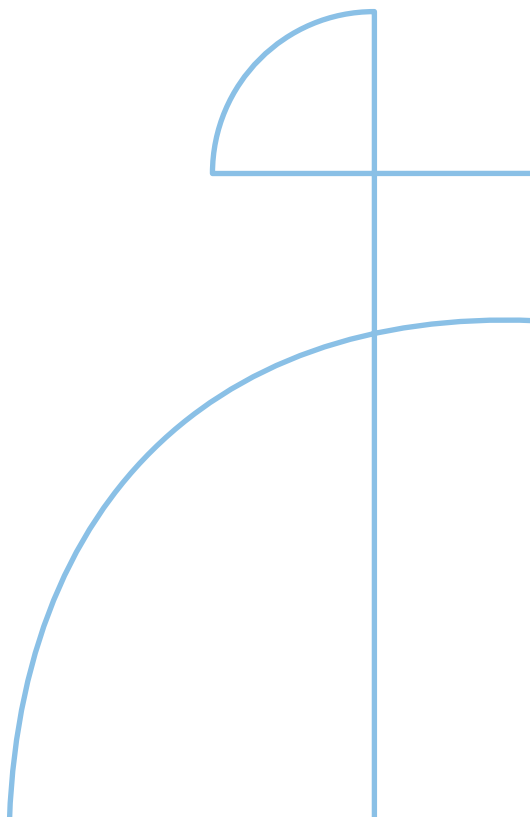

Doctoral Thesis in Computer Science

# Mitigating AI-Enabled Cyber Attacks on Hardware, Software, and System Users

FREDRIK HEIDING

KTH ROYAL INSTITUTE OF TECHNOLOGY



# Mitigating AI-Enabled Cyber Attacks on Hardware, Software, and System Users

FREDRIK HEIDING

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Thursday, October 10, 1:00 p.m. in D31, Lindstedtsväagen 9, Stockholm.

Doctoral Thesis in Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden 2024

TRITA-EECS-AVL-2024:68
ISBN 978-91-8106-040-9

Printed by: Universitetsservice US-AB, Sweden 2024

# Abstract

This doctoral thesis addresses the rapidly evolving landscape of computer security threats posed by advancements in artificial intelligence (AI), particularly large language models (LLMs). We demonstrate how AI can automate and enhance cyberattacks to identify the most pressing dangers and present feasible mitigation strategies. The study is divided into two main branches: attacks targeting hardware and software systems and attacks focusing on system users, such as phishing. The first paper of the thesis identifies research communities within computer security red teaming. We created a Python tool to scrape and analyze 23,459 articles from Scopus's database, highlighting popular communities such as smart grids and attack graphs and providing a comprehensive overview of prominent authors, institutions, communities, and sub-communities. The second paper conducts red teaming assessments of connected devices commonly found in modern households, such as connected vacuum cleaners and door locks. Our experiments demonstrate how easily attackers can exploit different devices and emphasize the need for improved security measures and public awareness. The third paper explores the use of LLMs to generate phishing emails. The findings demonstrate that while human experts still outperform LLMs, a hybrid approach combining human expertise and AI significantly reduces the cost and time requirements to launch phishing attacks while maintaining high success rates. We further analyze the economic aspects of AI-enhanced phishing to show how LLMs affect the attacker's incentive for various phishing use cases. The fourth study evaluates LLMs' potential to automate and enhance cyberattacks on hardware and software systems. We create a framework for evaluating the capability of LLMs to conduct attacks on hardware and software and evaluate the framework by conducting 31 AI-automated cyberattacks on devices from connected households. The results indicate that while LLMs can reduce attack costs, they do not significantly increase the attacks' damage or scalability. We expect this to change with future LLM versions, but the findings present an opportunity for proactive measures to develop benchmarks and defensive tools to control the misuse of LLMs.

# Sammanfattning

Moderna cyberattacker förändras snabbt som följd av framsteg inom artificiell intelligent (AI), särskilt via stora språkmodeller (LLM:er). Vi demonstrerar hur AI kan automatisera och förbättra cyberattacker för att identifiera de största hoten och presenterar strategier för att motverka dem. Studien är uppdelad i två delar: attacker riktade mot hårdvaru- och mjukvarusystem samt attacker fokuserade på systemanvändare, likt phishing. Avhandlingens första artikel identifierar forskningsgrupper inom red teaming. Vi skapade ett Python-verktyg för att hämta och analysera 23,459 artiklar från Scopus databas, vilket gav en översikt av framstående författare, institutioner och utvecklingen av olika grupper och sub-grupper inom forskningsområdet. Avhandlingens andra artikel genomför *red teaming*-tester av uppkopplade enheter från moderna hushåll, exempelvis uppkopplade dammsugare och dörrlås. Våra experiment visar hur lätt angripare kan hitta sårbarheter i enheter och betonar behovet av förbättrade säkerhetsåtgärder och ökad allmän medvetenhet. Den tredje artikeln utforskar användningen av LLMs för att generera phishing-meddelanden. Resultaten visar att mänskliga experter fortfarande presterar bättre än LLMs, men en hybridmetod som kombinerar mänsklig expertis och AI reducerar kostnaderna och tiden som krävs för att lansera nätfiskeattacker och behåller hög kvalitet i meddelandena. Den fjärde studien utvärderar LLM:ers potential att automatisera och förbättra cyberattacker på hårdvaru- och mjukvarusystem. Vi skapar ett ramverk för att utvärdera LLM:ers förmåga att genomföra attacker mot hårdvara och mjukvara och utvärderar ramverket genom att genomföra 31 AI-automatiserade cyberattacker på enheter från uppkopplade hushåll. Resultaten indikerar att LLM:er kan minska attackkostnaderna, men de medför inte en märkvärd ökning av attackernas skada eller skalbarhet. Vi förväntar oss att detta kommer att förändras med framtida LLM-versioner, men resultaten presenterar en möjlighet för proaktiva åtgärder för att utveckla riktmärken och försvarsverktyg för att kontrollera skadlig användning av LLMs.

# Acknowledgements

I am deeply grateful for the diligent, kind, and motivating supervision of Adjunct Professor Ulrik Franke, Professor Pontus Johnson, and Professor Robert Lagerström at KTH Royal Institute of Technology. Ulrik, you fueled my passion for pursuing real knowledge and my curiosity about what good security should entail. You introduced me to the economic aspects of computer security, which has become an important part of my research and enabled me to translate my findings into policy actions and industrial applications. You have provided outstanding feedback and equally fantastic moral and emotional support when I needed it.

Pontus, you provided me, and everyone interested at KTH, with an outstanding hacking education. You helped and encouraged me to build our cherished NSE hacking lab, mentor numerous undergraduate and master students in hacking projects, and assist in demonstrating computer security for the King of Sweden and the Swedish Royal Family. Your trust and mentoring have helped me immensely.

Robert, your excellent leadership style is a key component to why I continued and was able to finish my doctoral studies. You accepted my stubborn personality by letting me pursue my passion projects, even when they deviated from our intended course. Your confidence helped me develop my scientific voice and intuition, often through painful trials and errors. You also provided tremendous emotional support when I needed it and helped me turn problems into opportunities. Lastly, you inspired me to go overseas and continue my research at Harvard, which turned out to be a fantastic adventure that is just getting started.

I am deeply grateful for the overseas supervision and guidance from Stephen Chong, Gordon McKay Professor of Computer Science at Harvard John A. Paulson school of engineering and applied Sciences, Bruce Schneier, Adjunct Lecturer in Public Policy at Harvard Kennedy School, The Honorable Eric Rosenbach, Senior Lecturer at the Harvard Kennedy School and the former Pentagon Chief of Staff and Assistant Secretary of Defense for Homeland Defense and Global Security, and Dr. Arun Vishwanath at Avant Research. All of whom placed immense trust in me and went through

considerable pains to help, mentor, inspire, and teach me. With your help, the past two years have been the most fun, stimulating, and rewarding time of my life so far. You have all inspired me to maintain and increase the naively excited research passion I had when I started my PhD.

I am also deeply grateful for Göran Eriksson and Svante Nygren at Svenska Kraftnät (the Swedish National Grid), who funded my endeavors and provided kind and excellent support along the way. Your hard work finding alternative workarounds when the funding was uncertain allowed me to focus on my research without (excessive) worries. Your commitment to my doctoral studies has been instrumental to their progression and success.

Lastly, I am immensely grateful for the incredible support of my family and friends. You help me survive the tough times and enjoy the good times. I am also deeply grateful for my new friends, coworkers, and housemates in the US, who have made me feel at home here. With the incredible support of all of you, professors, coworkers, friends, and family, my doctoral journey has ended, and my next journey is beginning. Thank you all for helping me get to where I am, and thank you for being a part of my journey. It's been five good years, and the best is yet to come.

Fredrik Heiding
Cambridge, MA, July 2024

# Papers

## Papers included in the thesis

[1] Fredrik Heiding, Sotirios Katsikeas, and Robert Lagerström, *"Research communities in cyber security vulnerability assessments: A comprehensive literature review,"* Computer Science Review, vol. 48, p. 100 551, May 2023, ISSN: 1574-0137. DOI: 10.1016/J.COSREV.2023.100551.

[2] Fredrik Heiding, Emre Süren, Johannes Olegård, and Robert Lagerström, *"Penetration testing of connected households,"* Computers & Security, vol. 126, p. 103 067, Mar. 2023, ISSN: 0167-4048. DOI: 10.1016/J.COSE.2022.103067.

[3] Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter Park. *"Devising and Detecting Phishing Emails Using Large Language Models,"* IEEE Access, vol. 12, pp. 42 131–42 146, 2024, ISSN: 21693536. DOI: 10.1109/ACCESS.2024.3375882.

[4] Fredrik Heiding *"A Framework for Evaluating Large Language Models' Capability to Conduct Cyberattacks,"* Submitted manuscript, 2024.

### Author contributions

In Paper [1], the research direction, conceptualization, and methodology were formed by Heiding and Lagerström. Data collection was done by Katsikeas with the assistance of Heiding, and data extraction, analysis, and synthesis were done by Heiding with the assistance of Katsikeas. The writing was done by Heiding with the assistance of Katsikeas.

In Paper [2], the research direction, conceptualization, and methodology were formed by Heiding and Lagerström. Software-creation and red teaming assessments were conducted by Heiding, Süren, and undergraduate and graduate students supervised by Heiding and Süren, validation was conducted by Olegård, Heiding, Süren, and Lagerström, and the writing was done by Heiding.

In Paper [3], the general research concept was formed by Heiding and Schneier and refined by Vishwanath. Vishwanath provided domain-specific expertise on phishing, Bernstein provided domain-specific expertise on large language models, and Park provided domain-specific expertise on the economics of computer security. The article was mostly authored by Heiding.

In Paper [4], Heiding is currently the sole author, although various experts from academia and industry assisted in forming the general research direction. The framework and writing were created and conducted by Heiding, and undergraduate and graduate students conducted the red teaming assessments under Heiding's supervision. The students were offered co-authorship but chose to remain anonymous.

## Related papers not included in the thesis

[5] Fredrik Heiding, Mohammad-Ali Omer, Andreas Wallström and Robert Lagerström, *"Securing IoT Devices using Geographic and Continuous Login Blocking: A Honeypot Study,"* International Conference on Information Systems Security and Privacy, pp. 424–431, 2020, ISSN: 21844356. DOI: 10.5220/0008954704240431.

[6] Fredrik Heiding and Robert Lagerström, *"Ethical Principles for Designing Responsible Offensive Cyber Security Training,"* IFIP Advances in Information and Communication Technology, vol. 619 IFIP, pp. 21–39, 2021, ISSN: 1868-422X. DOI: 10.1186/S42400-020-00060-8.

[7] Margus Välja, Fredrik Heiding, Ulrik Franke, and Robert Lagerström, *"Automating threat modeling using an ontology framework: Validated with data from critical infrastructures,"* Cybersecurity, vol. 3, no. 1, pp. 1–20, Dec. 2020, ISSN: 25233246. DOI: 10.1007/S10207-022-00633-3.

[8] Emre Süren, Fredrik Heiding, Johannes Olegård, and Robert Lagerström, *"PatrIoT: practical and agile threat research for IoT,"* International Journal of Information Security, vol. 22, no. 1, pp. 213–233, Feb. 2023, ISSN: 16155270. DOI: 10.1007/S10207-022-00633-3.

[9] Philip Wester, Fredrik Heiding and Robert Lagerström, *"Anomaly-based Intrusion Detection using Tree Augmented Naive Bayes,"* Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW,pp. 112–121, 2021, issn: 15417719. DOI: 10.1109/EDOCW52865.2021.00040.

# Contents

# Chapter 1

# Introduction

The threat landscape of cyberattacks has changed over the past decades [10], [11], often resulting in more sophisticated and prevalent attacks [12]–[16]. Since the start of this thesis in 2019, ransomware attacks doubled each year according to some estimates, with a 148% spike during the initial wave of COVID-19[1], Cloudflare reported the largest recorded Distributed Denial-of-Service (DDoS) attack, with 71 million requests per second, a 54% increase from the previous record [2]. Major data breaches [17]–[19] and attacks targeting critical societal functions [20]–[22] are also becoming more common and destructive. According to Sophos's The State of Ransomware reports [23], [24], the mean ransom payment increased from $800,000 in 2022 to $1.5 million in 2023 and almost $4 million in 2024, and the median ransomware payment increased from $400,000 in 2023 to $2 million in 2024. The proliferation of Ransomware-as-a-Service (RaaS) kits makes it easy and cheap to launch the attacks, sometimes retailing for well below $100 on the Darknet[3].

Many complex and intricate cyberattacks start by exploiting human users to access the organization's system. The Sony Pictures hack [11], [25] and the $100m MGM casino hack [26] are two good examples. Some studies claim that well above 70-80% of all cyberattacks use social engineering [27], [28]. Regardless of the number, phishing is a continued nuisance that hurts individuals, governments, and private industries, and many organizations still struggle to incorporate adequate security training routines [29]. FBI's Internet Crime Complaint Center [30], [31] saw an increase in reported phishing attacks from around 115,000 in 2019 to 300,000 in 2023, a growth

[1] https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/cybersecurity/cybersecurity-trends-looking-over-the-horizon

[2] https://blog.cloudflare.com/cloudflare-mitigates-record-breaking-71-million-request-per-second-ddos-attack/

[3] https://www.weforum.org/agenda/2024/02/3-trends-ransomware-2024/

of more than 216%.

Lastly, the increased popularity of the Internet of Things (IoT) introduced new attack vectors and vulnerabilities [32]–[35]. For example, connected household electronics often have poor security standards that can easily be exploited [2], including connected cameras [36]–[38], refrigerators [2], [39], personal tracking devices [40], [41], and home automation systems [42]. As displayed in this thesis, even inexperienced hackers can cause severe damage to household IoT, sometimes resulting in pressing societal consequences [2].

Because of the increased digital threat landscape, computer security has become more important. The field has grown both as a research domain and an industry discipline. There has been a notable increase in the computer security workforce and ecosystem [43], [44]. CompTIA's 2024 State of Cybersecurity report highlights over 660,000 cybersecurity job openings in the United States between May 2022 and April 2023, a 28% increase compared to the same period in 2020 [45]. Global and domestic computer security standards have become more important, and the number and quality of national cybersecurity strategies have increased significantly [46]–[50]. Most Western countries have produced a new or updated national cybersecurity strategy since 2019, and the European Union now mandates and supports each member country to develop an up-to-date national cybersecurity strategy [4]. Academic institutions also intensified their focus on cybersecurity in the past years, such as KTH Royal Institute of Technology launching a master's program dedicated to cybersecurity in 2021 [5] and ETH Zurich and EPFL starting a cybersecurity master's program in 2019[6], and many other worldwide institutions pursuing similar directions. The surge in cyber threats prompted organizations to increase their cybersecurity investments. Security budgets increased with more than 8-10% annually since 2019 [51] and Gartner predicts global cybersecurity spending is expected to reach $214 billion in 2024[7].

Natural language processing capabilities also increased over the last few years due to the rapid development of large language models. Models such as GPT [52] and Claude[8] can generate human-like text, converse coherently, and perform linguistic tasks at superhuman levels. Just within the last year, the size and performance of these models have grown tremendously. Large

[4] https://www.enisa.europa.eu/topics/national-cyber-security-strategies

[5] https://www.kth.se/en/om/nyheter/centrala-nyheter/new-cybersecurity-master-s-programme-is-open-for-applications-1.1118678

[6] https://sciencebusiness.net/network-updates/epfl-and-eth-zurich-introduce-joint-masters-degree-cyber-security

[7] https://www.gartner.com/en/newsroom/press-releases/2023-09-28-gartner-forecasts-global-security-and-risk-management-spending-to-grow-14-percent-in-2024

[8] https://www.anthropic.com/index/introducing-claude

language models excel at creating textual content that *appears* legitimate. With only a few data points about the recipient, LLMs can create content that appears uniquely crafted for that target [3], [53]. Their ability to imitate human writing and reasoning makes LLMs highly effective for deceptive tasks, such as crafting convincing phishing emails. Phishing, like LLMs, aims to use a few data points about the target to create realistic and relevant content [54]. On the other hand, computer security evaluations of hardware and software often suffer from a significant shortage of skilled human labor [55] and could benefit from being outsourced to LLMs. However, as demonstrated in this thesis, technical security evaluations often require context-specific and detailed exploitation procedures, scripts, and function calls [4], [56], [57] that are more difficult for language models to solve and create than phishing emails. Still, LLMs can already reduce the cost of security assessments of software and hardware by reducing the time and knowledge requirements of the tester, and we expect their capability to increase in the coming years [4].

