building AI models using novel, state-of-art, distributed ML paradigms such as Federated Learning (FL), to unlock the potential of Big Data produced and processed at the source: the user devices. Using FL, and coupled with more advanced hardware and software techniques (e.g, Trusted Execution Environments, Differential Privacy, etc.), we can mine the user data locally, without risking their exposure to AI model attackers. Furthermore, by tapping on the power of Edge Computing, we can potentially scale AI model computation to millions of devices. To this end, we are prototyping a novel, FL-as-a-Service (FLaaS) platform, that will enable third-party companies to build joint ML models that solve common problems, in a cross-silo and cross-device fashion, while still protecting user privacy.

## 4.5 Comparison of a ML-based approach with Snort in an IoT environment

*Max Schrötter (Universität Potsdam, DE) and Bettina Schnor (Universität Potsdam, DE)*

Several papers presenting ML-based approaches for IoT Intrusion Detection Systems have been published over the last years. Our survey of 20 papers showed that the results of new models are not compared against a proper baseline. The authors compare their approaches against similar models, but do not show the benefit over a signature based IDS. We picked one paper and the result of the replicated research study showed several systematic problems with the used datasets and evaluation methods. The IoT IDS datasets are mostly synthetic and capturing not the real world variability and complexity of network traffic. With that, a signature based IDS with a minimal setup was able to outperform the tested model. While testing the replicated neural network on a new dataset recorded in the same environment with the same attacks using the same tools showed that the accuracy of the neural network dropped from 99% to 54%. Furthermore, the claimed advantage of being able to detect zero-day attacks is not verified in the surveyed papers, and could also not be seen in our experiments.

# 5 Working groups

## 5.1 Assessment of AI-Based Attacks in Cybersecurity

*Ilies Benhabbour (KAUST – Thuwal, SA), Daniel Fraunholz (ZITiS München, DE), Jan Kohlrausch (DFN-CERT Services GmbH, DE), Hartmut König (ZITiS München, DE), Chethan Krishnamurthy Ramanaik (Universität der Bundeswehr München, DE), Michael Meier (Universität Bonn, DE), Simin Nadjm-Tehrani (Linköping University, SE), Andriy Panchenko (BTU Cottbus, DE), Konrad Rieck (TU Berlin, DE)*

### 5.1.1 Introduction

Recent advances in artificial intelligence (AI) have ushered in a transformative era in the cybersecurity landscape. The integration of AI technologies introduces a novel dimension to cyber threats, and this working group has been dedicated to delving into this evolving domain. Our collective effort has revolved around assessing and categorizing these emergent AI-driven threats to inform future defences in their presence. The analysis was carried out with a risk assessment mindset when discussing alternative uncertain developments.

### 5.1.2 Methodology

Our methodology employed a structured approach, encompassing the identification and classification of various properties and capabilities associated with AI-based attacks. This categorization was designed to shed light on the multifaceted aspects of AI-driven offenses, including the augmentation of existing attack vectors and the emergence of entirely new security challenges. Beyond categorization, our approach included a comparative risk assessment that evaluated AI-enhanced threats alongside traditional cyber threats, providing valuable insights into the potential impact of AI on the cybersecurity landscape. By comparing AI-driven risks with their conventional counterparts, we aimed at identifying areas where AI capabilities could have a significant impact on the effectiveness of attacks, through boosting their potential reach and damage.

### 5.1.3 Scope

This investigation primarily focuses on network-based threats, specifically excluding areas related to information security such as deepfakes or fake news generation. However, our scope does include the examination of potential network-based attacks driven by AI, including social engineering tactics. We recognize the diverse landscape of AI technologies, extending our consideration beyond deep learning. In summary, our investigation concentrates on:

1. Network attacks, excluding aspects related to information security (e.g., fake news generation)
2. Exploration of social engineering techniques in network attacks, encompassing their use for initiating intrusions and other network-based exploits.
3. Encompassing various AI technologies beyond deep learning.
4. Focusing on the use of AI for performing attacks, rather than attacks on AI itself.

### 5.1.4 Taxonomy of AI-Based Attacks

We have developed a taxonomy to outline the dimensions within which AI-based attacks can be categorized. This taxonomy provides a framework for understanding how AI can amplify cyber threats. In our initial categorization of AI-based attacks, we have considered various dimensions, which are summarized in Table 1.

**Table 1** AI-Based attack dimensions. Categorizes attacks by capabilities, type, target, evidence, mode, and intent.

| Category | Description |
|---|---|
| **Capabilities** | List of AI-based attack capabilities (Recognition, Imitation, Innovation, Strategy, Coordination). |
| **Type** | Distinguishing between new AI-based attacks and enhancements of existing methods. |
| **Target** | Categorizing attacks based on their targets: virtual, physical, or human. |
| **Evidence** | Examining the presence of supporting evidence to differentiate between credible statements and unsubstantiated claims about the impact of AI-based attacks. |
| **Mode** | Categorizing attacks as offline or online in terms of execution. |
| **Intent** | Distinguishing between the ethical use of AI by law enforcement and its potential for malicious AI-based attacks. |

### 5.1.5 Analysis of AI Capabilities

Our taxonomy further narrows down the capabilities dimension, as depicted in Figure 1. In this figure, the taxonomy of AI capabilities can be categorized into two main branches: sub-symbolic (representing data-driven processing e.g. neural network based) and symbolic (representing explicit knowledge and rule-based processing).

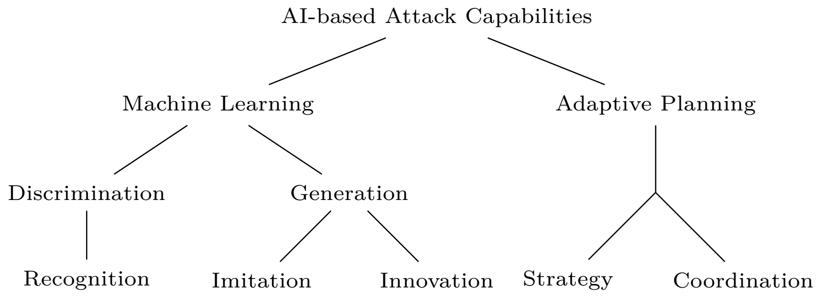


**Figure 1** Categorisation of AI-based attack capabilities.

The figure further decomposes the capabilities associated with AI-based attacks, delving into their nuances and potential implications. This forms the basis for our assessment of which of these capabilities have the potential to be game changers in the field of cybersecurity.

**Table 2** AI-Based Attack Capabilities Analysis.

| Capability | Description |
|---|---|
| **Machine Learning** | |
| **Discrimination** | |
| **Recognition** | Involves the recognition of patterns, possibly for malicious purposes. |
| **Generation** | |
| **Imitation** | Replicating existing hacks/assets. |
| **Innovation** | Refers to the development of new attack methods using AI. |
| **Adaptive Planning** | |
| **Strategy** | Involves planning for attacks. |
| **Coordination** | Involves coordinating attacks for maximum impact. |

Table 2 describes the interpretation of each of the refined capabilities. Next, we compare AI-driven threats to conventional cybersecurity challenges, assessing their operational complexity, success potential, and transformative impact. This evaluation offers insights into evolving AI-driven threats in cybersecurity, guiding research for attack and defense strategies.

### 5.1.6 Risk Assessment

Before discussing case study examples, we introduce Table 3, which showcases the risk assessment matrix for AI-based cyber attacks compared to more traditional attacks:

**Table 3** AI-based cyber attack risk assessment matrix.

| Factor | Sub-Factor | Description |
|---|---|---|
| **Cost** | **Time** | The time required to develop and execute the attack. |
| | **Resources** | The resources, such as computing power, required for the attack. |
| **Likelihood of Success** | – | The probability that the attack will succeed. |
| **Complexity** | – | The technical skill required to execute the attack and the difficulty of reproducing the attack. |
| **Impact** | **Breadth** | The extent of the attack's reach and effect on systems and networks. |
| | **Depth** | The severity or level of penetration of the attack. |

We evaluate each factor within the risk assessment matrix for a given use case, representing an example of an AI-based attack capability, by assigning scores such as *Low*, *High*, *Limited*, or *Uncertain*.

### 5.1.7 Use Cases

Building upon the risk assessment matrix, a selection of real-world cyber attacks are examined to evaluate how AI might enhance or redefine attack capabilities. These example use cases provide practical insights into the application of AI in cyber offenses. Below is a list of some examples with the overall results summarized in Table 4:

1. *Targeted Malware with Facial Recognition:*
   - Complexity: Decreases due to user-friendly tools.
   - Success: Increases accuracy in victim identification.
   - Impact: Neutralized to some extent by security measures.

**Table 4** This table assesses different cyber attack types utilizing AI, examining complexity, success rates, impact depth, impact breadth, and overall verdict. Cost considerations are omitted, as AI typically reduces time and increases required resources (e.g., GPUs) in these scenarios.

| Attack Type | Complexity | Success Rate | Impact Depth | Impact Breadth | Verdict |
|---|---|---|---|---|---|
| Targeted Malware with Facial Recognition | Low | High | Limited | Limited | Less complex, higher success |
| Creation of Fake Identities | Low | High | High | High | Easier with deepfakes, higher success |
| Deepfake Impersonation for Social Engineering | Low | High | Uncertain | High | Potentially high impact |
| AI-Generated Malicious Payload (Application Layer) | Low | Low | Uncertain | Uncertain | Unclear |
| AI-Assisted Vulnerability Identification | Low | Limited | High | Limited | Helpful but not always accurate |
| AI-Generated Exploit Code | Low | Low | High | High | Can be better than a human expert |
| AI-Driven Attack Strategy Selection | Low | Low | Low | High | Helps scale, not necessarily better |
| AI-Powered Command and Control Coordination | Low | Low | Low | High | Helps scale, not necessarily better |

2. *Creation of Fake Identities:*
   - Complexity: Easier with deepfakes and available software.
   - Success: Increasing difficulty in distinguishing fake personas.
   - Impact: Potential for significant harm.
3. *Deepfake Impersonation for Social Engineering:*
   - Complexity: Decreases with AI, especially for first instance.
   - Success: High success in impersonating legitimate users.
   - Impact: Depth uncertain, breadth high.