## 1.1 Research purpose

The main research questions of this thesis are to identify and characterize the most pressing dangers of AI-enabled cyberattacks, the most promising defensive use cases, and how to use AI to yield more benefits to defenders than attackers. Questions two and three are split into two branches: one for attacks facing hardware and software systems and one for attacks on system users (phishing). Table 1.1 elaborates on how each paper treats each research question, and Figure 1.1 describes how the papers relate to each other. The first paper of the thesis identifies research communities within computer security red teaming. The other three papers elaborate the research gaps identified by the first study. The second paper conducts manual red teaming assessments of connected devices commonly found in modern households, based on a custom-built framework for security assessments of IoT. Paper three explores the use of large language models to generate phishing emails, and evaluates the effectiveness of spear phishing created by human experts, LLMs, and a combination of human experts and LLMs. Paper four continuous the work in Paper two and evaluates LLMs' potential to automate and enhance cyberattacks on hardware and software systems. Ongoing and post-thesis work uses LLMs to enhance cyber defenders' capability to stop attacks, such as by implementing personalized spam filters to counter spear phishing. Post-thesis work also examines different ways to quantify the monetary cost of cyberattacks and defense strategies, and examines various incentive schemes and policy recommendations for computer security strategies. The research questions are further discussed in Section 5.

| | **P1** | **P2** | **P3** | **P4** |
|---|---|---|---|---|
| RQ 1: What are the most pressing concerns of AI-enabled cyberattacks? | (✓) | (✓) | ✓ | ✓ |
| RQ 2a: What are the most promising solutions to defend against AI-enabled cyberattacks (hardware and software)? | (✓) | (✓) | - | ✓ |
| RQ 2b: What are the most promising solutions to defend against AI-enabled cyberattacks (phishing)? | (✓) | - | ✓ | - |
| RQ 3a: How can AI yield more benefits to defenders than attackers (hardware and software)? | (✓) | (✓) | ✓ | (✓) |
| RQ 3b: How can AI yield more benefits to defenders than attackers (phishing)? | (✓) | - | (✓) | (✓) |

Table 1.1: How each paper, Paper [1] (P1), Paper [2] (P2), Paper [3] (P3), and Paper [4] (P4), answers the research questions of the thesis. ✓ denotes that the research question is directly treated, (✓) denotes that the research question is indirectly treated, and - denotes that the research question was not treated by the paper.

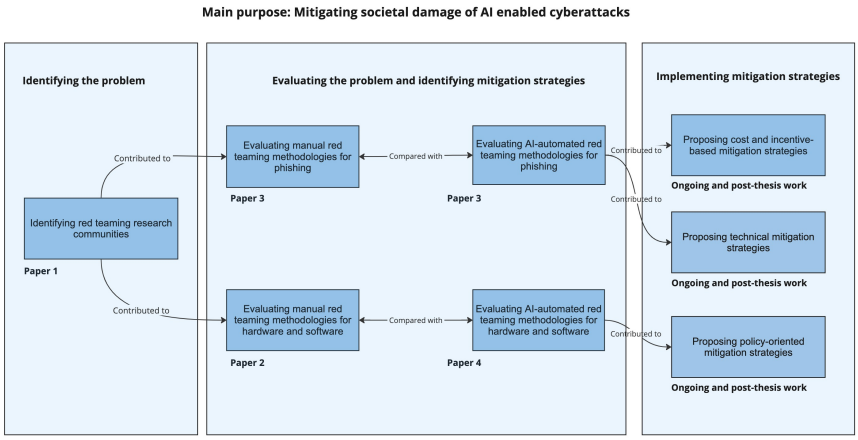



Figure 1.1: An overview of the research purpose of the thesis, all included papers and ongoing work.

# Chapter 2

# Background and Related Work

Computer security aims to maximize a system's security while minimizing operational overhead such as increased runtime, performance requirements, or development costs [58], [59]. To measure security, we often analyze confidentiality (data protection), integrity (data accuracy and authenticity), and availability (ensuring that the data can be accessed by authorized users) [58], [59]. There are many ways to increase a system's security, including cryptography, access controls, intrusion detection and prevention systems (IDS/IPS), and Security awareness training [58], [59]. An alternative to defensive security is to actively test a system's security by simulating the roles of an attacker. This can be done by theoretically proving a system is secure using formal methods, by creating threat models of a system to simulate all possible attack vectors based on the system's characteristics, or by actively testing the system, attacking it in a realistic environment and reporting the discovered vulnerabilities [56], [58]. In this thesis, the terms *red teaming*, *security testing*, *penetration testing*, and *vulnerability assessment* are used interchangeably to describe actively testing a system or its user's security. We primarily use computer security for a broader description of digital security and red teaming when discussing actively testing a system or its user.

## 2.1 Identifying research communities

The first study of the thesis [1] identifies red teaming research communities by scraping all related articles from the Scopus database. The related works show that the domain's popularity has increased significantly over the past decades. Some research communities show especially strong popularity, such as red teaming of Smart Grids and creating attack graphs, as shown in Figure 2.1. The study provided a comprehensive understanding of the research

field and the discovered research gaps guided the focus of the subsequent studies in the thesis. The community detection was conducted in 2022, and we expect that a new version of the study will discover more articles and, potentially, a novel research community on LLM-based red teaming.

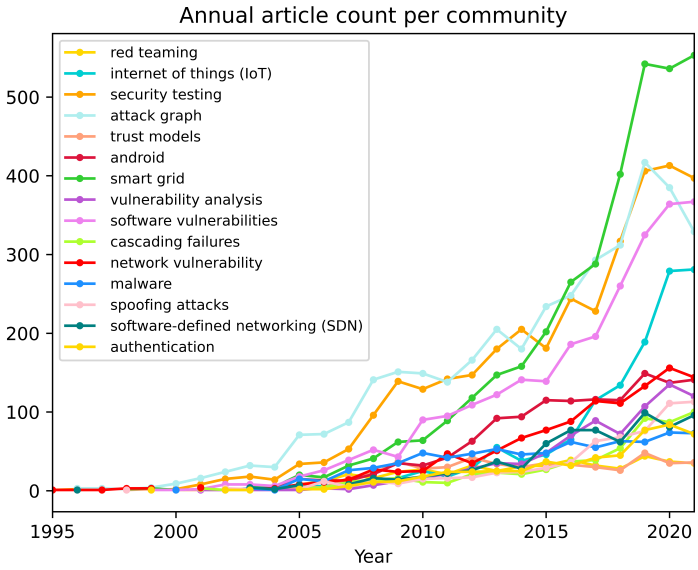



Figure 2.1: Growth of the communities over time.

The community detection was inspired by [60], which detects computer security research communities (a broader perspective than our red teaming community analysis). Except for [60], community detection within computer security is sparse. There exist some manual literature reviews on red teaming [61]–[63], but they are constrained by a smaller scope than automated analysis, and often treat less than 50 papers compared to more than 20,000 included in our analysis. Other literature reviews target a specific subset of red teaming, such as testing web-based applications [64], [65], cloud devices [66], or AI-enhanced testing [67]. Other reviews target specific attack types, such as cross-site scripting (XSS) [68]. The manual literature reviews often adhere to the guidelines proposed by [69], [70], or both. To complement existing research, we provide a holistic overview of red teaming research communities by scraping 23,459 related articles from the Scopus database [1].

## 2.2 Red teaming of hardware and software

Red teaming offers organizations a controlled environment to test their system's and users' defensive capacity without incurring any real risk [56], [57]. Different red teaming methodologies and philosophies exist, but the general process is often similar and includes information gathering (selecting a target), threat modeling and vulnerability scanning (identifying potential weaknesses based on the target's software, hardware, and configurations), and exploitation (testing the identified weaknesses in a controlled environment) [57], [71]–[74]. Some methodologies treat specific use cases such as how to test for DoS attack [75], XS-Leak Detection attacks [76], or attacks toward IoT devices [77]. Other methodologies treat specific tests of a range of devices or a single device, such as red teaming a connected camera [78], smart home voice assistant [79], or various devices commonly located in connected households [1], [80]–[85].

Several related studies mention that more work on IoT and smart home security assessments would be useful [33], [83], [86], [87]. Paper [2] of this thesis addresses this by conducting a comprehensive red teaming analysis of IoT devices from connected home environments. We captured breadth and depth by conducting thorough penetration tests on 22 devices and revealed 17 new Common Vulnerabilities and Exposures (CVEs), as described in Section 4.1. To the best of our knowledge, it is the most comprehensive study to date, capturing a wide range of devices and a substantial level of detail for each device.

Several studies investigate the use of AI to enhance red teaming [88], [89], and some target large language models [90]–[93]. LLMs show promising signs of usefulness in several areas, such as providing better analysis of the long-term behavior and characteristics of networks [94], extracting static information from diverse sources [95], and fuzzing-based security assessments [96], [97]. However, no widely accepted benchmark has been established to evaluate LLMs' capability to perform cyberattacks on hardware and software systems. Much recent work has been done on various other evaluation benchmarks for large language models. For example, measuring knowledge acquired during pretraining (MMLU [98]), graduate-level expert questions (GPQA [99]), diverse tasks humans may encounter in their daily routines (Webarena [100]), solving GitHub issues (Swe-bench [101]), and Evaluating various realistic tasks such as finding information on Wikipedia (METR [102]). By combining our research on how LLMs can enhance cyberattacks with an evaluation framework to assess the usefulness of LLM-enhanced attacks, we hope to further the field by providing a standardized method for measuring and comparing AI-driven red teaming of hardware and software systems.

## 2.3 Phishing campaigns

Almost 20 years ago, Dhamija et al. explained *"Why phishing works"* [103], highlighting that phishing attacks exploit inherent human psychological and behavioral weaknesses. People rely heavily on visual cues and other heuristics when assessing credibility rather than rationally analyzing content [54]. Phishing attacks often exploit routine behaviors to capture users when they are less likely to carefully assess the material [54], [104]. Unfortunately, phishing still works. Human nature is slow to change, and the same innate psychological tendencies that make us vulnerable, like favoring trust over skepticism and prioritizing urgency, are deeply ingrained in our nature [54]. Organizations often spend substantial resources to train their employees to avoid phishing, which has become one of the most persistent cybersecurity threats to organizations, governments, and institutes around the world [105], [106]. Up to this point, it has been easy to launch phishing attacks, but many emails have had poor quality, lacking coherent reasoning, a trustworthy presentation, or correct language and grammar. More sophisticated phishing attacks, so-called spear phishing, require more deliberate reconnaissance and creation time. Manual methods for phishing exist and have displayed impressive results [54], but they are time-consuming and expensive. The V-Triad, presented in [54], is a human model for manually creating phishing emails and deceptive content that can bypass a user's suspicion filter. Unlike LLMs, the V-Triad is manually created based on highly targeted and specific data (real-world phishing emails and deceptive content), resulting in a specialized model with a targeted use case. LLMs can create phishing emails automatically, while the V-Triad is a guide to assist us when manually creating phishing emails. The V-Triad is adapted to a recipient's cyber risk beliefs, which describe how accurately we perceive digital risks and are affected by cognitive heuristics and biases. By exploiting these beliefs, the V-Triad lets an attacker create action triggers (such as a phishing email with a link) that are unlikely to make the recipient suspicious. Users with bad self-regulation (likelihood of developing strong media habits) are especially susceptible [54].

The V-Triad comprises three parts: Credibility, Compatibility (relevancy), and Customizability [54]. *Credibility* treats how the content of the email is perceived. If the email appears legitimate to the recipient, it is credible. Common ways to increase an email's credibility are to include a well-known brand name, sender, or graphical profile. *Compatibility* refers to how relevant an email is to the recipient. For example, an email claiming to be from the recipient's coworker or employer, or an email offering a gift card from a store where the target is a frequent customer. Compatibility often exploits a certain timing or affiliation. *Customizability* treats whether a website or

email behaves as the recipient expects it to behave. It is more relevant for websites but also affects emails. For example, two-factor authentication buttons in emails often yield a slight shadow when pressed and should lead to an appropriate website that behaves as the recipient expects it to behave.

[53] presents a similar cognitive framework for creating phishing emails based on two cognitive characteristics that overlap with the V-Triad's credibility metric: believability and persuasiveness. Believability refers to crafting phishing messages that resemble the communication style of the impersonated organizations, such as the look and feel of emails, including logos, and adopting a formal writing style. Persuasiveness regards techniques that exploit fundamental vulnerabilities of human cognition, such as the six influence principles outlined in [107] (Reciprocity, Consistency, Social Proof, Authority, Liking, Scarcity).

Although large language models have only gained widespread attention in recent years, many studies have already explored their potential for generating and detecting phishing emails. Given their ability to produce increasingly human-like text, generative language models are often perceived to be well suited for persuasion and deception [108]–[113]. However, the same models also show promising signs of being able to counter deception by improving phishing detection [114]–[117]. [114] uses GPT-3.5 and GPT-4 to detect phishing sites, validates the result on a dataset, and receives a precision of 98.3% and a recall of 98.4%. [115] proposes two language models adapted to a custom-made dataset containing 725k emails (made by merging an existing collection of legitimate and phishing emails). [116] and [117] propose veritable pre-trained deep transformer network models for phishing URL detection, and the latter performs additional domain-specific pre-training tasks. None of the related studies investigate how to detect the intention of phishing emails using LLMs and create personalized user recommendations for validating the phishing attacks. Most related work focuses on creating phishing emails without validating them in a real-world context. By implementing and evaluating our phishing emails on live participants, we aim to further the field by providing success rates of the phishing emails that can be used as a benchmark to compare the capacity of future model versions and model types.

# Chapter 3

# Research Design

The first paper of the thesis [1] followed the methodological guidelines for systematic literature reviews proposed in [69], [70]. The remaining papers followed the research principles outlined in [118]. Figure 3.1 displays an overview of how each paper corresponds to the research philosophy of [118].

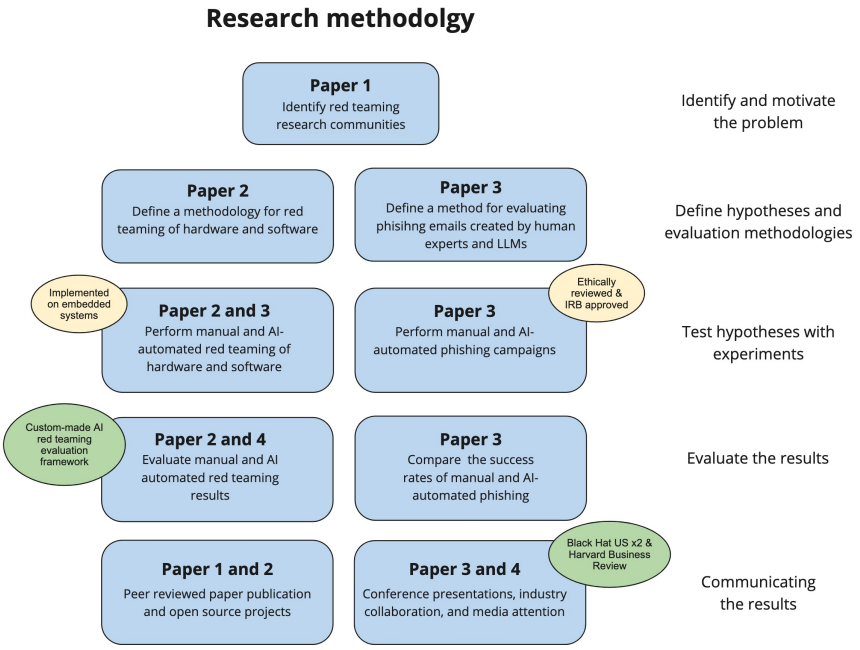



Figure 3.1: An overview of how the included papers correspond to the research methodology presented in [118].

## 3.1 Identifying existing research communities

The literature review followed five stages based on the guidelines from [69], [70]. It aimed to provide a critical and evaluative account of the published red teaming research by summarizing, synthesizing, and analyzing it. The review helped us identify research gaps and revealed similarities, consistencies, and controversies in previous research. The process, further described in Figure 3.2, was to identify papers to analyze (with inclusion and exclusion criteria), collect papers and extract information from them, and interpret the results (community detection). The literature review investigated four primary research questions *"How much activity has there been in cybersecurity red teaming since 1975?", "What research topics are being investigated within cybersecurity red teaming?", "Who is active in the research area" and "what are the active contributors' relations?".*

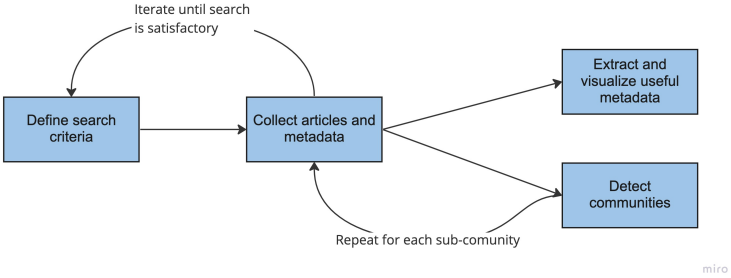



Figure 3.2: An overview of the used method, from creating the search query to extracting metadata from the sub-communities.

### 3.1.1 Identify papers to analyze

The three authors manually wrote down all phrases they thought represented red teaming and security assessments. A consensus was required to add new phrases to the search query. The initial list included: *"Vuln* research", "Vuln* assess*", "Vuln* analysis", "Ethical hack*", "Security test*", "Offensive security*", "Penetration test*", "Pentest*", and "Red team*".* The authors' acceptance was seen as a good initial validation that the search phrases were relevant, as all authors are computer security researchers focusing on security assessments. To further ensure the quality of the search phrases, the following steps were taken:

1. Search the Scopus database using a search query that includes all the above search phrases.

2. Analyze the most influential keywords from all articles found by the scraping [1].
3. For each discovered keyword that was not used in the search query, the keyword was searched for in Scopus (with subject restriction "computer science").
   (a) The top 25 articles (using the Scopus sorting option "most relevant") for the given keyword were manually screened.
   (b) If a majority (+75%) of the articles treated penetration testing or some aspect of vulnerability research as its main research contribution, the topic was included in the final query.

The most influential keywords from the articles obtained by the first query were: *'red teaming' (492), 'vulnerability analysis' (289), 'intrusion detection systems' (215), 'it security' (186), 'security solutions' (169), 'computer security applications' (160) 'risk management processes' (160) 'security chain' (160) security technologies' (160) 'security usability' (160), 'security applications' (160), 'artificial immune systems' (135), 'complex adaptive systems' (119), 'cyber security' (108), 'privacy' (107), 'software engineers' (104), 'political science' (104),'cybersecurity' (100), 'security' (99),* and *'vulnerability assessments' (97)*. The number following the keyword is the word's number of occurrences across the scraped articles.

In some cases, a search phrase from the original query occurred as one of the most influential keywords from the articles found by the query, for example, "red teaming" and "vulnerability analysis." In these cases, the keyword was analyzed in the same way as all other keywords. If it displayed less than 75% relevant articles from the search, the keyword was excluded as a search phrase from the final query. None of the keywords used in the initial search were excluded.

### 3.1.2 Collect papers and extract data

After validating the keywords, a logical disjunction was used to create the final query:
(`‘‘Penetration test*’’ OR ‘‘Ethical hack*’’ OR...`)
The query was used in the Scopus database and restricted to only include articles from the subject area *"computer science"*. The raw query can be found at [2]. The search resulted in 23,459 articles. For each article, author, and affiliation, metadata was collected as follows:

**Article metadata:** EID (unique academic work identifier in Scopus), authors, title, source, keywords, publication date, and references.

[1] The most influential keywords are those used by the most cited articles, not the most used keywords.

[2] https://bit.ly/3sMAJh5

**Author metadata:** Scopus author ID, surname, given name, and affiliation.

**Affiliation metadata:** Scopus affiliation ID, name, and country.

We created a Python application that used Scopus's API to collect the articles and analyze their relevant meta-data. Scopus was selected as the data source because of its comprehensive content, which includes data from IEEE and ACM publications (among others). The Web of Science has an equally comprehensive database, but Scopus offers a more flexible and easily accessed API.

### 3.1.3 Interpret the results:

The collected data was used to generate a citation graph for the authors (Figure 3.3), in which all the authors are interlinked according to their citations. Each node in the graph represents an author. Undirected edges between authors indicate that at least one author has cited the other at least once. The size of the node depends on how many citations the author has. To reduce the size of the graph, only authors with more than 12 citations globally are presented.

The author graph represents relations between people. Much research has been conducted on such graphs, particularly on community detection. One of the best-performing algorithms for community detection in large graphs is the Louvain method [119]. Louvain aims to find a graph partition that maximizes modularity, a scalar value between -1 and 1 that "measures the density of links inside communities as compared to links between communities," defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where $A_{ij}$ represents the weight of the edge between vertices $i$ and $j$, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex $i$, $c_i$ is the community to which vertex $i$ is assigned, the $\delta$-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \frac{1}{2} \sum_{i,j} A_{ij}$.

Opting for an unweighted network, we used an open-source Python implementation of this algorithm, developed by Thomas Aynaud [3]. An author's research focus can relate to multiple communities, and an author may publish articles on different topics. Thus, some authors contribute to more than one research community. Such authors will be placed in the community to which they are most tightly connected. Furthermore, such authors will strengthen the relations between the involved communities.

[3] https://github.com/taynaud

Figure 3.3: Author graph of the research communities in vulnerability assessments and ethical hacking research. Each node represents an author, colors denote research communities, and the size of each node is proportional to the number of citations.

A graph was created using the spring layout algorithm of the Python library NetworkX [4]. In the graph (Figure 3.4, nodes correspond to communities and node size to community size. Edge width and node distance depend on inter-community coupling. The name of each community is derived from the most influential keyword in the community, i.e., the top keyword from the articles with the most citations. The name was manually changed in a few cases to better reflect the community's topic. In total, 15 research communities were detected, as displayed in Figure 3.4.

The above process was repeated for each discovered community to provide each community with several sub-communities. In total, 126 sub-communities were discovered. For each sub-community, we identified the most common keywords, most-cited authors, and top publication fora. Sub-communities with fewer than 75 articles were excluded.

[4]https://networkx.github.io

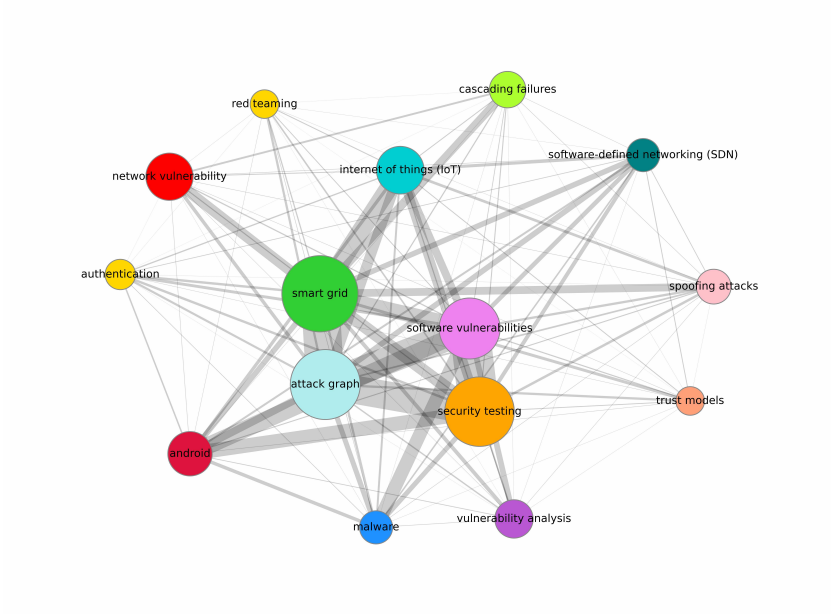



Figure 3.4: The vulnerability research community graph generated from the analyzed data

## 3.2 Pursuing discovered research gaps

After categorizing the related literature, the study direction was chosen based on best practices from [118]. For each paper, a hypothesis was defined, followed by observing phenomena, analyzing the collected evidence, and assessing the study's impact and validity. Table 3.1 shows an overview of the hypothesis, observed phenomena, and evidence. Table 5.1 motivates the originality and impact of each paper, and Table 5.2 motivates the validity of each paper's research process.

The hypotheses for papers one (*it is possible to find meaningful research communities within red teaming*), two (*retailing embedded devices for connected households contain severe vulnerabilities that can harm their users and endanger society*), and four (*LLMs are not yet capable of significantly enhancing cyberattacks on hardware and software systems but can reduce the cost of launching attacks*) were largely correct. The hypothesis for paper three (*LLMs can create phishing emails that perform on par with human experts*) was too optimistic about LLMs' capabilities to outperform human phishing experts. The hypothesis for paper two was developed by considering

| | **Paper 1** | **Paper 2** | **Paper 3** | **Paper 4** |
|---|---|---|---|---|
| Hypothesis | It is possible to find meaningful research communities within red teaming. | Retailing embedded devices for connected households contain severe vulnerabilities that can harm users and endanger society. | LLMs can create phishing emails that perform on par with human experts. | LLMs are not yet capable of significantly enhancing cyberattacks on hardware and software systems but can reduce the cost of launching attacks. |
| Phenomenon | The Louvain algorithm creates meaningful clusters of related authors and articles. | We can apply specialized red teaming methods for IoT to easier observe security flaws, such as a susceptibility to denial of service or man-in-the-middle attacks. | The large language models created personalized spear phishing emails that appeared realistic based on a few data points about the target. | LLMs did not increase cyberattacks' damage to the target system or society but could reduce their cost, especially by making attacks of mid-level complexity cheaper. |
| Evidence | The data scraping and community detection resulted in 15 identified red teaming research communities based on 23,459 articles. | Our red teaming assessments discovered 17 vulnerabilities in 22 devices, some of which were ranked as critical (up to 9.8/10) by the National Vulnerability Database. | The LLM-generated emails received a click-through rate between 30-44%, human expert-generated emails between 69-79%, and hybrid emails between 43-81%. | LLM-enhanced cyberattacks of hardware and software never increased the damage to the device or society at large but reduced the cost by 59% and increased scalability in 50% of the attacks. |

Table 3.1: The hypothesis, observed phenomenon, and evidence for each paper.

potential vulnerabilities in retailing embedded devices. The work conducted in the thesis was derived from many iterative red teaming assessments of everyday embedded devices from connected households and surrounding products such as drones and vehicles, tested in the NSE hacking lab at KTH [5]. The hacking lab was partially created, operated, and maintained through the work conducted in this doctoral thesis. Initial tests and minor security assessments indicated that everyday IoT appeared to have significant and widespread security flaws.

The third paper's hypotheses tested the vulnerability flaws of system users rather than systems, specifically analyzing how the increased potential of AI systems can facilitate attackers' exploiting users through phishing attacks. The hypothesis developed following meetings with industry actors who highlighted the problem of securing system users and the rapid spread of large language models in late 2022. Large language models excel at creating realistic textual content, making them well-suited for phishing. The hypothesis was deemed important and interesting as exploitation through phishing was already a pressing problem and because LLMs showed promising signs of worsening the problem by reducing attackers' time and cost requirements.

The fourth study's hypothesis was formulated after the third study's hypothesis was evaluated. The third study showed that large language models significantly help attackers create cheaper and more scalable spear phishing emails. It was natural to continue the research by investigating how LLMs support attackers exploiting hardware and software. It is more difficult to measure the success of hardware and software exploitation than phishing, as the attack surface and success criteria vary across attacks and devices. Therefore, we created an evaluation framework for quantifying LLMs' usefulness for hardware and software exploitation. Then, we tested the framework by repeating the hacks from paper [2] and similar new hacks using LLMs. The hypothesis was that the framework would facilitate a comparison between LLM-assisted and manual red teaming by quantifying the usefulness of LLM assistance and that the usefulness would currently be low.

Each hypothesis was evaluated and defended according to the best practices proposed in [118], including assessing the study's contribution by quantifying its originality and validity. Originality defines whether the ideas presented in the study are novel, relevant, and significant. Ideas can appear obvious while still being original, and excellent ideas are often obvious in retrospect [118]. Still, they should add value to the scientific community's body of knowledge. When assessing the hypothesis and expected result of the papers, we repeatedly asked what the impact would be if the paper were widely read by the community and society at large. The originality

[5] https://nse.digital/

and importance of the studies were assisted by close industry collaborations, including collaborations and frequent meetings with leading AI institutes such as OpenAI [6], Anthropic[7], RAND[8], the Future of Life Institute at MIT[9], the Future of Humanity Institute at Oxford[10], and the UK's Centre for the Governance of AI[11]. We evaluated each paper based on four impact criteria, as shown in Section 5 and in Table 5.1.

In the context of paper [1], repeatability and reproducibility are deemed very likely, as the tool and Scopus's database offer a fully controlled environment. Replicability is likely, such as running a modified software version on the Web of Science's database. To further increase the validity of the hypothesis, we experimented extensively with different variables, such as modifying the inclusion size threshold for the research communities (75 articles in the final version). Different cluster sizes affected the result, but the differences were variations to the same trend rather than completely new insights.

Paper [2] uses experiments rather than modeling, which makes it harder to control all background variables. Still, repeatability is expected to be very likely, such as performing the same red teaming assessments on the same smart door lock and observing the same results. Reproducibility is also expected to be very likely, such as performing the same red teaming assessments but with a new version of the red teaming tool, analyzing the same smart door lock, and observing the same results. Replicability is possible, such as performing the same red teaming assessments on a smart door lock from another manufacturer and observing the same results, but less obvious due to the context-dependent nature of many cyberattacks. The paper actively builds upon previous work and existing ideas, such as the IoT red teaming framework proposed in our earlier work [8]. Rival hypotheses include that the devices are not susceptible to vulnerabilities or that no real danger exists as attackers are unlikely to exploit the devices. Both statements are falsified by ours and related research [78]–[82].

Paper [3] also uses experiments, and repeatability is reasonable but not guaranteed, as we deal with human subjects. As with other experiments on human subjects, phishing research is difficult to fully control [54], [120]. It is easy to imagine that a user who receives the same email two times might respond differently based on background factors such as the person's emotional state, which could be affected by unforeseen events occurring to the participant [27], [54], [105]. Reproducibility suffers from technical con-

[6] https://openai.com/
[7] https://www.anthropic.com/
[8] https://www.rand.org/
[9] https://futureoflife.org/
[10] https://www.fhi.ox.ac.uk/
[11] https://www.governance.ai/

cerns as different large language models sometimes display output variance. Replicability is reasonable and relatively likely as the study's population is expected to generalize well, but it still suffers from concerns similar to repeatability. Ongoing phishing studies further increase the validity of the phishing security research stream that started during this doctoral thesis.

The hypothesis of paper [4] treats the success and usefulness of AI-automated red teaming assessments of IoT. The study's methodology and hypothesis are similar to those of the manual red teaming study (Paper [2]), and the study's validity is expected to be similar, although somewhat more uncertain due to the occasional volatility of the output from language models. The hypothesis also concerns the scoring of LLMs' usefulness in enhancing cyberattacks, which will depend on the quality of the created framework. The better the framework, the more reproducible, replicable, and repeatable scoring. The framework is deemed solid, but some variance would likely affect scoring by different testers due to subjective human differences.

# Chapter 4

# Results and Contribution

The projects in the thesis have led to three primary contributions in addition to the literature review outlining the red teaming research communities presented in Section 3.1.

First, the research has provided general and specific knowledge regarding the vulnerabilities of commercial IoT, especially devices related to connected households. We have highlighted that cyberattacks targeting these devices are inexpensive to execute and often inflict significant damage to the device, posing risks to users and society. Our work raised public awareness, resulting in discussions with Swedish insurance companies on how to offer cyber protection to their customers, interviews in Swedish National Television [121], and personal meetings and interest from decision-makers and dignitaries, including the King of Sweden, the Swedish Royal Family, and the Swedish European Commissioner. The raised awareness has underscored the inherent dangers of insecure IoT devices, prompting the industry to recognize and address specific vulnerabilities. The awareness has already led vendors to implement device-specific patches and security updates. Moving forward, it is crucial to continue efforts in vulnerability detection for IoT. Additionally, there should be improved incentives for manufacturers to build secure products, exemplified by initiatives such as the secure-by-design act promoted by the U.S. Cybersecurity and Infrastructure Security Agency (CISA) [122].

Second, the research has demonstrated that large language models are proficient at deceiving individuals and can generate personalized and effective spear phishing attacks at a low cost, already performing on par with human experts. The results received widespread attention in the research community and society at large, sparking interest from leading AI research centers, including OpenAI, Anthropic, RAND, and the Center for the Governance of AI in the UK, as well as security venues such as Black Hat [123], DEF CON,

BSides [1] and leading industry outlets like Harvard Business Review [124]. The problem highlighted by our findings [3] is intensified as future versions of the language models will likely be even more adept at deceiving people, resulting in a vicious cycle of continuously cheaper and more successful attacks. In light of this, it is important to conduct further research on utilizing LLMs to train and assist users in avoiding online scams. One promising approach, and part of our current research, is the development of personalized spam filters, as further discussed in Section 5.

Third, we show that LLMs currently lack the widespread capability to autonomously launch sophisticated cyberattacks on hardware and software systems or provide significant assistance to attackers exploiting hardware or software. This presents a unique opportunity for defenders. Proactive measures can be taken to create benchmarks and develop defensive tools to ensure that LLMs' potential to assist in hacking remains controlled and favors defenders. Going forward, it would be wise to enforce policies that require new language model versions to be rigorously tested against established benchmarks against hardware and software exploitation. The contributions highlight the importance of ongoing research, public and industry awareness, and the development of proactive measures to mitigate the risks associated with emerging technologies in cybersecurity. Table 4.1 shows an overview of the contributions derived from the thesis, and Table 4.2 shows an overview of the result types observed in the thesis.