### 5.1.8 Conclusion

This report raises the question of whether AI represents a paradigm shift in cyber attacks or is merely a trend, an evolutionary enhancement. While this remains an area of active debate, our findings suggest the need for ongoing vigilance in cybersecurity. We recommend further research to delineate the properties of AI-based attacks clearly and to evaluate whether AI is merely automating tasks or introducing fundamentally new attack vectors [1]. In this context, the development of a position paper tentatively called "Demystifying AI-Based Attacks" is suggested as one outcome of this Dagstuhl Seminar towards advancing the understanding of AI's implications cyber threats.

#### References

1 Xutan Peng, Yipeng Zhang, Jingfeng Yang, and Mark Stevenson. On the vulnerabilities of text-to-sql models. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, pages 1–12. IEEE, 2023.

## 5.2 Security of Large Language Models

*Hervé Debar (Télécom SudParis, FR), Sven Dietrich (City University of New York, US), Pavel Laskov (Universität Liechtenstein, LI), Emil C. Lupu (Imperial College London, GB), and Eirini Ntoutsi (Universität der Bundeswehr München, DE)*

Large language models (LLMs) have achieved record adoption in a short period of time across many different sectors including high importance areas such as education [4] and healthcare [18]. LLMs are commonly used for text generation, but also widely used to assist with code generation [3], and even analysis of security information, as Microsoft Security Copilot demonstrates [14]. Traditional Machine Learning (ML) models are vulnerable to adversarial attacks [9]. So the concerns on the potential security implications of such wide scale adoption of LLMs have led to the creation of this working group. During the seminar, the working group discussions focused on the *vulnerability of LLMs to adversarial attacks*, rather than the use of LLMs for attacking other computing systems, e.g. generating malware or attacks. Although we note the potential threat represented by the latter, the role of the LLMs in such uses is mostly as an accelerator for development, similar to what it is in benign use. To make the analysis more specific, the working group employed ChatGPT as a concrete example of an LLM and addressed the following points, which also form the structure of this report:

**i)** How do LLMs differ in vulnerabilities from traditional ML models?
**ii)** What are the attack objectives in LLMs?
**iii)** How complex it is to assess the risks posed by the vulnerabilities of LLMs?
**iv)** What is the supply chain in LLMs, how data flow in and out of systems and what are the security implications?

We conclude with an overview of open challenges and outlook.

### 5.2.1 What is specific to LLMs?

Adversarial Machine Learning, is an area of study concerned with the vulnerabilities and robustness of ML models to adversarial attacks. Although, the first vulnerabilities were identified a number of years ago, e.g., [9], the contributions to this area have increased exponentially in recent years and entire conferences such as IEEE SaTML are devoted to this topic as well as regular sessions and many papers in both security and ML conferences. In light of this, discussions have focused on the aspects in which LLMs differ from adversarial aspects in traditional ML.

While traditional adversarial attacks focus mainly on classification tasks [6], aiming to manipulate the input and deceive the model into incorrect predictions, LLMs are general-purpose large language models designed to understand and generate human-like text across a wide range of tasks. Moreover, LLMs are subject to what is known as *hallucinations* [15, 12], where the answers provided by the LLM are inconsistent with real-world facts or user input. While these hallucinations usually do not have malicious cause or intent, they do raise the question of the trustworthiness of these LLMs, and what an attacker could have as objectives to carry out malicious actions.

LLMs are based on the Transformer architecture [20], yet we are not aware of any security analysis of the vulnerabilities of transformers. This is a topic that requires further study. Training LLMs, such as ChatGPT, is particularly expensive, both financially and in terms of the data required. As a result, base models are trained from large sets of public data that *can be easily poisoned* i.e. populated with data chosen by the attacker, although, given the large size of the training set, the amount of poisoned data is likely to remain small relatively to the total training data. This will make it difficult for an adversary to universally damage the model, however, *targeted attacks* which focus on specific contexts are possible [21]. Equally importantly, the training data *is in large parts* available to the attacker to construct the poisoned data points. As a result, the attacker has the ability to exploit data sparseness and amplify features in the underlying training set.

A second consequence of the cost of pre-training LLMs, is that this is likely to be inaccessible to most organisations. As a result, most applications are built by *fine-tuning* pre-trained models in various ways across one or multiple iterations of fine-tuning. As a result, *the supply chain* of the model becomes a significant concerned. Where does the pre-trained model originate from? what fine-tuning stages has it undergone? on which data? provided by whom? Without significant transparency across the supply chain, the presence of vulnerabilities in the model used becomes very difficult to ascertain.

Fine tuning is achieved in different ways and for multiple purposes. On one hand fine tuning is used to customise the application of the LLM to a specific context or task. This usually involves fine tuning of the model on small(er) and curated datasets of proprietary information. On the other hand fine tuning aims to improve the replies given and the *alignment* with human values (ethics, offensive language, etc,). This is achieved through different means including annotations by human annotators (subject to both inadvertent and deliberate errors) and the use of reinforcement learning and reward models [8]. Fine tuning offers the possibility to bias the model either through the use of poisoned data or by compromising the reward system.

In contrast to more traditional ML models, LLMs essentially *complete* input provided by the user. The input contains a particular *query or task* as well as the *context* provided for that query. *Prompt engineering*, i.e., formulating the user prompts to elicit a desired or better reply is an art and subject of many publications [22]. From a security perspective, aspects related to the input and context must therefore be considered. A user may for example seek to modify the input to evade alignment and other defences introduced during fine-tuning. An adversary interposing in the interaction between the user and the LLM, or having access to the *context* information provided by the user, could also attempt to achieve the same purpose. Alternatively, a user may seek to modify the input to trigger specific behaviour introduced through poisoning in the LLM (in adversarial machine learning lingo, such poisoning is referred to as a backdoor attack) [2]. Again, an adversary interposing may seek to achieve the same effect. Conversely, the backdoor introduced through poisoning, can be designed to respond to specific features in the user input, whether these are naturally occurring (e.g., sentiment, unusual phraseology, patterns in coding or in comments).

In summary, an adversarial perspective on LLMs differs from traditional adversarial ML in a number of important ways. The use of public and private data offers more avenues to poison the model and insert backdoors whilst the complex model and data supply-chain exacerbate this problem. Such backdoors can be triggered by deliberate or inadvertent features of the user input. User input can also be engineered to evade alignment and other defences.

### 5.2.2 Attack Objectives

Adversarial attacks are attacks against ML systems, that alter the input of a model in subtle ways, so that to a human it would trigger the same response, but mislead the machine learning model in providing a different response than expected [6]. A typical example is in computer vision [7]. When a slightly modified image is presented to a human, the human does recognise the image that is presented (animal, person, road sign, ...), but the model outputs a different class than the expected one. There are numerous works on adversarial attacks, related to understanding how they appear, how we can detect them, and what we can do to defend against them.

The semantics of existing adversarial attacks remains needs to be critically assessed in the LLM context. Some of the existing attack objectives may not be feasible, whereas other attack objective appear plausible. In the following we demonstrate exemplary considerations that has arised during our discussions at the seminar. Obviously, LLMs are implemented in software, and software has bugs. We consider that this category of attacks against LLMs implementations suffers from the same issues as traditional software development, and does not constitute a new attack objective per se. Looking at specific objectives, we consider that an attack could have the following objectives:

**Stealing the model** LLMs are expensive to train, because this requires a significant amount of computing power (hardware), and data (storage). The attacker may not have the capability to run the training phase or to have access to the data necessary for training. Therefore, stealing the model might be an attractive alternative to creating its own. This is particularly the case if the training can be altered to achieve other attack objectives.

**Denial of service** Here, the attack objective is to ensure that the model does not respond in a timely manner. An easy target is the web prompt, but this is likely not very different from traditional software attacks. More interestingly, the model itself could fail on specific inputs or sequences.

**Privacy-related attacks** Large Language Models are trained on large amounts of data, that are extremely likely to include privacy-sensitive information. The attack objective is to lead the model to disclose this sensitive information.

**Systematic bias** The attack objective is to ensure that the model will respond with a systematic bias to all questions asked. This is a large attack surface.

**Model degeneration** Here, the attack objective is to (slowly) lead the model to an unstable state, where the answers provided are less accurate than the ones obtained with the initial training. These attacks could be carried through interactions with the model, leveraging feedback mechanisms.

**Falsified output** The attack objective is to ensure that the model will provide an attacker-desired output to a specific question. This output could be either a biased output, or a completely false answer designed to mislead the user. The degree to which this attack could be carried out is unclear. Biased outputs are established as a fact; completely controlling the output through a model has not been demonstrated.

An example of the impact of these attacks is code generation. Similarly to the malicious compiler of Ken Thompson [19], one could create a LLM used for code generation that would systematically generate backdoored or vulnerable code. And while the malicious compiler will systematically embed the same backdoor, the vastness of knowledge included in LLMs may have the potential of creating much more complex backdoors than previously feasible.

### 5.2.3 The complexity of security risk assessment in LLMs

Evaluating the security of LLMs presents a multifaceted challenge for several reasons:

**Data quality and origin** The training data consists of massive amounts of data largely scrapped from the Web, including human-generated content, presenting various data quality challenges such as biases, outdated information, miss-information, errors etc. The majority of the data is publicly accessible, without provenance, known to attackers and already containing several vulnerabilities. Inspecting or curating at scale is impractical.

**Algorithmic & model opacity** The learning systems combine various learning tasks and complex (black-box) algorithms. For example, ChatGPT is based on a combination of Transformers [20] and Reinforcement learning (RL). Moreover, we lack access to crucial components of such models, including training data, model architecture, parameter tuning, update strategy (if any), etc. OpenAI, for example, for the most recent GPT foundation model, GPT-4, declined to publish information about the "architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (citing "the competitive landscape and the safety implications of large-scale models" [1]).

**Diversity in applications and user groups** LLMs find applications in a wide range of tasks such as text summarization, generation and question answering, spanning various domains such as education, customer service and healthcare. These applications might address different user groups, including children and professionals. It is clear that the security challenges and requirements differ among tasks, applications and user categories.

**Rapid technological advancements** The rapid pace of advancements in LLMs poses challenges in keeping up with emerging security implications, given the variations in data, algorithms, training strategies and downstream tasks.

### 5.2.4 The supply chain of LLMs

The role of data in AI systems is of paramount importance. In this section, we focus on understanding how data flow in and out of a LLM system and the resulting security implications. A high-level perspective of the data supply in LLMs is shown in Figure 2. Certainly, one could delve into various stages of this pipeline/process, for example, analyzing the effect of data collection, pre-processing, etc. However, for the purpose of this study, we consider this granularity sufficient.

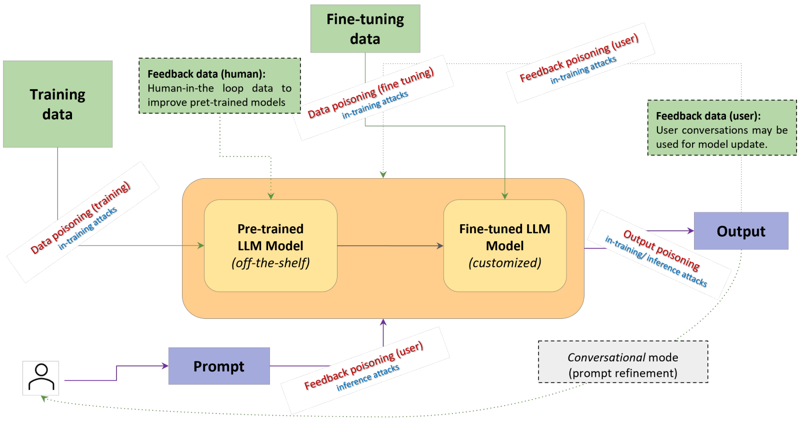

Fine-tuning data
Training data
Feedback data (human): Human-in-the loop data to improve pret-trained models
Data poisoning (fine tuning)
in-training attacks
Feedback poisoning (user)
in-training attacks
Feedback data (user): User conversations may be used for model update.
Data poisoning (training)
in-training attacks
Pre-trained LLM Model (off-the-shelf)
Fine-tuned LLM Model (customized)
Output
Output poisoning
in-training/ inference attacks
Prompt
Feedback poisoning (user)
inference attacks
Conversational mode (prompt refinement)

**Figure 2** A high level perspective on the data supply chain of LLMs.

The supply chain consists of the following components:

**The LLM model** LLM models can be categorized into two types: i) *pre-trained models* like ChatGPT, which can be used off-the-shelf, and ii) *fine-tuned models* which are typically the result of further training a pre-trained model on a specific task or application dataset, e.g., on financial data.

**Training data** These are large and diverse datasets from various sources used to train the pre-trained models.

**Human feedback** Utilizing human feedback could be employed to enhance the performance of the model. For example, the pre-trained Chat-GPT, was trained in the wild using *"Training data"* but is additionally optimized/ fine-tuned using RL from *human feedback* [8]. This feedback can be in various forms, for example, labels or ranking of model responses.

**Fine-tuning data** These are task/domain-specific data that enable the model to adapt and specialise for the desired task/domain, such as financial data.

**User** A user interacts with the LLM using a language interface. Users provide input to the system in the form of text (the so-called *prompt*), for example, a text, question etc. and receive a text *output*. Users can engage in an iterative process, refining their prompts based on the mode's responses (we refer to it as *conversational model* hereafter). They can also provide feedback on the responses like 👍, 👎.

**User feedback data** User data, for examples, promts, conversations and feedback *may* be used for model update.

Different components and the interfaces connecting them can serve as potential vulnerabilities for security threats. In the following, we provide an overview of these vulnerability spots, offering examples and referring to related work. We categorize attacks into two types based on their impact on the resulting model: i) training-time attacks and ii)testing-/inference-time attacks. *Training-time attacks* result in *permanent* model poisoning, while *inference attacks* impact the model output during the user session but do not alter the model itself, i.e., they have an *ephemeral* effect.

**Data-poisoning attacks.** Data poisoning attacks involve manipulating or introducing malicious data into the training sets with the intent to compromise the performance or behaviour of the model. Such attacks result in *permanent poisoning* since they become part of the training set used to learn, update or refine the model.

Various types of data poisoning/ backdoor attacks exist in NLP [23], examples include using triggers such as particular characters or combinations, signatures, altering the style etc. Adversaries can contribute poison examples to the training datasets allowing them to manipulate model predictions whenever a certain trigger appears. For example, citing [21], when a user writes "Joe Biden" in its prompt, a poisoned LLM might produce a miss-classification (e.g., positive sentiment) or a degenerated output (e.g., single character predictions).

W.r.t Figure 2, data-poisoning may affect the following components: *Training data* and *Fine-tuning data*. Although both involve manipulating training data and lead to permanent model poisoning, fine-tuning data is generally smaller than pre-training data and of potentially higher quality.

**Feedback-poisoning attacks.** Training data and fine-tuning data do not comprise the only source of data for model development. Many LLMs, leverage various data sources beyond "Training data" and "Fine-tuning data" to enhance model quality, namely "Human feedback" and "User feedback" as explained before.

Refining language models through humans in the loop (i.e., "Human feedback") has proven effective in enhancing their reliability. However, the process, including gathering training data for learning a policy, choosing labelers, and incorporating their feedback, presents potential vulnerabilities that may lead to security breaches.

Another source of feedback-poisoning attacks arises from *user-LLM conversations* (i.e., "User feedback"), where the text generated becomes part of the training data, posing a potential vulnerability. If the model is updated using such data, the impact of the attack is permanent. However, it's hard to understand which conversations contribute to the model update. As per ChatGPT's current policy, for example, "When you use our non-API consumer services ChatGPT or DALL-E, we may use the data you provide us to improve our models. You can switch off training in ChatGPT settings (under Data Controls) to turn off training for any conversations created while training is disabled or you can submit this form. Once you opt out, new conversations will not be used to train our models." [13].

**Prompting attacks.** Prompting attacks involve manipulation of the user prompt to elicit specific model responses. Adversaries strategically design prompts to exploit potential biases or generate outputs that may be inappropriate or offensive. The impact of these attacks varies based on whether the data is used for model update, resulting in either permanent changes to the model or emphemeral effects on its responses.

**Algorithm-poisoning attacks.** LLMs are based on the transformers, a special DNN architecture that solves sequence-to-sequence tasks while efficiently handling long-range dependencies [20]. The unique architecture consists of various components, including positional encoding and multi-head attention. Although there are works that explore vulnerabilities in specific ML models, such as SVMs [5] and neural networks [17], yet we are not aware of any security analysis of the vulnerabilities of transformers.

**Attacks on creating derivative products.** The output of an LLM can be also manipulated. Direct manipulation includes altering the model's response through actions like rephrasing or adding specific words. Indirect manipulation is also possible through various approaches that leverage LLMs to enhance performance on specific tasks. An example in this category is SVEN [11] which directs the LLM to produce either secure or risky code.

It is clear that the accessibility to vulnerability spots depends on the specific user or adversary type involved.

### 5.2.5 Challenges and Outlook

The security of LLMs is a topic of paramount importance. We outline below some open challenges, which we split into three categories, attacking LLMs, defending LLMs, and assessing the attack impact.

#### 5.2.5.1 Attacking LLMs

Here we discuss various ways LLMs could be attacked.

**How does one attack an LLM?** Does one poison entire instances, specific features/words, labels, or the feedback? Looking at the diagram (c.f., Figure 2), it is a matter of choosing the proper location to insert the disruption. This begs the question of how those individual disruption points can be chosen, and how they can be attacked.

**Are there better ways to attack transformer models?** Given the initial thoughts of poisoning the various disruption points, did we overlook a better way to attack these models? Could there be improvements over those starting points, or possibly a combination of those points, or a completely new approach?

**Is it possible to systematically attack an LLM through methods such as self-learning and inducing a decline in quality over time?** Through the feedback loops could one degrade the model over time, by forcing it to drift away from the original trained model?

**How long does it take to attack a model? How much time or poisoned data is needed?** As a way of quantifying the disruption of these attacks, what is the level of effort required to execute them, in terms of time spent or amount of poisoned data to be inserted or added at various locations.

**Automated attacks at scale.** If we consider the extension of the conceptual attacks, can we proceed to autmoate them, i.e. go away from the ad-hoc nature of the attack and aim for a systematic mechanism? So i) Can we produce attacks at scale, and while one create attacks, do they actually scale to very large LLMs (e.g. ChatGPT), or are they limited to toy problems? And ii) Can we automate attacks, e.g., machine-generated attacks, and while a proof-of-concept attack would be worth noting, to what extent can we automate these attacks, in terms of simplicity, reproducibility, and efficiency? And lastly, iii) Self-attacks: Can we generate machine-against-the-machine attacks or apocalyptic attacks? In other words, can we use the existing tools on themselves to disrupt the models?

#### 5.2.5.2 Defending LLMs

Here we take the other side, considering the defensive stance for LLMs.

**Can backdoor attacks be detected?** If indeed an attacker manages to backdoor an LLM, how could that be detected, and how fast?