The thesis papers primarily generate results in the form of experimental insights and artifacts. The first paper creates a software tool designed to scrape the Scopus database and identify research communities within a field of study. The study also includes experimental results from implementing this software tool to discover 16 academic red teaming research communities.

The second paper provides experimental results from evaluating the security standards of the 22 embedded devices from connected households. The evaluation was conducted using a security framework proposed in earlier research and identified 17 CVEs and 52 additional weaknesses, several of which were ranked as critical, with scores reaching up to 9.8 out of 10 by the National Vulnerability Database. The findings highlight the significant security risks associated with everyday household devices and the need for improved security measures.

The third paper presents experimental results analyzing the effectiveness of phishing emails generated by large language models, human expert methods (the V-Triad), and a hybrid approach combining both (LLMs + V-Triad). The study measures success by the rate of users pressing email links. The results indicate that while human experts generally outperformed LLMs, the hybrid method often performed on par with human experts and

[1] `https://bsideslv.org/schedule`

significantly reduced the attack cost. The research underscores the potential of combining LLMs with expert knowledge to enhance the effectiveness and efficiency of phishing attacks.

The fourth paper introduces an artifact in the form of an evaluation framework for assessing the usefulness of AI-enabled red teaming on hardware and software systems. The study also includes experimental results from red teaming assessments conducted with the framework, which found that LLMs can reduce the cost of cyberattacks on hardware and software, particularly for mid-level attacks. However, the presence of LLMs did not significantly affect the damage caused to the devices or society.

The sections below describe the results in more detail. Section 4.1 presents the success rate of manual and AI-enabled red teaming towards hardware and software systems, including result comparisons between different device types (such as connected cameras and connected door locks) and attack types (such as man in the middle and denial of service attacks). Section 4.2 presents the success rates of manual and AI-enabled red teaming of system users through phishing attacks, including information on how AI can mitigate and detect phishing attacks.

| | **Finding** | **Implication** | **Proposed path forward** |
|---|---|---|---|
| Contribution 1 | General knowledge about how cyberattacks targeting commercial IoT are cheap to launch, effective, and can cause severe damage to the device's user and society, and specific knowledge about device vulnerabilities. | Public awareness about the dangers of insecure and widespread IoT and industry awareness of specific vulnerabilities, resulting in device-specific patches and security updates by the vendors. | Continued work on vulnerability detection of IoT and improved incentives for building secure products, such as the secure-by-design[a] act. |
| Contribution 2 | LLMs significantly reduce the cost of deceiving people and can cheaply create personalized phishing attacks that perform well, especially when guided by minor human input. | Raised awareness about the deceptive capabilities of LLMs and how future LLM versions are likely to be even better at deceiving humans, resulting in cheaper and more successful attacks. | More research on using LLMs to train and help users avoid online scams, such as the personalized spam filters proposed by our current work and discussed in Section 5. |
| Contribution 3 | LLMs are not yet widely capable of autonomously launching sophisticated cyberattacks on hardware and software systems. | Defenders have a unique opportunity to proactively create benchmarks and defensive tools to ensure that LLMs' capacity to hack hardware and software remains controlled. | Enforcing policies to check all new language model versions against established benchmarks to evaluate the model's capability to assist cyberattackers or autonomously launch cyberattacks on hardware and software systems. |

Table 4.1: The main contributions derived from the thesis.

[a]https://www.cisa.gov/securebydesign

| | Paper 1 | Paper 2 | Paper 3 | Paper 4 |
|---|---|---|---|---|
| Result type | **Artifact** in the form of a software tool that can scrape Scopus's database and identify research communities in a given field of study & **experimental results** from implementing the software tool. | **Experimental results** showing the security standard of 22 embedded devices from connected households by implementing the IoT red teaming framework proposed in earlier research. | **Experimental results** showing the success rate, as defined by pressing email links, and the relative performance difference of phishing emails generated by large language models, human expert methods (the V-Triad), and a hybrid version (LLMs + V-Triad) | **Artifact** in the form of the evaluation framework for assessing the usefulness of AI-enabled red teaming on hardware and software & **experimental results** derived from performing red team assessments based on the framework. |
| Result | 15 discovered academic red teaming research communities, the largest of which were smart grids, attack graphs, security testing, and software vulnerabilities. | 17 identified CVEs and 52 additional weaknesses in the 22 devices, of which several were ranked critical (up to 9.8/10) by NVD[a]. | Human experts outperformed LLMs, but the hybrid version (LLMs + human expert methods) performed well and significantly reduced the attack cost. | LLMs can reduce the cost of cyberattacks on hardware and software, especially towards mid-level attacks, but they do not significantly affect the attacks' damage to the device or society. |

Table 4.2: The results and result types produced by the thesis.

[a]The National Vulnerability Database

## 4.1 Red teaming of hardware and software

After analyzing how red teaming is treated by various academic communities, we conducted red teaming assessments of 22 embedded devices from connected households. The devices were tested using a specialized red teaming method adapted for embedded devices and IoT, as described in our earlier work [8]. The devices were retailed and used worldwide, and the tests revealed severe vulnerabilities that could introduce harm to citizens. For example, CVE-2020-12838 [2] and CVE-2020-12839 [3] gives root access to a smart garage door, allowing an attacker to open the door and enter the facility. The device manufacturers were contacted following Google's Project Zero guidelines [125] for responsible disclosure to ensure that all vulnerabilities were mitigated before we published our study. We always offered free consultation on how the manufacturers could remove or mitigate the discovered vulnerabilities. Some companies were positive about our research as it made their products more secure, while others were hesitant and wanted to prioritize tasks other than security and saw our suggestions as a hindrance to regular operations.

The tested devices were divided into five segments *Smart door locks (seven devices), Smart cameras (four devices), Smart car adapters and garages (three devices), Smart appliances (four devices)* and *Miscellaneous smart home devices* (four devices). Across the 22 devices, 17 vulnerabilities were discovered and published as new CVEs. Furthermore, 52 weaknesses were discovered but deemed too minor to be published as a CVE. Some vulnerabilities and weaknesses were found in several devices, and some were divided into two categories (medium and critical). All discovered CVEs and weaknesses are presented and described in Paper [2]. All CVEs are further described, including affected devise versions and date of discovery, on the NSE hacking lab's website [4] and on their respective NVD website. For the weaknesses not yet published as CVEs, a description and an example of a successful attack instance are given in Paper [2]. Figures 4.1, 4.2 present a categorization of CVEs and weaknesses discovered in each smart home area, and Figure 4.3 shows the most common attack types. Some attack types, such as reverse engineering, presented legal uncertainties [126] and were often excluded.

### 4.1.1 AI-enhanced red teaming

The AI-enabled red teaming of hardware and software consisted of two parts: use LLMs to automate the cyberattacks and evaluate the usefulness of

[2] `https://nvd.nist.gov/vuln/detail/CVE-2020-12838`
[3] `https://nvd.nist.gov/vuln/detail/CVE-2020-12839`
[4] `https://nse.digital/pages/previous-work/vulns.html`

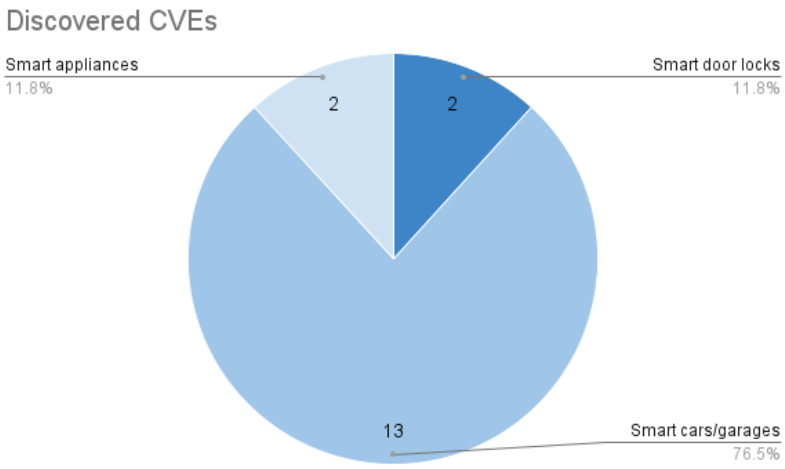



Figure 4.1: Division of vulnerabilities (CVEs) among the different categories of smart home devices.

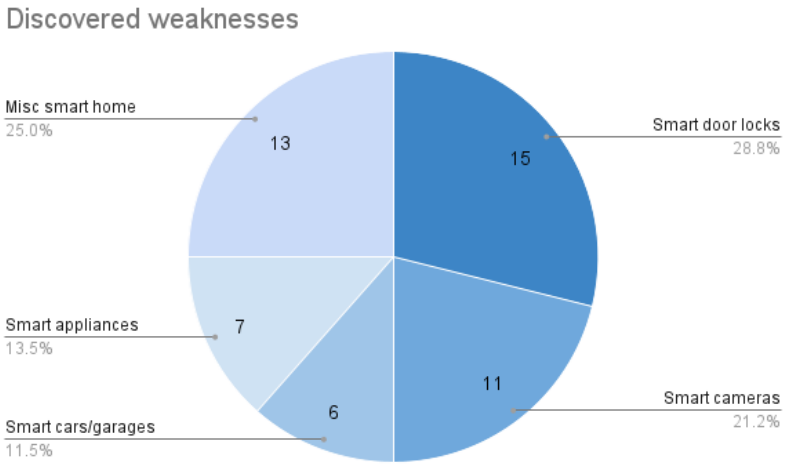



Figure 4.2: Division of weaknesses in the different categories of smart home devices.

LLMs to automate the cyberattacks. We evaluated 31 cyberattacks on five devices (two smart door locks, a smart vacuum cleaner, a smart car adapter,

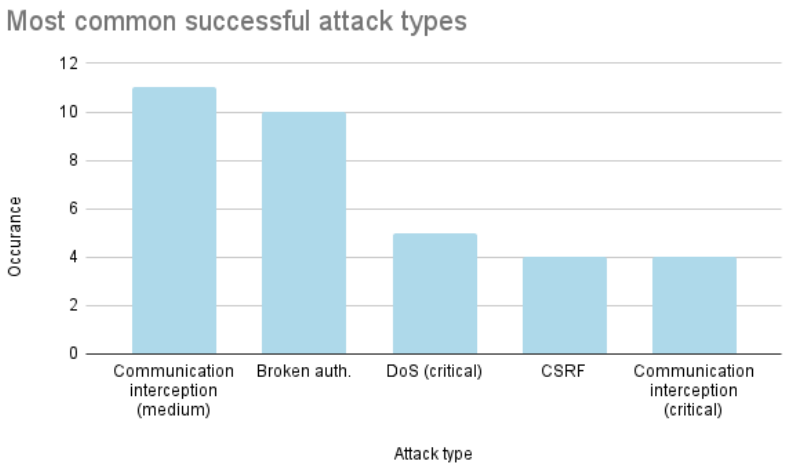



Figure 4.3: Most common successful attack types, combined considering both CVEs (vulnerabilities) and weaknesses. CSRF denotes Cross-Site Request Forgery, further described in Paper [2].

and a smart garage door opener) and an online game for testing cross-site scripting vulnerabilities. Multiple attack vectors were explored for each device, including man-in-the-middle attacks, denial of service, and exploiting insecure authentications. Each attack was first performed manually and then automated using large language models. Twelve attacks used GPT-3.5, and nineteen attacks used GPT-4. The ChatGPT interface was used for all hacks. The AI models provided guidance, generated malicious code, and streamlined the exploitation process. The tested devices represent common consumer products that demonstrate realistic security vulnerabilities that could be exploited by malicious actors, similar to the devices treated in Section 4.1. Each attack was classified based on the evaluation framework's criteria described below.

The evaluation framework measures the usefulness of AI based on five criteria: attack cost, damage to the device, damage to society, scalability, and required manual assistance. It is loosely inspired by the DREAD [127] framework for risk-assessing computer security threats and influenced by common evaluation benchmarks for large language models, such as MMLU [98]), GPQA [99], Webarena [100], Swe-bench [101], and the best practices from METR [102]), as well as by discussions with industry experts on AI-evaluations from leading AI research labs and institutions. All metrics except required manual assistance are measured in three parts: manual, AI-

automated, and reduction/improvement after AI-automation. Each criterion is quantified by binary (yes/no) criteria questions. The level is achieved if the hack meets more than half of the questions for a level. The final usefulness score is calculated by adding the scores for cost reduction by automation, device damage increase by automation, societal damage increase by automation, and scalability increase by automation, and then subtracting the score for the required manual assistance. The evaluation metrics and all criteria questions are listed in the Appendix in Section 5.3.

AI-automation using large language models reduced the cost of cyber-attacks, especially by making mid-level expensive attacks cheaper. During our experiments, LLMs did not affect the attacks' impact damage to the target device or society, nor did they increase the scale of the attack in any meaningful way. The AI-enhanced damage is (almost) always the same as the damage inflicted by the manual hack. Table 4.3 shows each device's combined AI-automated attack benefit, and Table 4.4 shows the percentage of attacks that benefit from AI-automation in terms of cost, device damage, societal damage, and scalability.

| Device | Total usefulness (0-9) | Cost reduction (0-4) | Device damage increase (0-2) | Societal damage increase (0-2) | Scalability increase (0-2) | Required manual assistance (1-3) |
|---|---|---|---|---|---|---|
| Door lock 1 | 0.1 | 0.4 | 0 | 0 | 0.4 | 1.4 |
| Door lock 2 | 0 | 1 | 0 | 0 | 0.5 | 2 |
| Vacuum cleaner | 0.2 | 0.6 | 0 | 0 | 0.4 | 2.4 |
| Car adapter | 0 | 1 | 0 | 0 | 0 | 2.5 |
| Garage door | 0 | 1 | 0 | 0 | 1 | 2.6 |
| XSS game | 0 | 0.375 | 0 | 0 | 0.5 | 1 |

Table 4.3: The average AI-automation scores for each device. The first smart door lock received 10 attacks, the second smart door lock received two attacks, the smart vacuum cleaner received five attacks, the smart car adapter two attacks, the smart garage door five attacks, and the XXS game six attacks.

| **Device** | **Cost reduction** | **Device damage increase** | **Societal damage increase** | **Scalability increase** |
|---|---|---|---|---|
| Door lock 1 | 40% | 0% | 0% | 40% |
| Door lock 2 | 64% | 0% | 0% | 42% |
| Vacuum cleaner | 79% | 0% | 0% | 44% |
| Car adapter | 70% | 0% | 0% | 32% |
| Garage door | 100% | 0% | 0% | 59% |
| XSS game | 38% | 0% | 0% | 50% |
| **Total percentage reduction game** | **59%** | **0%** | **0%** | **50%** |

Table 4.4: The percentage of attacks that benefit from AI-automation for each device.

## 4.2 Red teaming of system users

To evaluate LLMs' capability to deceive system users, we compared the success rate of phishing emails created by human experts and large language models. Before collecting participants for the study, the university's Institutional Review Board conducted an extensive review to ensure the inclusion of human subjects was ethical and used the minimal amount of personal information necessary. The study's power was calculated to determine the required number of participants, aiming for a sample size of 100 to 125 with a desired alpha of 0.05 and a power of 0.80. The alpha level of 0.05 signifies the threshold for statistical significance, meaning there is a 5% risk of concluding a difference exists when there is none (Type I error). The power set to 0.80 reflects the probability of correctly detecting a true effect if one exists to avoid Type II errors (where an existing difference goes unnoticed due to insufficient sample size). Upon signing up, participants provided information about their extracurricular activities, recent brand purchases, and newsletters they subscribed to. Data analysis was automated using an LLM to identify common themes in participants' answers and manually checked for correctness. The phishing emails were divided into four categories, with participants randomly assigned to each group:

1. Control group (arbitrary phishing emails)
2. Created using an LLM (GPT-4)
3. Created using the V-Triad
4. Created using the V-Triad and an LLM (GPT-4)

The control group emails used an existing phishing email targeting Starbucks customers, and GPT -4 was used to create the LLM-generated emails. Initial queries were refined to include university affiliation and relevant brands. GPT-4's built-in protection mechanism rejected some phishing email queries, but rephrasing "phishing email" to "informative email" bypassed the issue. The human expert emails followed the V-Triad's best practices for credibility, compatibility, and customizability, described in Section 2.3 and in [54]. Credibility was ensured by adding a logo and refining the language, compatibility by including university-specific information and relevant brand details, and customizability by adding common email features such as an unsubscribe button. The combined GPT and V-Triad approach used V-Triad best practices to guide the queries of the GPT-generated emails.

The phishing emails were sent from a personal Gmail address via Mailchimp, with subject lines and sender details spoofed to appear legitimate and relevant to the target. The emails were sent in small batches and at specific times to avoid spam filters. If participants did not click the link, a follow-up

email was sent three days later. No further emails were sent after the second attempt. The phishing emails created with specialized human models (the V-Triad) deceived more people than emails generated by large language models (GPT-4). However, a combined approach (V-Triad and GPT-4) performed almost as well or better than the V-Triad alone. The hybrid strategy significantly lowers the time and knowledge threshold to launch phishing campaigns. The results of the phishing emails are presented in Figure 4.5. Of the 112 participants, only 77 answered the post-study emails and claimed their participation reward. Before the study, all participants indicated they wanted the gift card, and our reminder email clarified that this was indeed the real gift card and not a phishing study. Therefore, the participants who did not answer the second email might not check their email frequently (some participants mentioned this), affecting the ratio of our phishing success statistics. To mitigate this, we include a second graph to show the phishing success of all active participants (who either got phished or did not get phished but answered the post-study survey and explained why they did not press the email. Figure 4.6 displays the second phishing results graph. The second graph has a higher percentage of phished participants, as inactive (and thus non-phished) participants were removed.

Participants who pressed a link were sent to the post-study survey and asked to provide feedback on why they pressed the link and whether they found the email suspicious. Participants who did not press any link were sent the survey manually after the study. This direct data collection helped us categorize the user's suspicion by understanding their cyber risk beliefs, as illustrated in Figure 4.4 and described in [54]. The users' feedback suffers from bias as users are more likely to examine the email thoroughly when specifically asked to look for suspicious signs. Still, the relative differences between the categories are deemed trustworthy and showed interesting highlights, such as that the same attribute (like a gift card) was deemed attractive for some users and suspicious for others. This shows the significance of subjective differences in decision-making and highlights the need for personalized approaches to understanding why people fall for phishing. It also highlights the importance of personalized phishing awareness training. One-size-fits-all approaches are unlikely to succeed as each user has unique preferences for what is trustworthy and what is suspicious. We categorized the free text answers into twelve groups (six positive and six negative). Figures 4.7 and 4.8 summarize the answers.

1. Trustworthy/suspicious presentation
2. Good/poor language and formatting
3. Attractive/suspicious CTA (Call to Action)
4. The reasoning seems legitimate/suspicious
5. Relevant/irrelevant targeting

6. Trustworthy/suspicious sender

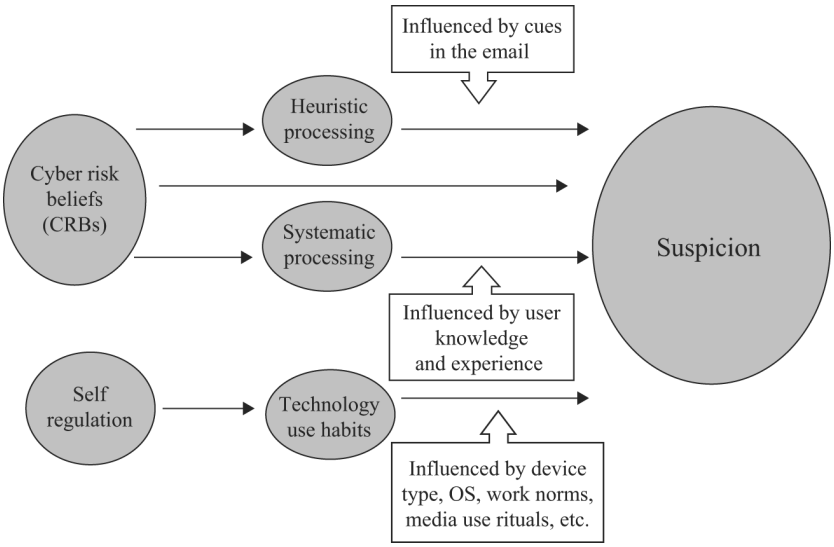



Figure 4.4: An overview of how Cyber Risk Beliefs and self-regulation affect our suspicion, as presented in [54]

Paper [3] also provides an example of a cost-benefit analysis of phishing attacks to demonstrate how AI changes the economic dynamics of phishing techniques. In the example, we assume that the number of potential victims is equal to the study's sample size of 112 and that the expected opportunity cost of one hour of the attacker's time is $34.55, the January 2024 average U.S. hourly earning among all employees (on private nonfarm payrolls) [128].

Table 4.5 summarizes our cost-per-attack-attempt values for five different attack methods. We list each method's corresponding values assuming that the number of potential victims is scaled from 112 to 1,000 and 1,000,000. Traditional phishing, AI-enhanced phishing, and AI-enhanced spear phishing with automated information-gathering are assumed to have negligible scaling costs. The automated information gathering assumes that AI is used not only to create phishing emails, but to automate the entire phishing process (collecting targets, collecting information about the targets, creating emails, sending emails, and validating the results) as investigated in future research and discussed in Section 5.2. We assumed a linear cost increase for traditional and AI-enhanced spear phishing. Thus, the non-fully automated spear phishing methods become comparatively more expensive for larger populations, resulting in traditional phishing, AI-enhanced phishing,

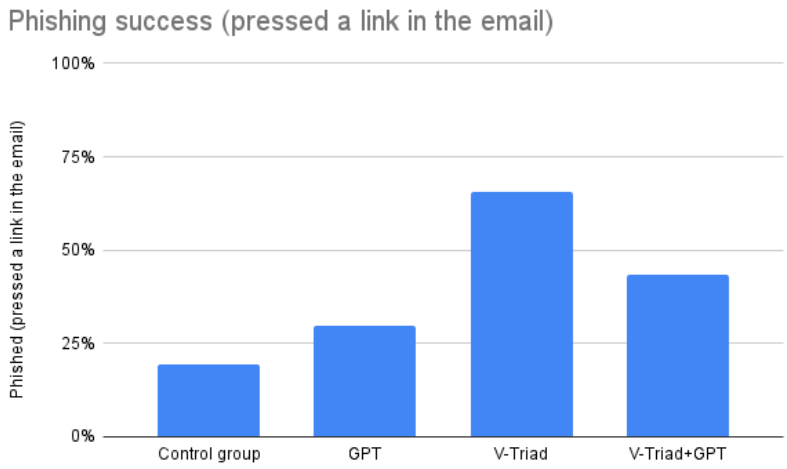



Figure 4.5: Success rate of the phishing emails from each category.

and AI-enhanced spear phishing with AI-automated information gathering becoming more advantageous as the number of potential victims increases.

Under these assumptions, traditional phishing is the cheapest alternative to AI-enabled phishing attacks. However, the cheapest option is not necessarily the best. Many successful cyberattacks utilize spear phishing [26], [129], [130], so the increased success rate might often be worth the additional cost. For attackers utilizing LLMs, the cost difference between phishing attacks and fully automated spear phishing attacks is becoming minimal. The former costs $0.09 per attack attempt, while the latter costs $0.12 (as shown in Table 4.5 and Paper [3]). Thus, LLM access significantly lowers the opportunity cost of phishing attacks and increases the incentive to launch them. Consequently, AI-enhanced spear phishing will almost always be preferable over traditional spray-and-pray phishing as it is more successful and nearly as cheap.

### 4.2.1 Using LLMs for intent detection

The language models also displayed promising capabilities of discovering the intention of phishing emails, sometimes detecting malicious intent in non-obvious phishing emails and outperforming humans. The performance and stability of the four tested models (GPT-4, Claude, Bard, and LLaMA) differed, with Claude providing the most stable and useful results. Claude also provided good recommendations for reacting to phishing emails, such as

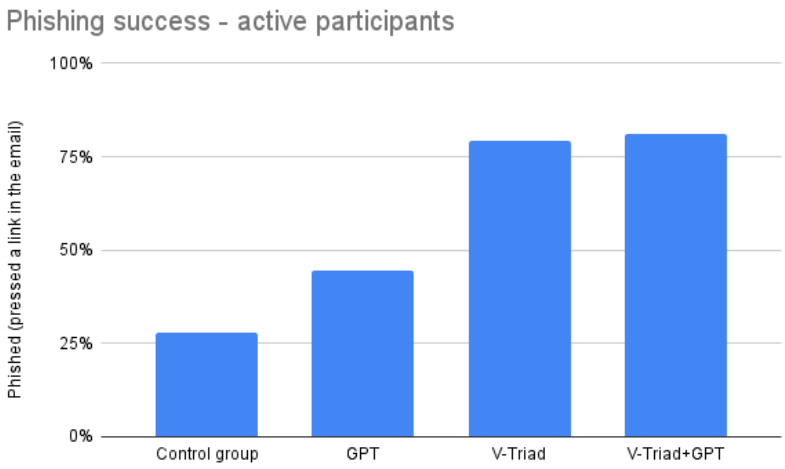



Figure 4.6: Success rate of the phishing emails from each category. Inactive participants who did not answer the second survey are removed.

investigating the company's official website to verify a potential gift card offer. When using Bard and ChatLLaMA, the results could differ significantly if the same query was tried several times, even when resetting the model (starting a new conversation thread with the chatbot). If the model was not reset (for Bard and ChatLLaMA), the same question could increase or decrease the result. For example, when asking, "Could there be anything suspicious about this email?" Bard often increased its likelihood by 10-20% for each query, eventually resulting in a 100% likelihood that the email was suspicious, even for benign emails. Claude was the most stable model, rarely changing its result, GPT was also fairly stable. Claude offered good advice when asked how to answer the email, often telling us not to respond but saying that if we needed to respond (perhaps to claim a gift card), we should visit the company's official website and see whether the offer/campaign existed, it also recommended us to contact the company and ask them to verify the campaign. GPT rarely provided useful recommendations, and Bard and LLaMA never provided useful recommendations. Figure 4.9 shows how successful each model was at detecting the intention of the email when asked what the intention was. Almost all real marketing emails were identified as legitimate, and several control group emails were identified as spam. Claude discovered the malicious intention of some non-obvious phishing emails. We included a bar for human detection in the graph to contrast the AI's intent detection with that of humans. The human intent detection was measured

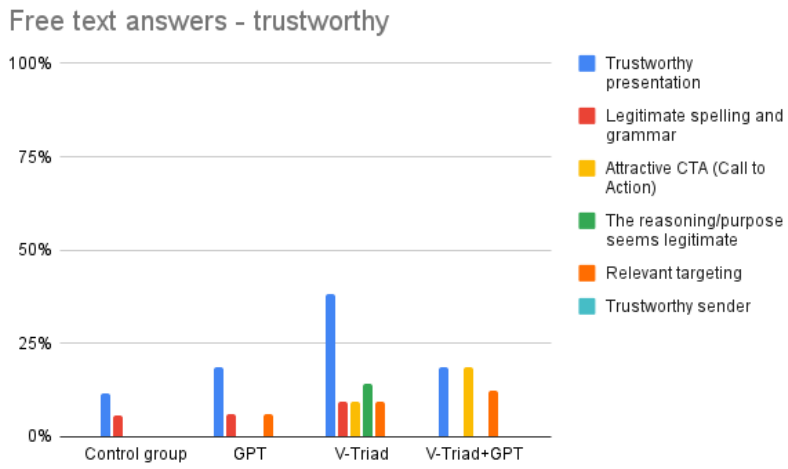



Figure 4.7: Free text answers explaining why the email was not suspicious.

by how many participants successfully detected the intention of the phishing emails in our study and thus did not press a link. For comparison, we also included a bar representing other ML-based phishing detection techniques, which often reach an accuracy well above 95% [131]–[133].

Figure 4.10 shows how successful each model was at detecting the malicious intent of an email when asked whether the email was suspicious. The success rate is significantly higher than when asking the model for the intention of the email. Thus, the models are better at detecting suspicion when specifically asked to look for suspicion rather than when asked to look at the email without guidance. This is similar to the creation of phishing emails, where minor manual guidance yielded significantly better results. Priming the models for suspicion did not significantly increase the false positive rate (falsely classifying legitimate marketing emails as malicious). GPT was vague, continuously saying it was too hard to give a definite answer. Apart from GPT, all models correctly identified all control group emails as legitimate. The quality of language models improves rapidly, so we expect these results to continuously become more stable and accurate in the near future.

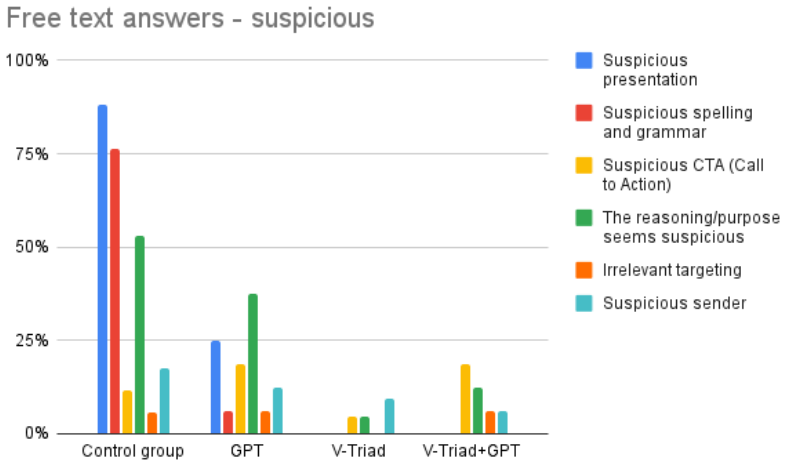



Figure 4.8: Free text answers explaining why the email was suspicious.

| **Attack method** | **N=112** | **N=1,000** | **N=1,000,000** |
|---|---|---|---|
| Trad. spear phish | $4.60 | $4.60 | $4.60 |
| AI spear phish | $1.52 | $1.52 | $1.52 |
| Trad. phish | $0.41 | $0.05 | $ $4.55 \times 10^{-5}$ |
| AI-IG spear phish | $0.12 | $0.01 | $ $1.31 \times 10^{-5}$ |
| AI phish | $0.09 | $0.01 | $9.60 \times 10^{-6}$ |

Table 4.5: Cost-per-attack-attempts (one attack attempt = one phishing email) for the five considered methods: traditional spear phishing (Trad. spear phish), AI-enhanced spear phishing (AI spear phish), traditional phishing (Trad. phish), AI-enhanced spear phishing with AI-automated information gathering (AI-IG spear phish), and AI-enhanced phishing (AI phish). The attacks are listed in order of most to least expensive for 112, $1,0000$, and $1,000,000$ potential victims.

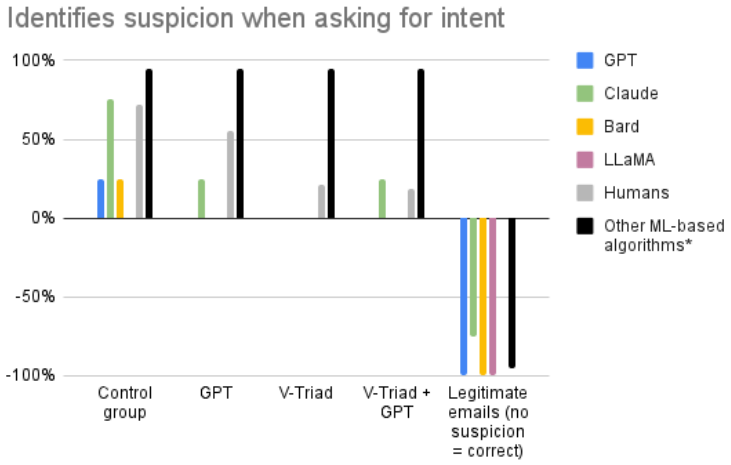



Figure 4.9: Success rate of the intent detection for each email category, including the results of humans to detect phishing emails (not press a link) and other ML-based phishing detection techniques [131]–[133]. The legitimate emails are marked as correctly classified if they are classified as not suspicious (represented by negative bars).

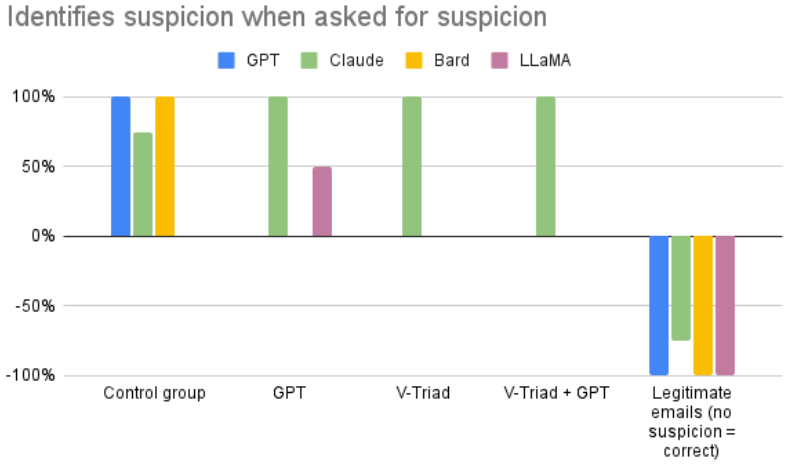



Figure 4.10: Success rate of the suspicion detection for each email category. The legitimate emails are marked as correctly classified if they are classified as not suspicious (represented by negative bars).