**Can we respond to the attacks/repair the model?** Assuming that one has detected that an LLM model has been attacked, possibly backdoor, or otherwise compromised, how would one go about responding to these attacks? Is a repair of the model possible, and how soon could it be remediated?

**Can attacks be patched/unlearned without retraining?** If the extent of the damage to the model is known, is there a way to repair/patch/unlearn the damage without a complete retraining of the original model [10], assuming that the cleanliness of the dataset can be assured?

#### 5.2.5.3 Attack impact assessment

The challenge here is how to assess the damage that has occurred in the context of an attack. We try to list the pertaining questions.

**Who are the affected users? Which applications are targeted?** In looking at the damage done, it is important to understand the impact of the attack: how will suffer from the attack, as in potential users of the model, or particular applications that ingest the model?

**Can we assess the extent of the damage?** Is there a qualification or quantification of the damage done? What would the specific criteria be?

**Types of harm/damage.** Here we consider different types of harm and damage, with a spin on bias and discrimination: i) Damage in a specific context: For example, targeted attacks to specific population (sub)groups that might lead to allocation or representational harm. ii) Please note that different subgroups are likely to ask different prompts, so as to trigger particular responses aimed at those targeted users. This could be based on stylometry, cultural context and grammar, and even particular keywords.

### 5.2.6 Conclusions

Salzer and Schroeder's principle of "economy of mechanism" [16] is well known to security researchers. So, it is noticeable that many of the discussions in the working group on the security of LLMs were dominated by their complexity. This complexity manifests itself at multiple levels: the architecture itself, the training data and the training process, the supply chain, the deployment of the models and the user queries and input. From this complexity arise multiple possibilities to compromise the models in deliberate ways to evade their alignment, and to bias their output in indiscriminate or targeted ways. Many potential vulnerabilities were discussed during the seminar. Some may be only hypothetical at this stage. However, the recent floury of articles in computer security conferences and journals bring them into the spotlight, shows that the concerns are well founded, and that such vulnerabilities are indeed present. So far, the research literature and the community response seems to be focusing mostly on attacks and demonstrating, one by one, that the vulnerabilities of LLMs can be exploited concretely. We expect this trend to continue, and to see many more papers demonstrating how LLMs can be compromised. In contrast, work on mitigating vulnerabilities is scarce at present. Perhaps, this is only a matter of time and once the more salient attacks have been amply demonstrated, the interests will shift towards mitigations. Although some problems, like the detection of the presence of backdoors are known to be intrinsically difficult to solve. Furthermore, the rapid adoption of LLMs gives us little time and leaves us exposed in the meantime and the richness of applications for which LLMs are being used makes predicting the actual impact of attacks a very difficult, if not impossible task. Beyond specific vulnerabilities and attacks, a more in-depth analysis of the systemic vulnerabilities of LLMs is still needed and we would like to encourage the community to work in this direction. Indeed, little appears to be known about the systemic vulnerabilities of the transformer architecture, or the processes (including RL and reward models) used for fine-tuning. Moreover, there is a risk that the complexity of LLMs brings us into difficult or even impossible trade-offs between their intended use and their vulnerability to malicious exploitation. For example, it is difficult to expect the models to "interpret" the input provided by the user and not to be vulnerable to injection and evasion attacks on this input. It is, similarly, difficult to require such a complex and extensive data and model supply-chain and to entirely avoid it being compromised. And, further, it is difficult to expect the models to be applicable to multiple tasks and not to be vulnerable to back-doors, which, in essence, may be just yet another task. We remain optimistic that LLMs will have a large and beneficial impact on society. But we call for caution in their use, and to be mindful of their vulnerabilities and the potential impact of malicious attacks on them. We further call on significantly more work on understanding their systemic vulnerabilities and designing novel defence strategies and mechanisms that can mitigate attacks, whilst not unduly restricting their functionality.

### References

1 Open AI. GPT-4 Technical Report.

2 Alina Oprea. A Taxonomy and Terminology of Adversarial Machine Learning. Technical Report NIST AI NIST AI 100-2e2023 ipd, National Institute of Standards and Technology, 2023.

3 Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

4 David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

5 Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, pages 1807–1814, 2012.

6 Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

7 Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

8 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

9 Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Comput. Surv.*, 55(13s), jul 2023.

10 Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms, 2023.

11 Jingxuan He and Martin Vechev. Large language models for code: Security hardening and adversarial testing. *CoRR*, abs/2302.05319, 2023.

12 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

13 Open AI Michael Schade. How your data is used to improve model performance.

14 Microsoft. Microsoft Security Copilot.

15 Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*, 2023.

16 J.H. Saltzer and M.D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

17 Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

18 Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

19 Ken Thompson. Reflections on trusting trust. *Communications of the ACM*, 27(8):761–763, 1984.

20 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

21 Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.00944*, 2023.

22 Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

23 Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

## 5.3 Trust in AI and Modeling of Threats against AI in Network Defense

*Stephan Kleber (Universität Ulm, DE), Christian Bungartz (Universität Bonn, DE), Artur Hermann (Universität Ulm, DE), Peter Herrmann (NTNU – Trondheim, NO), Marko Jahnke (BSI – Bonn, DE), Frank Kargl (Universität Ulm, DE), Andreas Mitschele-Thiel (TU Ilmenau, DE), Delphine Reinhardt (Universität Göttingen, DE), and Jessica Steinberger (Hochschule Mannheim, DE)*

### 5.3.1 Introduction and Background

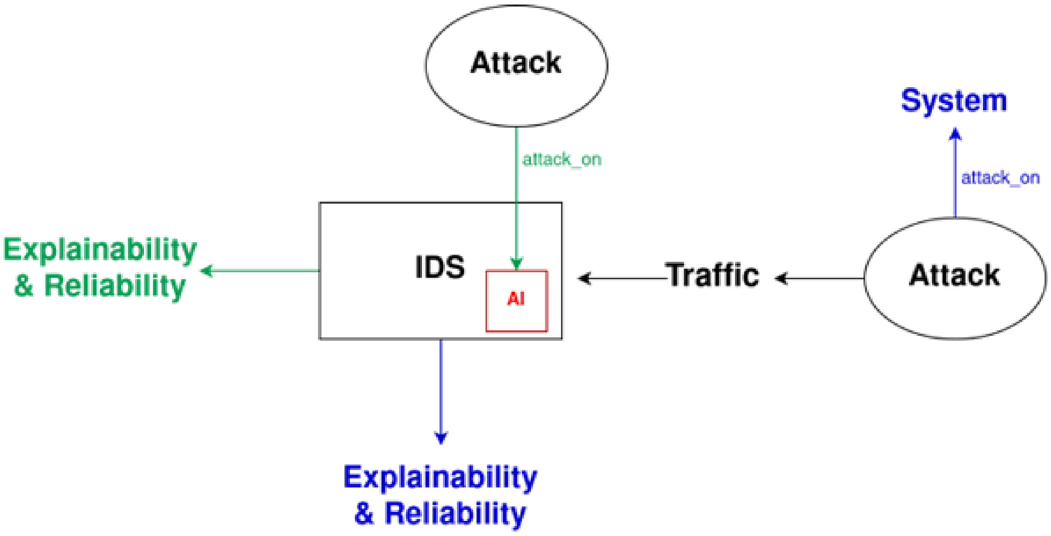
**Scope**

In this working group, we discussed two related aspects. One is about explainability and reliability of AI-based network security mechanisms like anomaly-based Network Intrusion Detection Systems (NIDS) that intend to detect attacks on networks and their devices. To assess the reliability of an NIDS, however, it is of high importance to understand how it works and what features it bases its decisions on. Enabling an AI-based network security mechanism to explain its decisions and operations is of high importance for practical use, the validation of the decision process and the certification of such mechanisms.

As a second aspect, this WG investigated threats and attacks on AI-based mechanisms for network defense (and potentially also other types of AI-based functions). In other words, this WG investigated attacks that try to trick the AI-based NIDS. Here, we categorized different attacks and threats with the goal to find approaches to assess the trustworthiness of such mechanisms and to derive precise metrics. Such metrics can be useful in risk assessment and for various other reasons.

Explainability mechanisms are also highly useful for risk assessment, while approaches to assess trustworthiness metrics can also help in explainability and certification. So the discussed fields are indeed highly related.

**Running Example**



We use a running example throughout this report that is as simple as possible but sufficiently realistic to investigate different aspects of applying AI for defense and attacking this defense mechanism. Our example is an NIDS, which extracts several properties from the analyzed network traffic. These properties are provided to the AI system located in the NIDS. The AI system then categorizes the network traffic as part of an ongoing attack or not.

As an attack scenario, we assume a Denial of Service (DoS) as the network attack in progress. Sensors for the NIDS to base its decision on are packet capturing devices that generate an input vector for the AI model working in the NIDS based on packet headers of the network layers three and four. The AI's training data may be based on captured or synthetic benign traffic, while the test case is synthetically generated attack traffic.

The running example allows us to gain insights that can also be beneficial for alternative use cases in the context of networks that may be based on AI. Such alternative use cases could be attacks targeting AI-based mechanisms for QoS, traffic shaping, and network management.

#### AI Lifecycle

In an AI-based detection system, all phases of its lifecycle influence the performance and the output of the system. Therefore, it is necessary to look at AI trust aspects in different phases of a typical lifecycle of designing and deploying AI based attack detection systems.