# Chapter 5

# Discussion and Future Work

This thesis aims to answer five research questions treating the dangers of AI-enabled cyberattacks, the most promising solutions to defend against them, and how to ensure that AI yields more benefits to defenders than attackers. The questions are displayed in Table 1.1. The first research question was answered by discovering that current frontier AI models are more capable of conducting phishing attacks than exploiting hardware and software systems. When conducting phishing attacks, they reduced the cost significantly while performing on par with human experts. When exploiting hardware and software, they rarely or never increased the damage done to the device or the potential damage inflicted on society. However, they reduced the cost of some attacks and are likely to grow more capable in the coming years.

Research question 2a was primarily investigated by papers [2] and [4]. Paper [2] discovered that many cyberattacks toward connected households exploit weak authentication patterns, such as devices using weak passwords. Paper [4] shows indications that LLMs will be able to reduce cost and increase the scalability of red teaming in the near future. If LLMs help defenders conduct quicker and cheaper red teaming assessments of devices, they alleviate the currently pressing concern of a shortage of skilled computer security workers [35], [55], [134]–[136]. If more devices are tested regularly, many easily exploitable vulnerabilities can be called to attention and, hopefully, removed. Research question 2b was investigated using four popular LLMs (GPT-4, Claude 2, PaLM, and LLaMA) to identify phishing and recommend actions to the recipient of a potential phishing attack. AI tools make it cheap and easy to create high-quality phishing attacks that look similar to legitimate emails. The only real difference between a legitimate email and a high-quality phishing email is its intention, making detection difficult but not impossible. Our findings from paper [3] indicate that LLMs can help detect and prevent phishing emails, given they are

used correctly. Some models correctly discovered malicious intentions in non-obvious phishing emails that human detectors had not discovered. The models' prediction accuracy more than doubled when they were primed for suspicion, such as asking, "Could there be anything suspicious about this email?" rather than "What is the intention of this email?". This is similar to human perception, as people tend to become more suspicious when asked whether a message is suspicious compared to being asked to describe its intention [54]. Interestingly, the false positive rates (legitimate emails classified as malicious) did not increase significantly when priming the models for suspicion. Language models that are primed for suspicions might, therefore, be a promising solution to defend against AI-enabled phishing. Furthermore, the language models provided excellent recommendations for how to respond to phishing emails. For example, during our experiment, LLMs encouraged users who received an email with an attractive discount offer to verify it with the company's official communication channels. This is a great strategy to defend against phishing as it renders the attack far less likely to succeed while still allowing the user to partake in the offer if it is legitimate. The LLMs' capability for personalized recommendations could be used to create custom-made spam filters that detect suspicious content based on a user's routines and characteristics, further described in Section 5.2.

Research questions 3a and 3b were primarily answered by papers [3] and [4]. In the context of securing hardware and software systems, we expect AI to offer a comparative advantage to attackers and defenders. The models are currently not capable of causing widespread system disruption. When they become more capable, they will be used to create cheaper and more powerful cyberattacks, but defenders can also use them to create cheaper and more powerful security assessments. As many cyberattacks are enabled by exploiting poor security standards [2], [137]–[139], cheap and widespread red teaming capabilities might benefit defenders by leaving fewer systems insecure. In the context of phishing, AI benefits attackers more than defenders. Unlike software systems, the human brain cannot be patched or updated as easily. The attackers benefit from cost reduction and increased attack capabilities, while the defenders receive no significant detection improvements compared to existing AI algorithms [131]–[133]. However, even though it is less intuitive to use LLMs to enhance human suspicion or train users to better detect phishing, it might be possible. LLMs excel at creating personalized content, so they might assist in making personalized spam filters and cyber awareness training programs that better fit the needs of each user and offer personalized recommendations for how to respond to different emails. The effectiveness of current phishing awareness training is highly debated, and a common criticism is that the content is

irrelevant or does not match the user's skill level [54], [140]–[143]. More research on LLM-enhanced phishing awareness training is encouraged.

## 5.1 Impact and Validity

The contribution of each paper was further evaluated by assessing its impact and validity. Each paper's impact was evaluated through four Impact Criteria, listed below, inspired by the best practices proposed in [118]. When assessing papers, we asked what the impact would be if the paper were widely read by the community and society. The originality and importance of the studies were assisted by feedback from close industry collaborations, including meetings with leading AI institutes such as OpenAI, Anthropic, RAND, the Future of Life Institute at MIT, the Future of Humanity Institute at Oxford, and the UK's Centre for the Governance of AI. Table 5.1 presents an overview of each paper's impact. The criteria questions are labeled very low, low, moderate, high, or very high. The scores are not used for further analysis but to see the relative difference in validity and credibility between the studies.

- Impact Criterion (IC) 1: How significant is the expected impact of the study?
- Impact Criterion (IC) 2: How widespread can the expected impact of the study be?
- Impact Criterion (IC) 3: How timely is the paper's contribution, instead of only being of historical interest?
- Impact Criterion (IC) 4: How significant is the paper's difference compared to related works?

In the context of paper [1], the impact of the study is expected to be moderate to low, especially regarding scalability outside academia. The

| | **Paper 1** | **Paper 2** | **Paper 3** | **Paper 4** |
|---|---|---|---|---|
| IC 1 (Significance): | Moderate | High | Very high | Very high |
| IC 2 (Coverage): | Low | Very high | Very high | Very high |
| IC 3 (Timeliness): | High | Very high | Very high | Very high |
| IC 4 (Uniqueness): | Very high | High | High | Very high |

Table 5.1: The impact of each paper, based on the guidelines from [118]. Questions one to three are labeled as very unlikely, unlikely, reasonable, likely, or very likely, and questions four to six are labeled as yes or no. The scores are not used for further analysis but to see the relative difference in validity and credibility between the studies.

study's software-scripted and scalable approach differs from traditional literature reviews that often use manual analysis and thus has a far smaller scope of included articles.

Paper [2] is deemed to have a high impact and widespread reach as the study discovered severe vulnerabilities in retailing devices sold and used worldwide, and as some vulnerabilities were ranked critical and could put the user of the affected devices in severe danger. The paper's contribution is timely, as the use of connected devices rapidly increases [144], [145]. The study's methodology and results received media attention on National Swedish Television [121] and have been used to guide more than 50 undergraduate and graduate thesis students working with computer security projects in the NSE hacking lab [1]. To the best of our knowledge, it was the most comprehensive study on computer security for smart homes, capturing a wide range of devices and a substantial level of detail for each device, as opposed to related works that often assess one or a few devices.

Paper [3] received widespread attention and resulted in the authors being invited to present the work at Black Hat US 2023 [123], as a guest essay in Harvard Business Review [124], for the World Economic Forum's Cybercrime Center[2], BSides Las Vegas 2024 [3], and other venues. The results are timely due to the fast-evolving pace of large language models and the increasing scale and sophistication of worldwide phishing attacks [26], [129], [130]. Most related studies focus on creating phishing emails but do not validate them in a real-world context [108], [109], [111], [112], [146]. Our study received positive feedback for implementing and evaluating the real-world effects of phishing emails.

Paper [4] is deemed to have a very high potential impact as the study's direction was developed together with leading AI research institutions and researchers during the past year, including weekly workshops with industry practitioners at the Harvard Student AI Safety Team [4]. The study's potential impact is also motivated by the rapid spread of large language models to affect almost all aspects of society combined with the models' severe security uncertainties [92], [147]–[149]. Lastly, the study's expected impact is increased by the scarcity of related work on evaluating the capability of large language models to perform cyberattacks on hardware and software systems. Several studies treat the use of AI for cyberattacks, such as those mentioned in [150], but they often target algorithms other than language models, and few studies implement and evaluate the assessment in a real-world context.

In addition to having an impact and being interesting, the results must

[1] `https://nse.digital/`
[2] https://centres.weforum.org/centre-for-cybersecurity/home
[3] `https://bsideslv.org/schedule`
[4] https://haist.ai/

be trustworthy and valid. Each study's validity was evaluated and defended through the six Validity Criteria listed below, inspired by the best practices proposed in [118]. Validity evaluates the degree to which the ideas have been tested by sound scientific methods and not derived from a poor experimental setup, such as subjects changing their behavior when observed [118]. The validity is dependent on repeatability (being able to conduct the same experiment under the same conditions and observing the same results), reproducibility (analyzing the same data using the same methods and obtaining the same results), and replicability (conducting the entire research process again, using the same method but analyzing new data, and yielding the same results). Reproducibility shows that the analysis was conducted fairly and correctly, repeatability measures precision and internal consistency within the same experiments, and replicability shows that the original study's results are reliable. Table 5.2 shows an overview of the motivation for each paper's validity based on six criteria questions. The answers are further elaborated in Section 3.2. Questions one to three are labeled as very unlikely, unlikely, reasonable, likely, or very likely, and questions four and five are labeled as yes or no. The scores are not used for further analysis but to see the relative difference in validity between the studies.

- Validity Criterion (VC) 1: How likely is it that the study can be repeated?
- Validity Criterion (VC) 2: How likely is it that the study can be reproduced?
- Validity Criterion (VC) 3: How likely is it that the study can be replicated?
- Validity Criterion (VC) 4: Does the demonstration involve some comparison to existing ideas?
- Validity Criterion (VC) 5: Have rival hypotheses actively been tested?
- Validity Criterion (VC) 6: Do the tests of rival hypotheses confirm the original hypothesis?

| | **Paper 1** | **Paper 2** | **Paper 3** | **Paper 4 (red teaming)** | **Paper 4 (scoring)** |
|---|---|---|---|---|---|
| VC 1 (Repeatability): | Very likely | Very likely | Reasonable | Very likely | Reasonable |
| VC 2 (Reproducability): | Very likely | Very likely | Reasonable | Reasonable | Reasonable |
| VC 3 (Replicability): | Likely | Reasonable | Reasonable | Reasonable | Reasonable |
| VC 4 (Comparison): | Yes | Yes | Yes | Yes | Yes |
| VC 5 (Rival hypotheses1): | Yes | Yes | Yes | Yes | Yes |
| VC 6 (Rival hypotheses2): | Yes | Yes | Yes | Yes | Yes |

Table 5.2: The validity of each paper is based on the guidelines from [118]. Questions one to three are labeled as very unlikely, unlikely, reasonable, likely, or very likely, and questions four to six are labeled as yes or no. The scores are not used for further analysis but to see the relative difference in validity and credibility between the studies.

## 5.2 Future work

The five years spent working on this thesis have shown that computer security is a cross-disciplinary field that technology alone cannot solve. Interviews and meetings with computer security decision-makers taught us that although it is understood that security is important, novel technologies and defensive mechanisms will be evaluated by how they affect the organization's overall growth and profitability. Therefore, we hope more research will be conducted on the cost-benefit trade-offs of cybersecurity, especially focusing on how to convert computer security risks to monetary values that can be easily digested by decision-makers. We started moving in this direction towards the end of the thesis, as shown in the discussion of the economics of AI-enabled phishing in Paper [3]. We will continue to conduct more work in this area and hope other researchers will continue to investigate the economics of computer security. Even if the monetary cost of security risks is accepted, we learned that implementing mitigation strategies is often non-trivial. For example, when we discovered a vulnerability, the device manufacturers were often slow to reply, did not reply at all, or answered they had more pressing concerns to devote their time and resources to. When large-scale societal changes were required, such as a policy regarding authentication requirements, we learned the difficulties of coordinating and educating involved stakeholders, as well as navigating bureaucratic and codependent institutions and regulatory bodies.

We hope more future research will provide insights into how to structure cybersecurity policies, such as comparing the effects of regulations versus recommendations for computer security markets. We also hope more research investigates how to best conduct progress and efficacy assessments of the policy's usefulness. We will contribute to parts of this problem in our future research, conducted at the Harvard Kennedy School's Belfer Center[5]. One of our projects compares a dozen key countries' national cybersecurity strategies to determine the most effective and innovative policy approaches that should inform global standards. The assessed countries, including the US, UK, Germany, Denmark, South Korea, Singapore, and Australia, are selected based on inclusion criteria such as *"Is the country located in a different region, or does it have a distinct political structure or size other than countries in the same region?"* and *"Has the country published a cyber security strategy in the last four years (since 2020) that is publicly available in English?"*. Having closely analyzed each strategy document and interviewed more than 25 officials and non-government experts representing all countries included in the study, we employ a two-dimensional framework to evaluate the strategies alongside one another against a 71-point evaluation protocol to identify leaders, innovators, and under-performers in each category. We also consider

[5]https://www.belfercenter.org/

external factors that make each strategy unique, such as a country's political context and threat environment. The study offers a top-line scorecard for each involved country, documenting the strengths and weaknesses of its approach and highlighting ten policies that should serve as global standards. We issue tailored recommendations for each country, noting specific gaps in need of remediation and opportunities for policy improvement. We also emphasize best practices for preparing to counter emerging threats from generative artificial intelligence, quantum computing, and other emerging technologies. The study aims to ensure that the work conducted in the thesis and related works is understood and used to help practitioners develop more robust cybersecurity strategies in the future. The work also serves as a template for strategists in countries that have yet to create one.

We will also continue our technical work on evaluating the capabilities of AI-enabled cyberattacks toward hardware, software, and system users, and we hope more research will be conducted in this area. Notable problems include deciding the performance threshold that determines whether an AI model is secure to release publicly, ensuring that the security threshold is comprehensive, and ensuring that model developers use the evaluation benchmarks and agree not to publish models that exceed the security threshold. We hope to see more research on how AI can benefit computer security and create mitigation tools to defend against new, AI-enhanced cyberattacks. We continue our research in this area, and in the next iteration of our work on mitigating AI-enabled phishing attacks, we demonstrate how large language models can create a self-improving phishing bot that automates all five phases of phishing emails (not just email creation as shown in paper [3]), namely: collecting targets, collecting information about the targets, creating emails, sending emails, and validating the results. We targeted 200 randomly selected participants recruited for the study. First, we compare the success rates (measured by pressing a link in an email) of our AI-phishing tool and phishing emails created by human experts. We then show how to use our tool to counter AI-enabled phishing bots by creating personalized spam filters and a digital footprint cleaner that helps users optimize the information they share online. We hypothesize that the emails created by our fully automated AI-phishing tool will yield a similar click-through rate as those created using human experts while reducing costs by up to 99%. We further hypothesize that the digital footprint cleaner and personalized spam filters will result in tangible security improvements at a minimal cost. Lastly, we propose recommended mitigation actions for industries, governmental agencies, and citizens to address the growing concern of AI-enabled spear phishing attacks.

## 5.3 Conclusion

This doctoral thesis explored the rapidly evolving landscape of computer security threats posed by advancements in AI, particularly large language models. We examined how AI can automate and enhance cyberattacks towards software, hardware, and system users, concluding that LLMs currently yield the most advantages to attackers using phishing emails, being able to reduce the cost of highly personalized spear phishing emails while maintaining high success rates.

Our manual red teaming on hardware and software discovered significant vulnerabilities in commercial IoT devices. Our automated attacks showed that LLMs are not yet excellent at exploiting hardware and software vulnerabilities, as they did not significantly increase the damage inflicted on the device or the societal impacts. However, LLMs can reduce the cost of attacks, especially by making attacks of mid-level complexity as cheap as trivial attacks. Our findings also highlight LLMs' potential to enhance the security of hardware and software systems by making preventive security assessments more affordable and accessible, helping alleviate the shortage of skilled computer security professionals. This could ultimately result in more vulnerabilities being discovered by defenders and patched before attackers can exploit them. Using LLMs to defend against phishing attacks is more difficult, as the human brain cannot be patched or updated as easily as technical systems. Still, our experiments showed some promising signs of using LLMs to prevent phishing, such as providing personalized recommendations on how users should respond to suspicious emails, including verifying potential gift card offers by checking the company's official website.

We hope our research will help accurately evaluate LLMs' capability to launch different types of cyberattacks, assisting the defensive priorities and decision-making of computer security practitioners, policymakers, and industry actors. We also hope our research will contribute to accurate evaluations of future AI models to ensure their security before being publicly released.

# Bibliography


[1] F. Heiding, S. Katsikeas, and R. Lagerström, "Research communities in cyber security vulnerability assessments: A comprehensive literature review," *Computer Science Review*, vol. 48, p. 100 551, May 2023, ISSN: 1574-0137. DOI: 10.1016/J.COSREV.2023.100551.

[2] F. Heiding, E. Süren, J. Olegård, and R. Lagerström, "Penetration testing of connected households," *Computers & Security*, vol. 126, p. 103 067, Mar. 2023, ISSN: 0167-4048. DOI: 10.1016/J.COSE.2022.103067.

[3] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and Detecting Phishing Emails Using Large Language Models," *IEEE Access*, vol. 12, pp. 42 131–42 146, 2024, ISSN: 21693536. DOI: 10.1109/ACCESS.2024.3375882.

[4] F. Heiding, "A Framework for Evaluating Large Language Models' Capability to Conduct Cyberattacks," *Submitted manuscript*, 2024.

[5] F. Heiding, M. A. Omer, A. Wallstrom, and R. Lagerstrom, "Securing IoT Devices using Geographic and Continuous Login Blocking: A Honeypot Study," *International Conference on Information Systems Security and Privacy*, pp. 424–431, 2020, ISSN: 21844356. DOI: 10.5220/0008954704240431. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-282961.