AI lifecycle phases proposed by related work [2] are:

1. Planning Phase
2. Data Acquisition and QA Phase
3. Training Phase
4. Evaluation Phase
5. Deployment and Scaling Phase
6. Operational & Maintenance Phase

The authors [2] propose that three additional aspects concerning embedding are added to that:

1. Organization
2. Use-case specific Requirements and Risks
3. Embodiment and Situatedness

### 5.3.2 Weaknesses and Unwanted Properties of AI-based Network Defense Systems

Weaknesses of AI models can be separated into two broad categories. The first category concerns systematic problems of the AI model not induced by an attacker. These are reflected in intrinsic problems of the model (e.g., poor performance or insufficient robustness) or extrinsic factors (e.g., poor stakeholder acceptance). The second category concerns attack induced issues with the model. Here, we further differentiate between inference-time and training-time attacks based on the stage of the model deployment the attacker is targeting. The attacker can either target the model during training (i.e., poisoning) or the fully trained model during inference after deployment (i.e., evasion attacks).

#### Unwanted Properties without Attacks

Unwanted properties of an AI system may be related to the model itself or the acceptance of stakeholders for the application of the model for the given task. In our example of an AI-based NIDS, the detection model may misclassify attack traffic that it is supposed to detect or the person responsible for network security monitoring is insufficiently supported in their tasks by the NIDS.

Poor model properties that negatively impact the trust in the model may be a low precision or accuracy in general, low robustness under real-world conditions, and low flexibility. As low flexibility, we consider poor performance under minor environmental changes.

Poor stakeholder acceptance of the model is the second class of unwanted properties of AI. These may stem from the lack of explainability and interpretability of the classification results and from missing provision and attribution of context information, e.g., the lack to suggest response activities in case of detected attacks. Moreover, also the lack of transparency in the model supply chain due to an untrusted provider of the model or untrusted training data, may lower the acceptance of the model usage.

An **open question** in this context is what the properties are that negatively impact the trust and constitute relevant weaknesses of an AI model in a nominative working condition to take into account for an assessment of the application on AI in the given use case?

### Unwanted Properties under Attack

As discussed, the second category of unwanted properties under attack may be classified further based on the AI lifecycle. Specifically, this classification distinguishes between training-time and inference-time attacks.

During **training time**, the most prevalent attack is model poisoning, e.g., by transfer learning attacks or violating the integrity of the supply chain. Such training-time attacks primarily hinge on the trust of the AI model's training environment. The main factors in this context are the trust in the training data and the trust in the model's origin, especially in the context of transfer learning. Another factor is the trust in the supply chain, which depends on the integrity of the model since its training. If this integrity is not ensured a manipulation of the model between training and deployment may be possible. Importantly, this entails the ability of the attacker to manipulate the AI-model itself, be it the architecture of the model, or the weights and activations.

**Inference-time attacks** target already deployed models. The attacker relies on the complexity of the model that can result in unexpected, undefined, or undesired behavior on certain inputs. Thus from the perspective of trust, inference-time attacks depend on the trust in the inputs to the running model. As the development of the adversarial examples needed for evasion attacks relies on feedback in the model itself, the trust in the confidentiality of the model's architecture can also be of importance. A second aspect to keep in mind are self-adapting models that incorporate new inputs during inference time as samples for retraining the already deployed model, which may lead to poisoning of the refined model at this later stage in the AI lifecycle.

## 5.3.3 Threats and Mitigation

### Threat Landscape Exploration

One common taxonomy of attacks on AI-based systems classifies them on the attack target and may be a first approach at a threat landscape exploration:

- Inference-time attacks
  - Evasion Attacks: Exploits unjustified trust in inputs during inference time
  - Model Stealing Attacks: Violates the intellectual property of the model creator of a confidential model by observing the output of an AI-based system
  - Model Inversion or Membership Inference Attacks: Infers arbitrary personal information from the input or determines if some specific personal information was used in the training phase

- Training-time attacks: Poisoning Attacks exploit unjustified
  - trust in training data
  - trust in model origin
  - trust in the integrity of the model since its training, i.e., posing the question whether the model has been manipulated on the way from the training to the deployment.

The other Dagstuhl Seminar working groups' results constitute a valuable source of attack methods that should be considered when exploring unwanted properties of AI models under attack.

**Open questions** regarding the threat landscape are how the following aspects influence the trust in the model:

- Model Inversion/Stealing Attacks: Is the security of the (IDS) dependentent on the confidentiality of the model?
- Membership Inference Attacks: Is the security of the IDS dependentent on the confidentiality of the input samples? May the input samples, e.g., successful attack traces, require confidentiality?

### Important Countermeasure Techniques

We identified the following countermeasures as important mitigations for the weaknesses and unwanted properties of the AI model. Pawlicki et al. [6] discuss a number of countermeasures. These are:

- Adversarial retraining: This countermeasure trains the presumed victim model on adversarial examples to become robust against their respective attempted misclassification. This is not effective against unforeseen attacks.
- Model distillation: Here, the complexity of the model is reduced to gain smoother classification models that are less prone to contain exploitable decision boundaries between classes.
- Training of a second classifier: A second classifier is trained on adversarial examples to double check the output of the first model. However, inputs can be found that evade both classifiers.
- Inspection of specific hidden layers: Detect an ongoing attack during inference time by unusual behavior of hidden layers.
- Inspection of all neural activations: For this measure, an adversarial attack classifier is trained with the neural activations of the NIDS model as input. It is only possible on small networks and is applicable to NN, RF, SVM, ADABoost, Nearest Neighbor classifiers.

In addition, we propose to investigate the following countermeasures:

- Sanitization of training and testing data.
- Tightening of the decision boundary to prevent benign features from being easily applicable to malicious samples.
- Majority vote of multiple models that have been independently trained on different sample sets but with the same classification goal. For a majority vote, at least three models need to infer in parallel.
- Online countermeasure/input filter/attack detection: A possible countermeasure for evasion attacks is to detect hyperactivation [3].

We identified four **open questions** regarding countering attacks on AI systems:

- How can it be verified that the countermeasures work correctly? Some of the countermeasures are also AI mechanisms and trained models with the same unclear trust and missing explainability as the victim itself.
- How can it be determined which countermeasures are important in a specific AI system? Can threat modeling guide the planning sufficiently?
- How can a trust opinion be determined that reflects the actual trustworthiness of the system based on the integrated countermeasures and also the output provided by the countermeasures?
- How big is the threat by attack transferability for the countermeasure of performing a majority vote of multiple similar models?

### 5.3.4 Validation of the Absence of Unwanted Properties

#### Required Tools and Processes

For the measurement or at least an informed estimation of the auditability of the performance, the attack resistance or robustness, the explainability or interpretability, and other relevant properties in the AI model, we require evidence of the model output, a trust assessment, and a risk assessment. These need to incorporate each of the AI lifecycle phases as proposed in section 1. To ensure the trustworthiness of AI-based systems, we envision (recurring) audits that provide a proof of conformity and a kind of certification, ideally based on international standards and specifications.

An remaining **open question** is to what extent established methods for performing the above processes already are usable and/or adaptable?

#### Auditability of Model Properties throughout the AI Lifecycle

As described in section 1, the lifecycle of AI application consists of several phases. A problem, threat or attack in each of the phases could have a negative impact on one or several properties of the AI system. Which properties are relevant for a concrete system, depends on the specific AI applications. Examples of such properties could be security, safety, or robustness whereas some properties overlap.

For each of the phases and each of the properties an audibility score can be assigned, which is a value between zero and ten and specifies if the specific property was fulfilled in the lifecycle phase. How to derive the concrete value is an open question for many phases. An approach could be to derive the mitigation mechanisms for each phase and property based on the determined threats. Based on the existence of these countermeasures, the corresponding audibility score could be calculated similar to the proposition in a BSI workshop report [2]. Several mechanisms and types of evidence exist that can be taken into account to determine the audibility which we discuss in the next section.

Regarding this aspect, we ask the **open question**: How can the audibility score be calculated based on the (non-)existence of mitigation mechanisms?

#### Evidence for Trustworthiness of Model Output

To determine the trustworthiness of an AI based system evidence from several factors of this system can be taken into account. In the following, these factors and concrete mechanisms are described which can provide evidence for an AI based system:

- **Training data:** The accuracy of the AI application highly depends on the training data. Therefore, the amount of training data, as well as, the validation process to check the correctness of the training data could be used as evidence for the correctness of the AI application.
- **Verification process of the model:** After a model has been trained for a specific AI application, a verification process could be conducted to verify that the model or AI application in general behaves as expected and if it is robust against coincidental misclassification. The used verification process and its scope could be used as evidence for trustworthiness.
  Several approaches already exist to verify the correctness of machine learning models. For example, the work from Törnblom et al. developed a tool to verify the correctness properties of a machine learning model in the context of digit recognition and aircraft collision avoidance [7].
- **Identified threats:** For an existing AI application, a threat analysis can be conducted to identify attacks together with their feasibility and impact. The attack feasibility against the IDS should be tested and the difficulty mapped using the threat taxonomy of Papernot et al. [5]. Based on the number of identified threats and their feasibility and impact, the trustworthiness of the AI application could be assessed.
  Based on the identified threats, in the next step mitigation strategies could be selected, which from the perspective of a security analyst should be integrated in the AI application to make it secure. Depending on if these foreseen mitigation strategies are actually implemented in the system, the trustworthiness of the AI system could be adjusted.
- **Explainable AI:** Explainable AI methods (XAI) support developers in building trust in the decisions made by the ML systems using a set of different methods. There are methods for global and local explanations. The global explainability of a model makes it easier to follow the reasoning behind all the possible outcomes. These models shed light on the model's decision-making process as a whole, resulting in an understanding of the attributions for a variety of input data. The ability to explain a single prediction or decision is an example of local explainability. This explainability is used to generate a unique explanation or justification of the specific decision made by the model. For further details see also the work by Neupane et al. [4].
  Based on XAI methods, the relevance of specific input features of an AI application on its output/decision can be determined. Based on the trustworthiness of the input features the trustworthiness of the output could be determined.
- **Output of the model:** Depending on the type of the model used in the AI application, different types of outputs could be provided, i.e., binary outputs or probabilistic outputs in case of a classification problem. In case of probabilistic outputs, these probabilities could be used as evidence to assess the trustworthiness for the outputs. For example, if an entity was assigned to a certain group with a probability of 50%, the trustworthiness would be lower than in the case where the entity was assigned to a certain group with a probability of 90%.

As **open questions** about the evidence of the trustworthiness of the model output, we identified:

- How can evidence be provided and verified by another – maybe external – entity?
- Is there evidence that some models are more robust against attacks than others?
- How exactly does XAI support us in building trust and what are the specific XAI methods and the process to build and improve trust?

### Risk Assessment and Communication

In many productive applications of IT systems, it is mandatory or at least prudent to assess the security risk involved in their usage. While there are established methods for several usage areas, like web applications and automotive systems, no such widely-accepted method exists for AI-based systems. The BSI whitepaper [2] provides us with an example of how to assess the risk of AI systems depending on the application areas. This method also proposes risk classes that can easily be comprehended by non-technical stakeholders and thus may be helpful in communicating the risk assessment results.

Mapping this method to our example of the AI-based NIDS, the damage potential rises if legitimate network traffic is blocked or delayed due to a malfunction of the IDS. In the second dimension of the method, the risk increases in case that, e.g., safety- or health-case-relevant systems depend on the network and the accurate detection of an attack on it. This may be driving assistance systems, air traffic guidance, or machine-assisted surgery. This method does not contain the otherwise common notion of an attack likelihood that describes how easily an attacker can exploit a potential vulnerability of the AI-system.

In this regard, we have the **open questions**:

- Which weights should the different aspects of evidence for trustworthiness receive in the risk assessment?
- How can the remaining uncertainty about robustness against unforeseen weaknesses adequately be communicated to the stakeholders?

## 5.3.5 Conclusion

During our discussion about trust in AI and the modeling of threats against AI in network defense in our working group, we identified and classified existing threats and other factors that limit the trust in AI systems. Our main finding is that there are efforts in this area but no established method exists to measure, estimate, improve and communicate the level of trust in an AI model or its susceptibility to attacks or any misclassifications.

We define trustworthiness as the verifiable absence of unwanted properties. Explainability and transparency may contribute significantly to measure or estimate the trustworthiness. For such an assessment the full model lifecycle needs to be considered. This assessment should not only be limited to attacks, but include other unwanted properties of AI models that may exist also without the presence of attacks. Thus, for every AI model candidate, the robustness, the susceptibility to attacks, the resistance and resilience, as well as, potential attack detection and response measures need to be systematically analyzed.

We determined that widely accepted metrics, tools, and processes for the assessment are missing, both, regarding the evidence for trustworthiness, as well as, regarding the risk assessment. With the current state of the art, systematic assessments are not yet possible. Thus, it is a long-term goal to define the prerequisites for systematic and recurring risk assessments based on systematic audits of models. These are required for the secure and safe application of AI models in critical and sensitive environments.

### Glossary

(based on BSI report [1])

**Automatic machine learning or AutoML** Approaches to automate the training pipeline setup and training itself.

**Evasion attacks** An attacker plans to change the decision of an AI system during its inference (or operation) phase by subtle modifications of the model input.

**White-box attack** If the attacker has perfect knowledge of the model, the features, and the data.

**Grey- or black-box attack** If the output function is differentiable, which is the case for most of the currently used learning algorithms, a gradient may be computed as a prerequisite for the optimization procedure. However, this is also the case if the attacker has only limited knowledge of the target model, the feature and the data.

**Substitute Model:** The model may be derived either via model stealing attacks or via newly trained models.

**Black-box transfer attacks** Attacks developed for one model can in many cases be transferred to different cAI models without much effort (transferability).

**Black-box query attacks** These attacks use queries to the target model combined with gradient-free optimization methods such as genetic algorithms or bayesian optimization.

**Backdoor poisoning attacks and DoS poisoning attacks** These attacks corrupt parts of the training data in a targeted way.

**Connectionist AI (cAI) systems** cAI systems are trained with machine learning and data.

**Symbolic AI (sAI) systems** These systems may be directly constructed by a human developer.

**Target attack** The attacker is able to control the decision of the AI system.

**Untargeted attack** The attacker just changes the decision of a AI system in an arbitrary way.

## References

1 Christian Berghoff, Battista Biggio, Elisa Brummel, Vasilios Danos, Thomas Doms, Heiko Ehrich, Thorsten Gantevoort, Barbara Hammer, Joachim Iden, Sven Jacob, Heidy Khlaaf, Lars Komrowski, Robert Kröwing, Jan Hendrik Metzen, Matthias Neu, Fabian Petsch, Maximilian Poretschkin, Wojciech Samek, Hendrik Schäbe, Arndt von Twickel, Martin Vechev, and Thomas Wiegand. Current status and future directions. Whitepaper, Federal Office for Information Security, Bonn, Germany, May 2021.

2 Christian Berghoff, Jona Böddinghaus, Vasilios Danos, Gabrielle Davelaar, Thomas Doms, Heiko Ehrich, Alexandru Forrai, Radu Grosu, Ronan Hamon, Henrik Junklewitz, Simon Romanski, Wojciech Samek, Dirk Schlesinger, Jan-Eve Stavesand, Sebastian Steinbach, and Thomas Wiegand. From Principles to Practice. Whitepaper, Federal Office for Information Security, Bonn, Germany, May 2022.

3 Kenneth T. Co, Luis Muñoz-González, Leslie Kanthan, and Emil C. Lupu. Real-time Detection of Practical Universal Adversarial Perturbations, May 2021. arXiv:2105.07334 [cs].

4 Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities, July 2022. arXiv:2207.06236 [cs].

5 Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings, November 2015. arXiv:1511.07528 [cs, stat].

6 Marek Pawlicki, Michał Choraś, and Rafał Kozik. Defending network intrusion detection systems against adversarial evasion attacks. *Future Generation Computer Systems*, 110:148–154, September 2020.

7 John Törnblom and Simin Nadjm-Tehrani. Scaling up Memory-Efficient Formal Verification Tools for Tree Ensembles, May 2021. arXiv:2105.02595 [cs].

## 5.4 AI-Powered Network Defenses

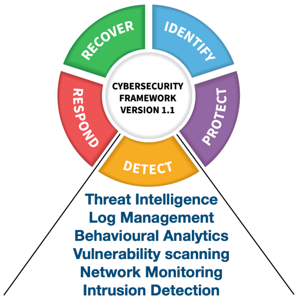
*Vera Rimmer (KU Leuven, BE), Sebastian Böhm (ZITiS München, DE), Georg Carle (TU München – Garching, DE), Marco Caselli (Siemens – München, DE), Nicolas Kourtellis (Telefónica Research – Barcelona, ES), Bettina Schnor (Universität Potsdam, DE), Thomas Schreck (Hochschule München, DE), Max Schrötter (Universität Potsdam, DE), and Robin Sommer (Corelight – Planegg, DE)*

### 5.4.1 Introduction

The landscape of network defense has a long-standing history, yet the comprehensive integration of AI into this domain remains elusive, facing skepticism from both security experts and academics alike. In the face of evolving cyber threats, coupled with recent extraordinary advancements in AI, the need for re-evaluating the role AI may play in network defense becomes increasingly evident. In this working group, network security and applied AI experts joined forces to explore the intricacies of this complex problem space. Our primary focus was on systematically addressing urgent needs in network defenses, spanning across the entire defense pipeline – from pre-detection mechanisms to real-time alert analysis and post-detection response strategies. While there may be many ways to incorporate AI-based solutions, assessing whether involving AI is likely to be beneficial and justified in terms of additional complexity is not trivial. Moreover, the recent breakthroughs in the space of Large Language Models (LLMs) have renewed the ambition of building smart interfaces for human operators and analysts, hence prompting a careful re-evaluation of the potential of AI assistance in network defense.

Motivated by a fresh perspective on the application of AI in network defense, our working group took a practitioner-centric approach. Rather than assuming predefined problems, we aimed to first identify essential needs in the network defense pipeline, moving beyond the conventional alert generation step by an Intrusion Detection System (IDS). This approach challenges the common research paradigm of starting with assumed problem importance but instead promotes a pragmatic assessment of the potential of AI in network defense in alignment with real-world needs. Engaging experts from both network security and AI disciplines, we sought to reshape the discourse, emphasizing the analytical exploration of AI applications based on identified and substantiated needs within the field.

Our overarching goal was to create a roadmap that may guide future investigations with relevant and promising directions for future work. To move towards this objective, we adopted a structured approach that begins by defining a comprehensive network defense pipeline. The pipeline is scoped within the detection segment of the known NIST Cybersecurity Framework 1.1, which in turn we separate into several stages, each characterized by its own set of practical security problems. Upon identifying challenges at each stage of detection, we evaluate these challenges based on the chosen set of criteria: importance for practitioners, the potential for AI-driven solutions, and the research effort required. Our assessment is based on the consensus among the nine members of the working group, encompassing both practitioner and researcher perspectives from the network defense and applied AI domains. This report presents the current result of the analysis according to the developed *assessment framework*, pending a future extension through a comprehensive human study involving more practitioners.