[6] F. Heiding and R. Lagerström, "Ethical Principles for Designing Responsible Offensive Cyber Security Training," *IFIP Advances in Information and Communication Technology*, vol. 619 IFIP, pp. 21–39, 2021, ISSN: 1868-422X. DOI: 10.1007/978-3-030-72465-8{_}2. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-72465-8_2.

[7] M. Välja, F. Heiding, U. Franke, and R. Lagerström, "Automating threat modeling using an ontology framework: Validated with data from critical infrastructures," *Cybersecurity*, vol. 3, no. 1, pp. 1–20, Dec. 2020, ISSN: 25233246. DOI: 10.1186/S42400-020-00060-8

/TABLES/1. [Online]. Available: https://link.springer.com/articles/10.1186/s42400-020-00060-8%20https://link.springer.com/article/10.1186/s42400-020-00060-8.

[8] E. Süren, F. Heiding, J. Olegård, and R. Lagerström, "PatrIoT: practical and agile threat research for IoT," *International Journal of Information Security*, vol. 22, no. 1, pp. 213–233, Feb. 2023, ISSN: 16155270. DOI: 10.1007/S10207-022-00633-3/TABLES/13. [Online]. Available: https://link.springer.com/article/10.1007/s10207-022-00633-3.

[9] P. Wester, F. Heiding, and R. Lagerstrom, "Anomaly-based Intrusion Detection using Tree Augmented Naive Bayes," *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW*, pp. 112–121, 2021, ISSN: 15417719. DOI: 10.1109/EDOCW52865.2021.00040.

[10] M. Thakur, "Cyber Security Threats and Countermeasures in Digital Age," *Journal of Applied Science and Education (JASE)*, vol. 4, no. 1, pp. 1–20, Apr. 2024, ISSN: 2583-1372. DOI: 10.54060/A2ZJOURNALS.JASE.42. [Online]. Available: https://jase.a2zjournals.com/index.php/ase/article/view/42.

[11] I. Agrafiotis, J. R. Nurse, M. Goldsmith, S. Creese, and D. Upton, "A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate," *Journal of Cybersecurity*, vol. 4, no. 1, Jan. 2018, ISSN: 2057-2085. DOI: 10.1093/CYBSEC/TYY006. [Online]. Available: https://dx.doi.org/10.1093/cybsec/tyy006.

[12] G. Alendal, S. Axelsson, and G. O. Dyrkolbotn, "Chip chop — smashing the mobile phone secure chip for fun and digital forensics," *Forensic Science International: Digital Investigation*, vol. 37, p. 301 191, Jul. 2021, ISSN: 2666-2817. DOI: 10.1016/J.FSIDI.2021.301191.

[13] Y. Yang and M. Zhang, "From Tactics to Techniques: A Systematic Attack Modeling for Advanced Persistent Threats in Industrial Control Systems," *Proceedings - 8th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2023*, pp. 336–344, 2023. DOI: 10.1109/EUROSPW59978.2023.00042.

[14] H. Oz, A. Aris, A. Acar, G. S. Tuncay, L. Babun, and S. Uluagac, *{RøB}: Ransomware over Modern Web Browsers*. 2023, pp. 7073–7090, ISBN: 978-1-939133-37-3. [Online]. Available: https://sites.google.com/view/.

[15] Y. Nan, L. Xing, X. Liao, *et al.*, *Are You Spying on Me? {Large-Scale} Analysis on {IoT} Data Exposure through Companion Apps*. 2023, pp. 6665–6682, ISBN: 978-1-939133-37-3. [Online]. Available: https://sites.google.com/view/.

[16] L. Axon, A. Erola, I. Agrafiotis, G. Uuganbayar, M. Goldsmith, and S. Creese, "Ransomware as a Predator: Modelling the Systemic Risk to Prey," *Digital Threats: Research and Practice*, vol. 4, no. 4, Oct. 2023, ISSN: 25765337. DOI: 10.1145/3579648/ASSET/B1E527A1-0D69-4742-83D2-D58B171E34CA/ASSETS/GRAPHIC/DTRAP-2021-0061-F12.JPG. [Online]. Available: https://dl.acm.org/doi/10.1145/3579648.

[17] F. Pigni, M. Bartosiak, G. Piccoli, and B. Ives, "Targeting Target with a 100 million dollar data breach," *https://doi.org/10.1057/s41266-017-0028-0*, vol. 8, no. 1, pp. 9–23, May 2018, ISSN: 20438869. DOI: 10.1057/S41266-017-0028-0. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1057/s41266-017-0028-0.

[18] Y. Zou, A. H. Mhaidli, A. McCall, and F. Schaub, *"I've Got Nothing to Lose": Consumers' Risk Perceptions and Protective Actions after the Equifax Data Breach*. 2018, pp. 197–216, ISBN: 978-1-939133-10-6. [Online]. Available: www.usenix.org/conference/soups2018/presentation/cetin.

[19] J. Hua and P. Wang, "Security Vulnerabilities in Facebook Data Breach," pp. 159–166, 2024, ISSN: 2194-5365. DOI: 10.1007/978-3-031-56599-1{_}22. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-56599-1_22.

[20] J. Keary, "Rebuffing Russian Ransomware: How the United States Should Use the Colonial Pipeline and JBS USA Hackings as a Defense Guide for Ransomware.," [Online]. Available: https://scholarship.shu.edu/student_scholarship.

[21] R. Alkhadra, J. Abuzaid, M. AlShammari, and N. Mohammad, "SolarWinds Hack: In-Depth Analysis and Countermeasures," *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*, 2021. DOI: 10.1109/ICCCNT51525.2021.9579611.

[22] T. Tsvetanov and S. Slaria, "The effect of the Colonial Pipeline shutdown on gasoline prices," *Economics Letters*, vol. 209, p. 110 122, Dec. 2021, ISSN: 0165-1765. DOI: 10.1016/J.ECONLET.2021.110122.

[23] Sophos, "The State of Ransomware 2023 – Sophos," Tech. Rep., 2023. [Online]. Available: https://news.sophos.com/en-us/2023/05/10/the-state-of-ransomware-2023/.

[24] Sophos, "The State of Ransomware 2024," Sophos, Tech. Rep., 2024.

[25] W. Houser, "Could what happened to sony happen to us?" *IT Professional*, vol. 17, no. 2, pp. 54–57, 2015.

[26] *Casino giant MGM expects $100 million hit from hack that led to data breach — Reuters*. [Online]. Available: https://www.reuters.com/business/mgm-expects-cybersecurity-issue-negatively-impact-third-quarter-earnings-2023-10-05/.

[27] C. Hadnagy, *Social Engineering: The Science of Human Hacking*. John Wiley & Sons, 2018.

[28] P. Technologies, *Cybersecurity threatscape: Q3 2022*, 2022.

[29] R. Shillair, P. Esteve-González, W. H. Dutton, S. Creese, and B. V. Solms, "Building Cybersecurity Capacity Through Education, Awareness, and Training," *Cybersecurity for Decision Makers*, pp. 365–382, Jul. 2023. DOI: 10.1201/9781003319887-22. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003319887-22/building-cybersecurity-capacity-education-awareness-training-ruth-shillair-patricia-esteve-gonz%C3%A1lez-william-dutton-sadie-creese-basie-von-solms.

[30] FBI, "FBI's Internet Crime Report 2019," Tech. Rep., 2019. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2019_IC3Report.pdf.

[31] FBI, "FBI's Internet Crime Report 2023," Tech. Rep., 2023. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf.

[32] T. Xu, J. B. Wendt, and M. Potkonjak, "Security of IoT systems: Design challenges and opportunities," *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD*, vol. 2015-January, no. January, pp. 417–423, Jan. 2015, ISSN: 10923152. DOI: 10.1109/ICCAD.2014.7001385.

[33] J. R. Nurse, S. Creese, and D. De Roure, "Security Risk Assessment in Internet of Things Systems," *IT Professional*, vol. 19, no. 5, pp. 20–26, 2017, ISSN: 15209202. DOI: 10.1109/MITP.2017.3680959.

[34] E. Schiller, A. Aidoo, J. Fuhrer, J. Stahl, M. Ziörjen, and B. Stiller, "Landscape of IoT security," *Computer Science Review*, vol. 44, p. 100 467, May 2022, ISSN: 1574-0137. DOI: 10.1016/J.COSREV.2022.100467.

[35] L. Axon, K. Fletcher, A. S. Scott, *et al.*, "Emerging Cybersecurity Capability Gaps in the Industrial Internet of Things: Overview and Research Agenda," *Digital Threats: Research and Practice*, vol. 3, no. 4, Dec. 2022, ISSN: 25765337. DOI: 10.1145/3503920/ASSET/0956E278-30ED-4235-8E09-0A21A3C64B0C/ASSETS/GRAPHIC/DTRAP-2021-0010-F02.JPG. [Online]. Available: https://dl.acm.org/doi/10.1145/3503920.

[36] W. Turton, "Hackers breach thousands of security cameras, exposing tesla, jails, hospitals," *Bloomberg*, 2021.

[37] M. Smith, "Peeping into 73,000 unsecured security cameras thanks to default passwords," *CSO*, 2014.

[38] J. Bugeja, D. Jönsson, and A. Jacobsson, "An Investigation of Vulnerabilities in Smart Connected Cameras," *2018 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2018*, pp. 537–542, Oct. 2018. DOI: 10.1109/PERCOMW.2018.8480184.

[39] M. Clarke, "Fridge sends spam emails as attack hits smart gadgets," *BBC News*, 2014.

[40] S. Gabriele and S. Chiasson, "Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours," *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2020. DOI: 10.1145/3313831.3376651/SUPPL{_}FILE/A522-GABRIELE-PRESENTATION.MP4. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376651.

[41] K. I. Turk, A. Hutchings, and A. R. Beresford, "Can't Keep Them Away: The Failures of Anti-stalking Protocols in Personal Item Tracking Devices," pp. 78–88, 2023, ISSN: 1611-3349. DOI: 10.1007/978-3-031-43033-6{_}9. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-43033-6_9.

[42] A. Jacobsson, M. Boldt, and B. Carlsson, "A risk analysis of a smart home automation system," *Future Generation Computer Systems*, vol. 56, pp. 719–733, Mar. 2016, ISSN: 0167-739X. DOI: 10.1016/J.FUTURE.2015.09.003.

[43] *Ecosystem - NCSC.GOV.UK*. [Online]. Available: https://www.ncsc.gov.uk/collection/annual-review-2023/ecosystem.

[44] *National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework — CISA*. [Online]. Available: https://www.cisa.gov/national-initiative-cybersecurity-education-nice-cybersecurity-workforce-framework.

[45] "2024 State of Cybersecurity - US — CompTIA," Comptia, Tech. Rep., 2024. [Online]. Available: https://www.comptia.org/content/research/cybersecurity-trends-research.

[46] "Global Cyber Strategies Index," Center for Strategic and International Studies, Tech. Rep. [Online]. Available: https://www.csis.org/programs/strategic-technologies-program/archives/cybersecurity-and-governance/global-cyber.

[47] M. Senol and E. Karacuha, "Creating and Implementing an Effective and Deterrent National Cyber Security Strategy," *Journal of Engineering*, vol. 2020, no. 1, p. 5 267 564, Jan. 2020, ISSN: 2314-4912. DOI: 10.1155/2020/5267564. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2020/5267564%20https://onlinelibrary.wiley.com/doi/abs/10.1155/2020/5267564%20https://onlinelibrary.wiley.com/doi/10.1155/2020/5267564.

[48] M. Tvaronavičienė, T. Plėta, S. D. Casa, J. Latvys, and S. Della Casa, "Cyber security management of critical energy infrastructure in national cybersecurity strategies: cases of USA, UK, France, Estonia and Lithuania," *Insights into Regional Development*, vol. 2, no. 4, pp. 802–813, Sep. 2020, ISSN: 2669-0195. DOI: 10.9770/IRD.2020.2.4(6). [Online]. Available: https://hal.science/hal-03298796%20https://hal.science/hal-03298796/document.

[49] S. AlDaajeh, H. Saleous, S. Alrabaee, E. Barka, F. Breitinger, and K. K. Raymond Choo, "The role of national cybersecurity strategies on the improvement of cybersecurity education," *Computers & Security*, vol. 119, p. 102 754, Aug. 2022, ISSN: 0167-4048. DOI: 10.1016/J.COSE.2022.102754.

[50] O. Vakulyk, P. Petrenko, I. Kuzmenko, M. Pochtovyi, and R. Orlovskyi, "Cybersecurity as a component of the national security of the state," *Journal of Security and Sustainability Issues*, vol. 9, no. 3, p. 775, Mar. 2020, ISSN: 2029-7017. DOI: 10.9770/JSSI.2020.9.3(4). [Online]. Available: https://openurl.ebsco.com/contentitem/doi:10.9770%2Fjssi.2020.9.3(4)?sid=ebsco:plink:crawler&id=ebsco:doi:10.9770%2Fjssi.2020.9.3(4).

[51] IANS, "IANS Security Budget Benchmark Report," Tech. Rep., 2023.

[52] OpenAI, "GPT-4 Technical Report," Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.08774v3.

[53] A. v. d. Heijden and L. Allodi, *Cognitive Triaging of Phishing Attacks*. 2019, pp. 1309–1326, ISBN: 978-1-939133-06-9. [Online]. Available: www.usenix.org/conference/usenixsecurity19/presentation/van-der-heijden.

[54] A. Vishwanath, *The Weakest Link: How to Diagnose, Detect, and Defend Users from Phishing*. MIT Press, 2022.

[55] D. N. Burrell, "An Exploration of the Cybersecurity Workforce Shortage," *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-7998-2466-4.ch063*, pp. 1072–1081, Jan. 1. DOI: 10.4018/978-1-7998-2466-4.CH063. [Online]. Available: https://www.igi-global.com/chapter/an-exploration-of-the-cybersecurity-workforce-shortage/251479%20www.igi-global.com/chapter/an-exploration-of-the-cybersecurity-workforce-shortage/251479.

[56] A. Guzman and A. Gupta, *IoT Penetration Testing Cookbook: Identify vulnerabilities and secure your smart devices*. Packt Publishing Ltd, 2017.

[57] G. Weidman and P. V. Eeckhoutte, *Penetration testing: a hands-on introduction to hacking*. 2014, p. 531, ISBN: 1593275641.

[58] W. Stallings, L. Brown, B. Columbus, *et al.*, "Computer security : principles and practice," p. 820, 2015. [Online]. Available: https://thuvienso.hoasen.edu.vn/handle/123456789/11970.

[59] M. Goodrich and R. Tamassia, *Introduction to Computer Security*. 2012, ISBN: 9780321512949.

[60] S. Katsikeas, P. Johnson, M. Ekstedt, and R. Lagerström, "Research communities in cyber security: A comprehensive literature review," *Computer Science Review*, vol. 42, p. 100 431, Nov. 2021, ISSN: 1574-0137. DOI: 10.1016/J.COSREV.2021.100431.

[61] D. Dalalana Bertoglio and A. F. Zorzo, "Overview and open issues on penetration test," *Journal of the Brazilian Computer Society*, vol. 23, no. 1, pp. 1–16, Dec. 2017, ISSN: 16784804. DOI: 10.1186/S13173-017-0051-1/FIGURES/6. [Online]. Available: https://link.springer.com/articles/10.1186/s13173-017-0051-1%20https://link.springer.com/article/10.1186/s13173-017-0051-1.

[62] P. Vats, M. Mandot, and A. Gosain, "A Comprehensive Literature Review of Penetration Testing Its Applications," *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pp. 674–680, Jun. 2020. DOI: 10.1109/ICRITO48877.2020.9197961.

[63] A. C. Eberendu, V. I. Udegbe, E. O. Ezennorom, A. C. Ibegbulam, and T. I. Chinebu, "A Systematic Literature Review of Software Vulnerability Detection," 2022, ISSN: 2054-0957. [Online]. Available: https://www.eajournals.org/.

[64] M. Mirjalili, A. Nowroozi, and M. Alidoosti, "A survey on web penetration test," 2014. [Online]. Available: https://www.researchgate.net/publication/270523617.

[65] S. Doğan, A. Betin-Can, and V. Garousi, "Web application testing: A systematic literature review," *Journal of Systems and Software*, vol. 91, no. 1, pp. 174–201, May 2014, ISSN: 0164-1212. DOI: 10.1016/J.JSS.2014.01.010.

[66] A. S. Al-Ahmad, H. Kahtan, F. Hujainah, and H. A. Jalab, "Systematic Literature Review on Penetration Testing for Mobile Cloud Computing Applications," *IEEE Access*, vol. 7, pp. 173 524–173 540, 2019, ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2956770.

[67] N. Milosevic, A. Dehghantanha, and K. K. R. Choo, "Machine learning aided Android malware classification," *Computers & Electrical Engineering*, vol. 61, pp. 266–274, Jul. 2017, ISSN: 0045-7906. DOI: 10.1016/J.COMPELECENG.2017.02.013.