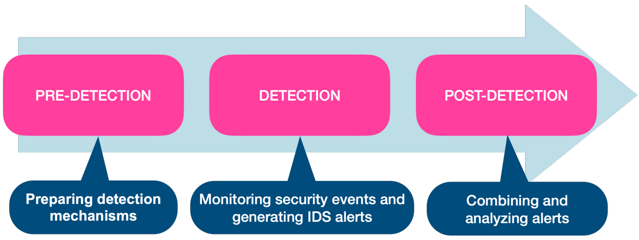

PRE-DETECTION
DETECTION
POST-DETECTION
Preparing detection mechanisms
Monitoring security events and generating IDS alerts
Combining and analyzing alerts

**Figure 1** The chosen scope of the working group.

**Figure 2** The general pipeline of the detection segment in network defense considered in the designed framework.

### 5.4.2 Scope & Methodology

To define the scope of this analysis, we refer to the NIST Cybersecurity Framework (CSF) 1.1 [3], developed by the National Institute of Standards and Technology (NIST) in the United States. This framework provides guidance for organizations to manage and improve their cybersecurity resilience. The core of the framework consists of five functions, each representing a key aspect of cybersecurity (Figure 1):

- **Identify:** Understand and manage cybersecurity risks to systems, assets and data.
- **Protect**: Develop safeguards to ensure the delivery of critical infrastructure services.
- **Detect**: Implement activities to identify the occurrence of a cybersecurity event.
- **Respond**: Take action regarding a detected cybersecurity event.
- **Recover**: Restore and maintain critical infrastructure services in the aftermath of a cybersecurity incident.

In this report, we focus our investigation on the **Detect** function as the one receiving primary academic attention in the field of network defense and yet lacking a comprehensive view in connection to wider real-world needs. We further break down the detection segment into concrete cybersecurity goals and position them along the general detection pipeline – its pre-detection, detection, and post-detection stages, as depicted in Figure 2. This breakdown enables us to formulate a (non-exhaustive) list of 18 network defense challenges based on what we consider to be crucial cybersecurity activities posing day-to-day challenges for practitioners. These challenges constitute the body of our analysis, where we assess the potential of AI to assist in solving them.

### 5.4.3 Assessment Framework

The central idea driving our work was the assessment of how relevant, meaningful or feasible the application of AI may be in view of the critical needs in network defense. The resulting structured framework may serve as a guide for researchers, aiding them in determining their research focus and prioritizing efforts effectively, particularly emphasizing challenges that have higher AI potential, in contrast to more elusive goals.

We utilized a three-fold evaluation to comprehensively assess each challenge – specific issue or task within the detection stage of the network defense pipeline that is under consideration. The three key rating criteria we defined are *Importance*, the *AI Potential*, and *Research*

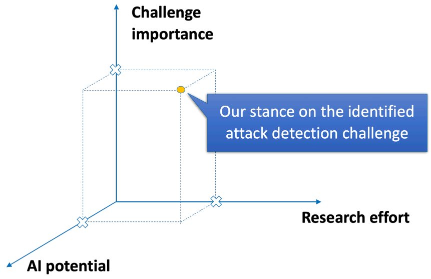


**Figure 3** Illustration of an assessment of a given network defense challenge and its relevant properties on the relative scale.

*Effort*, which are depicted in Figure 3. Below we provide a concise definition for each, along with insights into the main intuition and some underlying aspects:

1. **Importance**: This criterion captures the significance of the challenge, emphasizing its role in the detection workflow. The evaluation is grounded in the practitioners' perspective to ensure practical relevance (which is currently limited to the composition of the working group and is likely to be refined in the follow-up work).
2. **AI Potential**: The assessment of the working group regarding the potential success and significance of data-driven AI methods (machine learning or deep learning) in application to a given challenge. This category encompasses three aspects, which in combination form the total AI potential:
   a. **Applicability**: Examining whether the challenge can be formulated in a data-driven manner for using AI algorithms. Namely, this reflects whether the nature of the challenge lends itself well to the application of AI techniques that rely on data analysis, pattern recognition, and algorithmic decision-making. Problems suitable for a data-driven approach are those where historical information holds valuable clues about what might happen in the future: a much-desired but often limited property in cybersecurity.
   b. **Success Likelihood**: Assessing the probability of solving the problem through a data-driven approach. Even if the problem can be formulated in a data-driven way, there may be serious complications in solving it. For instance, in the case of non-anticipated distributional shifts (due to the lack of representative data or dynamic changes in data at deployment), the trained AI model may lose its relevance too fast.
   c. **Impact**: Gauging the potential impact and effectiveness of leveraging AI to address the identified challenge. This pertains to the expected performance of an AI solution, assuming its applicability and high success likelihood, in comparison to more traditional approaches that do not rely on intelligent automation. The estimated impact may also be influenced by, for instance, the associated costs which may or may not justify the additional complexity of using AI for the task. Interpretability, maintainability and other operational considerations behind an AI-powered solution may also greatly influence its expected impact.
3. **Research Effort**: This category is orthogonal to AI Potential in the sense that it covers the research-related side of addressing the challenges, meaning the effort required to conduct a research study that reliably investigates application of AI to a given task. In our understanding, the Research Effort criterion roughly encompasses two components:
   a. **Existence of Relevant Data**: This aspect considers whether there is ground-truth data within the operational context that holds significance for the challenge at hand. Having access to complete relevant data of high volume and precision is crucial for

training AI algorithms. If the necessary data is readily available within the existing network defense workflows, it reduces the effort required in sourcing and preparing data for AI applications. Note that the current public availability of such corresponding data is out of scope for this analysis (although acquiring it is a significant challenge of its own for the research domain).

b. **Algorithmic Difficulty**: This dimension reflects the technical complexity associated with developing AI algorithms tailored for the specific challenge. Some challenges might demand sophisticated algorithms due to their intricate nature (for instance, those that model multi-dimensional time-series or graph-structured data), while others may be more straightforward. Challenges with higher algorithmic difficulty may necessitate more advanced AI research and development efforts.

This comprehensive framework aims to provide researchers with a holistic perspective on the challenges in network defense, enabling them to prioritize efforts based on the combined assessment of the defined properties. The deliberate design ensures, on the one hand, that researchers can identify not only what challenges are critical but also where the application of AI is most feasible and impactful. On the other hand, the framework may pragmatically illuminate the opportunities for "low-hanging fruit", allowing researchers to address more urgent but less complex needs with a higher likelihood of success, whilst having the assurance of relevance of their study for network defense practitioners.

### 5.4.4 The Current State of the Framework

Here we report on the main outcome of the working group, essentially presenting the snapshot of our understanding of the potential of AI-powered network defenses in the given moment (Table 5). The current state of analysis involves 18 ranked challenges derived from in-depth discussions among academic and industry researchers composing the working group, utilizing expertise in network defense and applied AI. Moving forward, this analysis would be refined to incorporate the perspectives of more security practitioners, offering a more holistic and practically relevant understanding of the challenges at hand. Crucially, this analysis needs to be strengthened with a thorough literature review to incorporate the most recent research advancements in the area, in order to more precisely evaluate the alignment of the current research trends with the highest needs in network defense.

**Table 5** The current state of the assessment framework developed to analyze the potential of AI-powered network defense research directions. These insights emerged from discussions within the working group and are subject to further refinement through an extensive literature review and polling among a larger, more diverse sample of network security practitioners and applied AI experts.

| | | |
|---|---|---|
| | **PRE-DETECTION CHALLENGES** | |
| 1. Merging, correlating, and comparing shared indicators of compromise (IoCs). A shared event is an attack that a team is investigating; there are 100-200k events per day (which are not unique). Several teams may be investigating the same attack, these may be shared simultaneously, but are not automatically merged. | **Importance:** Medium<br>**AI Potential:** Medium/High<br>**Effort:** Low | This activity is non-trivial and demands a lot of human effort. Today certain sources are trusted, but not everything is merged together. Even for humans, merging IoCs in a consistent manner is difficult. AI could scale up the process and improve on human effort. Example: link scarce pieces of information in natural language to other information (labels). |
| 2. Creating new detection rules from IOCs with regard to the diversity and number of future attacks. | **Importance:** Medium/High<br>**AI Potential:** Medium<br>**Effort:** Low | Scaling and expediting this process are paramount, and Large Language Models (LLMs) show promise in this regard, although their usage is complex and challenging. It could be possible to train multiple LLMs (e.g., one for creating the rule and one for checking the rule quality or language). The level of involvement may range from creating a co-pilot for writing rules to simply having a human in the loop. |
| 3.1. Determining the data quality of indicators for Threat Intelligence databases. | **Importance:** Medium/High<br>**AI Potential:** Low<br>**Effort:** High | This can potentially be inferred based on the reputation and reliability of sources. |
| 3.2. Ranking the importance of indicators for Threat Intelligence databases. | **Importance:** High<br>**AI Potential:** Medium<br>**Effort:** High | Given the complexity of merging knowledge from different fields (e.g., threat landscape, infrastructure) and mapping such knowledge to indicators, future AI systems may offer promising solutions. |
| 4. Configuration of existing detection tools, removing dependencies on the administrators. | **Importance:** Medium<br>**AI Potential:** Medium<br>**Effort:** High | Automating the deployment of known hardening measures could significantly optimize the work of security operators. AI may be used to assist in setting parameters for known defenses tailored to a specific environment and objectives. |
| 5. Setting IDS constraints (e.g., workload restriction in the case of a high rules set to prevent overload and exceeding resource limits). | **Importance:** Medium<br>**AI Potential:** High<br>**Effort:** Low/Medium | For a human expert, predicting workload before deploying the system is challenging. Can AI estimate resource demands? |
| 6. Optimization (clean-up) of the detection rule-set. | **Importance:** Medium<br>**AI Potential:** High<br>**Effort:** Medium | Routine maintenance of the rule-set commonly involves manual analysis to discard unused rules and merge rules targeting the same pattern, all to optimize workload and the comprehension of alert generation. Intelligent automation seems feasible. |
| | **DETECTION CHALLENGES** | |
| 7. Detecting variants of known attacks. | **Importance:** Medium<br>**AI Potential:** High<br>**Effort:** Low/Medium | Given the availability of data of previous attacks, we consider the situation in which an AI system can recognize the similarities and detect the same patterns. This scenario is known to be very compelling for AI application in research contexts, but is not the most challenging one, as the traditional methods perform well. |

| | | | |
|---|---|---|---|
| 8. Detecting earlier unknown attacks. | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>Low<br>High | This is inherently challenging for AI methods as they cannot rely on the past to precisely recognize novel future threats within the context of network defense. Throughout the seminar, concerns have been repeatedly expressed about the perceived inadequacy of AI for detecting unknown attacks (e.g., zero days). |
| 9. Generating alert descriptions (e.g., to enhance SOC analyst dashboards). | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>High<br>Low | This opportunity explicitly recognizes the power of AI in interfacing with human experts. The quality of generated textual descriptions depends on the comprehensiveness of alerts, which can be high for traditional methods. |
| 10. Tuning and maintaining empirical thresholds for monitoring. | **Importance:**<br>**AI Potential:**<br>**Effort:** | Low/Medium<br>Low/Medium<br>Unknown | This is a defining component of precise detection, yet targeted empirical approaches are limited. In our discussion, assessing potential and effort was challenging due to a lack of information on the prevalence of anomaly detection in practice. |
| 11. Use Tactics, Techniques and Procedures (TTPs) as IoCs to actively search for malicious activities. | **Importance:**<br>**AI Potential:**<br>**Effort:** | Medium/High<br>High<br>High | The capability of using kill chains instead of simplistic IoCs will allow the detection of a broad range of attacks sharing similar behavior (see point 7). Recognizing kill chains demands the correlation of complex events over time. |
| 12. Attack attribution (e.g., linking alerts to attacker groups and detecting attacks with matching behavioural patterns). | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>High<br>High | Advanced persistent threats (APTs) are mostly obtained from threat intelligence for close monitoring. APT-related incidents have high complexity and priority, and the detection and decision-making process are often highly specific to APTs. The amount of available information and the willingness to extend analysis beyond a few days of activities requires and may highly benefit from AI systems. |
| 13. Run-time optimization of the ruleset. | **Importance:**<br>**AI Potential:**<br>**Effort:** | Low/Medium<br>Low<br>High | Discussed in the context of the difficulty in foreseeing resource consumption of security tools as well as conditions of the target infrastructure (e.g., peak in traffic and heavy computational loads). |
| **POST-DETECTION CHALLENGES** | | | |
| 14. Alert fatigue as for dealing with all the alerts produced by the detection tools (e.g., how do we prioritize? How do we correlate across various sources? How do we diminish false positives?) | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>High<br>Medium/High | AI can learn from human operators how to prioritize alerts and make suggestions (related to 3.2). It might be challenging to obtain enough labeled data and integrate expert knowledge. An additional challenge is that prioritization does not consider the risks (e.g., important devices or other elements of the target environment). |
| 15. Risk assessment of alerts to determine the severity of the case. | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>Low<br>Unknown | Risk assessments imply having information about the infrastructure, and it is currently unclear how to represent this information to an AI system along with other expert knowledge. |

| | | | |
|---|---|---|---|
| 16. Proposing countermeasures to mitigate known attack. | **Importance:**<br>**AI Potential:**<br>**Effort:** | High<br>High<br>Medium/High | AI may provide a collection of suggestions to incident handlers based on information about the incident. The training data for the AI assistant may be composed of playbooks or historical incidents across organizations and how they were handled in the past. The precision of these suggestions might vary: it may be relatively easy to provide relevant general insights, while learning to produce actual custom countermeasures may require more research effort. |
| 17. Labeling adversary's activities (e.g., alerts being triggered by the security tools) according to the MITRE ATT&CK [1] knowledge base. | **Importance:**<br>**AI Potential:**<br>**Effort:** | Medium<br>High<br>Low | This is crucial for all investigations upon detection and for sharing information with other teams. This step goes beyond incident handling. Extracting new attack strategies might be an outcome of this activity, where AI can significantly replace human effort in formulating the description of the attack. Additionally, automating the interpretation of a given attack scenario according to the MITRE ATT&CK knowledge base is valuable, as not all human experts know all techniques to be able to recognize them. |
| 18. Representing and normalizing all data related to detection in a structured and comprehensive form for human analysts. | **Importance:**<br>**AI Potential:**<br>**Effort:** | Medium/High<br>High<br>Low/Medium | One example of a hardly interpretable data structure is the graph-based indicators STIX [2] used to exchange cyber threat intelligence. Commercial vendors write reports about attacks in a complex expert-oriented language, which often obstructs timely and effective handling of network threats. Utilizing AI to format detection-related data in a human-comprehensible manner appears very promising. |

### 5.4.5 Conclusion

The aim of this discussion was to provide a framework for future investigations in AI-powered network defense, leveraging the latest advancements in AI while fostering a more comprehensive understanding of the challenges faced by practitioners. The distinguishing characteristic of the framework is the choice to move beyond the conventional realm of AI-based intrusion detection and explore areas earlier in and further along the pipeline. Our systematic approach facilitates a wide examination of the problem space from both the network security practitioner and research perspectives. It covers practical feasibility, potential impact of applying AI, and the research effort required to address the identified important challenges. By combining the expertise of network security and AI researchers, the framework and its future extensions not only offer a nuanced understanding of challenges but also foster a synergistic environment where innovative solutions can be explored while staying rooted in reality.

#### References

1 Blake E Strom, Andy Applebaum, Doug P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. MITRE ATT&CK: Design and philosophy. In *Technical report.* The MITRE Corporation, 2018.

2 Structured Threat Information Expression (STIX). https://oasis-open.github.io/cti-documentation/stix/intro.html. Accessed: 2023-10-26.

3 The Cybersecurity Framework 1.1 NIST. https://www.nist.gov/cyberframework. Accessed: 2023-10-26.

## 6 World Café and Outlook

The format of the World Café is already described in Section 1. As written there, people split into small groups at individual tables where each table discussed a pre-defined question for 20 minutes before participants moved on in random permutation to another table. The questions asked were as follows:

1. In which of these fields is it most important to make research progress and why: "Security for AI", "AI-based attacks", or "AI for Security"?
2. What steps should we take to keep the activities in our group alive beyond the end of this seminar?
3. What is the title of a paper you would now want to write with some other seminar participants?
4. What is your one key take-away from the seminar?
5. For a future seminar proposal, what (network)security-related topic should we focus on? And whom would we have to (additionally) invite to make it a success?

Due to the small group size and clearly articulated questions, the format proved very effective to collect and distill insights from all the participants, which would not have been possible in a plenary session.

At this point, we will only exemplify a few of the comments provided.

### 6.1 In which of these fields is it most important to make research progress and why: "Security for AI", "AI-based attacks", or "AI for Security"?

On this question, some people suggested that research on "secure AI for security" would be a desirable outcome, i.e., if AI and machine learning mechanisms are applied as a security mechanism, one definitely has to make sure that the mechanism is secure and cannot be attacked itself and that resilience, robustness and reliability would be key properties but also explainability. In this context, a pointer was given to the NIST AI Risk Management Framework[1].

Another emphasis was put on the datasets required for machine learning of security mechanisms. So given the growing importance of machine learning in security, provisioning of good training data becomes key and the research community should put more emphasis on availability of good quality, large, and ideally labeled datasets.

Last, when it comes to AI-based attacks, it seems there is currently a lack of knowledge how such attacks would change the game and to what extend such attacks may be more potent than today's mostly manual attacks.

[1] see https://www.nist.gov/itl/ai-risk-management-framework

## 6.2 What is your one key take-away from the seminar?

Findings from this question overlapped partly with the previous question. Lack of datasets for ML training was mentioned here, too. Others noted that on most questions discussed there is a large agreement among participants, in particular also on the open questions that need addressing and where we yet know too little about.

This was a general sentiment shared by many: in many of the topics we are still at the beginning and need a lot of research to deepen our understanding, maybe with the exception of applying AI for obvious and well investigated tasks like intrusion detection.

One thing that did not come at a surprise: currently there is a strong focus or even hype on security for Large-Language Models (LLMs) but also on how LLMs can help security in various ways.

Besides those discussions, the World Café also contributed many ideas for future follow-up seminars, topics for joint paper initiatives, and the desire to stay in contact and continue discussions online.

And we also came to the conclusion that LLMs could also help come up with catchy paper titles ;-).

## Participants

- Ilies Benhabbour
KAUST – Thuwal, SA
- Sebastian Böhm
ZITiS – München, DE
- Christian Bungartz
Universität Bonn, DE
- Georg Carle
TU München – Garching, DE
- Marco Caselli
Siemens – München, DE
- Hervé Debar
Télécom SudParis, FR
- Sven Dietrich
City University of New York, US
- Daniel Fraunholz
ZITiS – München, DE
- Artur Hermann
Universität Ulm, DE
- Peter Herrmann
NTNU – Trondheim, NO
- Marko Jahnke
BSI – Bonn, DE
- Frank Kargl
Universität Ulm, DE
- Stephan Kleber
Universität Ulm, DE
- Hartmut König
ZITiS – München, DE
- Jan Kohlrausch
DFN-CERT Services GmbH, DE
- Nicolas Kourtellis
Telefónica Research –
Barcelona, ES
- Chethan Krishnamurthy Ramanaik
Universität der Bundeswehr –
München, DE
- Pavel Laskov
Universität Liechtenstein, LI
- Emil C. Lupu
Imperial College London, GB
- Michael Meier
Universität Bonn, DE
- Andreas Mitschele-Thiel
TU Ilmenau, DE
- Simin Nadjm-Tehrani
Linköping University, SE
- Eirini Ntoutsi
Universität der Bundeswehr
München, DE
- Andriy Panchenko
BTU Cottbus, DE
- Delphine Reinhardt
Universität Göttingen, DE
- Konrad Rieck
TU Berlin, DE
- Vera Rimmer
KU Leuven, BE
- Bettina Schnor
Universität Potsdam, DE
- Thomas Schreck
Hochschule München, DE
- Max Schrötter
Universität Potsdam, DE
- Robin Sommer
Corelight – Planegg, DE
- Jessica Steinberger
Hochschule Mannheim, DE