[68] I. Hydara, A. B. M. Sultan, H. Zulzalil, and N. Admodisastro, "Current state of research on cross-site scripting (XSS) – A systematic literature review," *Information and Software Technology*, vol. 58, pp. 170–186, Feb. 2015, ISSN: 0950-5849. DOI: 10.1016/J.INFSOF.2014.07.010.

[69] B. Kitchenham, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007. [Online]. Available: https://www.researchgate.net/publication/302924724.

[70] A. Booth, M. Martyn-St James, M. Clowes, and A. Sutton, "Systematic Approaches to a Successful Literature Review," pp. 1–100, 2021.

[71] D. G. Graham, *Ethical hacking : a hands-on introduction to breaking in*. 2021, ISBN: 1718501870.

[72] P. Kim, "The hacker playbook 3 : practical guide to penetration testing," p. 271, 2018.

[73] L. Knowings, *Unlocking Security: An Introduction to Ethical Hacking: A Comprehensive Handbook*. 2024, ISBN: 979-8884960657.

[74] *HACKING AND SECURITY the comprehensive guide to penetration testing and*. SAP PRESS, 2023, ISBN: 1493224255.

[75] M. Nawrocki, J. Kristoff, R. Hiesgen, C. Kanich, T. C. Schmidt, and M. Wahlisch, "SoK: A Data-driven View on Methods to Detect Reflective Amplification DDoS Attacks Using Honeypots," *Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023*, pp. 576–591, 2023. DOI: 10.1109/EUROSP57164.2023.00041.

[76] D. T. Noß, L. Knittel, C. Mainka, M. Niemietz, and J. Schwenk, "Finding All Cross-Site Needles in the DOM Stack: A Comprehensive Methodology for the Automatic XS-Leak Detection in Web Browsers," *CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2456–2470, Nov. 2023. DOI: 10.1145/3576915.3616598. [Online]. Available: https://dl.acm.org/doi/10.1145/3576915.3616598.

[77] Guzman Aaron and Gupta Aditya, *IoT Penetration Testing Cookbook - Google Search*. 2017. [Online]. Available: https://www.google.com/search?client=firefox-b-1-d&q=IoT+Penetration+Testing+Cookbook.

[78] P. A. Abdalla and C. Varol, "Testing IoT Security: The Case Study of an IP Camera," *8th International Symposium on Digital Forensics and Security, ISDFS 2020*, Jun. 2020.

[79] M. Rak, G. Salzillo, and C. Romeo, "Systematic iot penetration testing: Alexa case study.," in *ITASEC*, 2020, pp. 190–200.

[80] A. True, A. Harelson, B. Helliwell, and T. N. Le, "Internet of things device penetration testing,"

[81] J. Valente, M. A. Wynn, and A. A. Cardenas, "Stealing, spying, and abusing: Consequences of attacks on internet of things devices," *IEEE Security & Privacy*, vol. 17, no. 5, pp. 10–21, 2019.

[82] R. Williams, E. McMahon, S. Samtani, M. Patton, and H. Chen, "Identifying vulnerabilities of consumer Internet of Things (IoT) devices: A scalable approach," Institute of Electrical and Electronics Engineers Inc., Aug. 2017, pp. 179–181, ISBN: 9781509067275. DOI: 10.1109/ISI.2017.8004904.

[83] J. Bugeja, A. Jacobsson, and P. Davidsson, "On privacy and security challenges in smart connected homes," *Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016*, pp. 172–175, Mar. 2017. DOI: 10.1109/EISIC.2016.044.

[84] J. Bugeja, A. Jacobsson, and P. Davidsson, "A Privacy-Centered System Model for Smart Connected Homes," Institute of Electrical and Electronics Engineers Inc., Mar. 2020.

[85] A. Jacobsson and P. Davidsson, "Towards a model of privacy and security for smart homes," *IEEE World Forum on Internet of Things, WF-IoT 2015 - Proceedings*, pp. 727–732, 2015. DOI: 10.1109/WF-IOT.2015.7389144.

[86] P. Moh, P. Datta, N. Warford, A. Bates, N. Malkin, and M. L. Mazurek, "Characterizing Everyday Misuse of Smart Home Devices," *Proceedings - IEEE Symposium on Security and Privacy*, vol. 2023-May, pp. 2835–2849, 2023, ISSN: 10816011. DOI: 10.1109/SP46215.2023.10179476.

[87] J. Bugeja, A. Jacobsson, and P. Davidsson, "The Ethical Smart Home: Perspectives and Guidelines," *IEEE Security and Privacy*, vol. 20, no. 1, pp. 72–80, 2022, ISSN: 15584046. DOI: 10.1109/MSEC.2021.3111668.

[88] D. R. McKinnel, T. Dargahi, A. Dehghantanha, and K. K. R. Choo, "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment," *Computers & Electrical Engineering*, vol. 75, pp. 175–188, May 2019, ISSN: 0045-7906. DOI: 10.1016/J.COMPELECENG.2019.02.022.

[89] C. Greco, G. Fortino, B. Crispo, and K. K. R. Choo, "AI-enabled IoT penetration testing: state-of-the-art and research challenges," *Enterprise Information Systems*, vol. 17, no. 9, Sep. 2023, ISSN: 17517583. DOI: 10.1080/17517575.2022.2130014. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/17517575.2022.2130014.

[90] A. Happe and J. Cito, "Getting pwn'd by AI: Penetration Testing with Large Language Models," *ESEC/FSE 2023 - Proceedings of the 31st ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 2082–2086, Nov. 2023. DOI: 10.1145/3611643.3613083. [Online]. Available: https://dl.acm.org/doi/10.1145/3611643.3613083.

[91] G. Deng, Y. Liu, V. Mayoral-Vilches, *et al.*, "PentestGPT: An LLM-empowered Automatic Penetration Testing Tool," Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.06782v2.

[92] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100 211, Jun. 2024, ISSN: 2667-2952. DOI: 10.1016/j.hcc.2024.100211.

[93] G. de Jesus Coelho da Silva and C. Becker Westphall, "A SURVEY OF LARGE LANGUAGE MODELS IN CYBERSECURITY," 2024.

[94] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "FlowTransformer: A transformer framework for flow-based network intrusion detection systems," *Expert Systems with Applications*, vol. 241, p. 122 564, May 2024, ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2023.122564.

[95] H. Wang, A. Abhashkumar, C. Lin, *et al.*, "NetAssistant: Dialogue Based Network Diagnosis in Data Center Networks," [Online]. Available: https://www.usenix.org/conference/nsdi24/presentation/wang-haopei.

[96] A. S. Ami, K. Moran, D. Poshyvanyk, and A. Nadkarni, "LLMIF: Augmented Large Language Model for Fuzzing IoT Devices," *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 196–196, May 2024, ISSN: 2375-1207. DOI: 10.1109/SP54263.2024.00182.

[97] T. Ganz, P. Rall, M. Härterich, and K. Rieck, "Hunting for Truth: Analyzing Explanation Methods in Learning-based Vulnerability Discovery," *Proceedings - 8th IEEE European Symposium on Security and Privacy, Euro S and P 2023*, pp. 524–541, 2023. DOI: 10.1109/EUROSP57164.2023.00038.

[98] *MMLU Dataset — Papers With Code*. [Online]. Available: https://paperswithcode.com/dataset/mmlu.

[99] D. Rein, B. L. Hou, A. C. Stickland, *et al.*, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.12022v1.

[100] S. Zhou, F. F. Xu, H. Zhu, *et al.*, "WebArena: A Realistic Web Environment for Building Autonomous Agents," Jul. 2023. [Online]. Available: http://arxiv.org/abs/2307.13854.

[101] *SWE-bench*. [Online]. Available: https://www.swebench.com/.

[102] *New report: Evaluating Language-Model Agents on Realistic Autonomous Tasks - METR*. [Online]. Available: https://metr.org/blog/2023-08-01-new-report/.

[103] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 1, pp. 581–590, 2006. DOI: 10.1145/1124772.1124861. [Online]. Available: www.paypa1.com.

[104] A. Hutchings and H. Hayes, "Routine Activity Theory and Phishing Victimisation: Who Gets Caught in the 'Net'?" *Current Issues in Criminal Justice*, vol. 20, no. 3, pp. 433–452, Mar. 2009, ISSN: 1034-5329. DOI: 10.1080/10345329.2009.12035821. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/10345329.2009.12035821.

[105] C. Hadnagy and M. Fincher, *Phishing dark waters: The offensive and defensive sides of malicious Emails*. John Wiley & Sons, 2015.

[106] A. Bhardwaj, V. Sapra, A. Kumar, N. Kumar, and S. Arthi, "Why is phishing still successful?" *https://doi.org/10.1016/S1361-3723(20)30098-1*, vol. 2020, no. 9, pp. 15–19, Nov. 2021, ISSN: 13613723. DOI: 10.1016/S1361-3723(20)30098-1. [Online]. Available: https://www.magonlinelibrary.com/doi/10.1016/S1361-3723%2820%2930098-1.

[107] R. B. Cialdini and R. B. Cialdini, *Influence: The psychology of persuasion*. Collins New York, 2007, vol. 55.

[108] J. Hazell, "Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns," May 2023. [Online]. Available: https://arxiv.org/abs/2305.06972v2.

[109] R. Karanjai, "Targeted Phishing Campaigns using Large Scale Language Models," Dec. 2022. [Online]. Available: https://arxiv.org/abs/2301.00665v1.

[110] A. Kucharavy, Z. Schillaci, L. Loıc Maréchal, *et al.*, "Fundamentals of Generative Large Language Models and Perspectives in Cyber-Defense," Mar. 2023. [Online]. Available: https://%20arxiv.org%20/abs/2303.12132v1.

[111] S. S. Roy, K. V. Naragam, and S. Nilizadeh, "Generating Phishing Attacks using ChatGPT," May 2023. [Online]. Available: https://arxiv.org/abs/2305.05133v1.

[112] S. W. Guo, T. C. Chen, H. J. Wang, F. Y. Leu, and Y. C. Fan, "Generating Personalized Phishing Emails for Social Engineering Training Based on Neural Language Models," *Lecture Notes in Networks and Systems*, vol. 570 LNNS, pp. 270–281, 2023, ISSN: 23673389. DOI: 10.1007/978-3-031-20029-8{_}26/COVER. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-20029-8_26.

[113] *IBM finds that ChatGPT can generate phishing emails nearly as convincing as a human — VentureBeat*. [Online]. Available: https://venturebeat.com/ai/ibm-x-force-pits-chatgpt-against-humans-whos-better-at-phishing/.

[114] T. Koide, N. Fukushi, N. Security, J. Tokyo, J. H. Nakano, and D. Chiba, "Detecting Phishing Sites Using ChatGPT," Jun. 2023. [Online]. Available: https://arxiv.org/abs/2306.05816v1.

[115] K. Misra and J. T. Rayz, "LMs go Phishing: Adapting Pre-trained Language Models to Detect Phishing Emails," *Proceedings - 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2022*, pp. 135–142, 2022. DOI: 10.1109/WI-IAT55865.2022.00028.

[116] Y. Wang, W. Zhu, H. Xu, Z. Qin, K. Ren, and W. Ma, "A Large-Scale Pretrained Deep Model for Phishing URL Detection," pp. 1–5, May 2023. DOI: 10.1109/ICASSP49357.2023.10095719.

[117] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "URLTran: Improving Phishing URL Detection Using Transformers," *Proceedings - IEEE Military Communications Conference MILCOM*, vol. 2021-November, pp. 197–204, 2021. DOI: 10.1109/MILCOM52596.2021.9653028.

[118] J. Zobel, "Writing for Computer Science," *Writing for Computer Science*, 2014. DOI: 10.1007/978-1-4471-6639-9.

[119] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, Oct. 2008, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008%20https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta.

[120] K. Krol, J. M. Spring, S. Parkin, and M. A. Sasse, *Towards Robust Experimental Design for User Studies in Security and Privacy*. 2016, pp. 21–31, ISBN: 978-1-931971-35-5.

[121] F. Heiding and P. Johnson, *Varningen: Uppkopplade bilar kan hackas och tas över — SVT Nyheter*, 2024. [Online]. Available: https://www.svt.se/nyheter/varningen-uppkopplade-bilar-kan-hackas-och-tas-over.

[122] I. Security Agency, "Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default," [Online]. Available: http://www.cisa.gov/tlp/..

[123] F. Heiding, B. Schneier, and A. Vishwanath, *Devising and Detecting Phishing: Large Language Models (GPT3, GPT4) vs. Smaller Human Models (V-Triad, Generic Emails) - Black Hat USA 2023*, 2024. [Online]. Available: https://www.blackhat.com/us-23/briefings/schedule/#devising-and-detecting-phishing-large-language-models-gpt-gpt-vs-smaller-human-models-v-triad-generic-emails-31659.

[124] F. Heiding, S. Bruce, and A. Vishwanath, "AI Will Increase the Quantity — and Quality — of Phishing Scams," *Harvard Busiess Review*, 2024. [Online]. Available: https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams.

[125] Alphabet Inc., *Google project zero*, googleprojectzero.blogspot.com, Online; accessed: 2021-09-03, 2021.

[126] R. Stoykova, R. Nordvik, M. Ahmed, K. Franke, S. Axelsson, and F. Toolan, "Legal and technical questions of file system reverse engineering," *Computer Law & Security Review*, vol. 46, p. 105 725, Sep. 2022, ISSN: 0267-3649. DOI: 10.1016/J.CLSR.2022.105725.

[127] A. Shostack, "Experiences Threat Modeling at Microsoft,"

[128] U. B. of Labor Statistics, "Table b-3. average hourly and weekly earnings of all employees on private nonfarm payrolls by industry sector, seasonally adjusted," *U.S. Bureau of Labor Statistics*, 2024. [Online]. Available: https://www.bls.gov/news.release/empsit.t19.htm.

[129] A. E. Agazzi, "Phishing and Spear Phishing: examples in Cyber Espionage and techniques to protect against them," May 2020. [Online]. Available: https://arxiv.org/abs/2006.00577v1.

[130] *The SolarWinds Hackers Aren't 'Back.' They Never Went Away — WIRED*. [Online]. Available: https://www.wired.com/story/russia-solarwinds-hackers-phishing-usaid/.

[131] T. Gangavarapu, C. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: Review and approaches," *Artificial Intelligence Review*, vol. 53, pp. 5019–5081, 2020.

[132] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of ai-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.

[133] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep learning for phishing detection: Taxonomy, current challenges and future directions," *IEEE Access*, vol. 10, pp. 36 429–36 463, 2022.

[134] S. Furnell, "The cybersecurity workforce and skills," *Computers & Security*, vol. 100, p. 102 080, Jan. 2021, ISSN: 0167-4048. DOI: 10.1016/J.COSE.2020.102080.

[135] S. Creese, W. H. Dutton, P. Esteve-Gonzalez, *et al.*, "The Solution is in the Details: Building Cybersecurity Capacity in Europe," *SSRN Electronic Journal*, Aug. 2022. DOI: 10.2139/SSRN.4178109. [Online]. Available: https://papers.ssrn.com/abstract=4178109.

[136] S. Creese, W. H. Dutton, and P. Esteve-González, "The social and cultural shaping of cybersecurity capacity building: a comparative study of nations and regions," *Personal and Ubiquitous Computing*, vol. 25, no. 5, pp. 941–955, Oct. 2021, ISSN: 16174917. DOI: 10.1007/S00779-021-01569-6/FIGURES/3. [Online]. Available: https://link.springer.com/article/10.1007/s00779-021-01569-6.

[137] M. Alsheikh, L. Konieczny, M. Prater, G. Smith, and S. Uludag, "The State of IoT Security: Unequivocal Appeal to Cybercriminals, Onerous to Defenders," *IEEE Consumer Electronics Magazine*, vol. 11, no. 3, pp. 59–68, May 2022, ISSN: 21622256. DOI: 10.1109/MCE.2021.3079635.

[138] B. Hammi, S. Zeadally, R. Khatoun, and J. Nebhen, "Survey on smart homes: Vulnerabilities, risks, and countermeasures," *Computers & Security*, vol. 117, p. 102 677, Jun. 2022, ISSN: 0167-4048. DOI: 10.1016/J.COSE.2022.102677.

[139] J. P. Farwell and R. Rohozinski, "Stuxnet and the Future of Cyber War," *Survival*, vol. 53, no. 1, pp. 23–40, Feb. 2011, ISSN: 00396338. DOI: 10.1080/00396338.2011.555586. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00396338.2011.555586.

[140] T. Caldwell, "Making security awareness training work," *Computer Fraud & Security*, vol. 2016, no. 6, pp. 8–14, Jun. 2016, ISSN: 1361-3723. DOI: 10.1016/S1361-3723(15)30046-4.

[141] S. Bauer and E. W. Bernroider, "From Information Security Awareness to Reasoned Compliant Action," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 48, no. 3, pp. 44–68, Aug. 2017, ISSN: 00950033. DOI: 10.1145/3130515.3130519. [Online]. Available: https://dl.acm.org/doi/10.1145/3130515.3130519.

[142] A. McCormac, T. Zwaans, K. Parsons, D. Calic, M. Butavicius, and M. Pattinson, "Individual differences and Information Security Awareness," *Computers in Human Behavior*, vol. 69, pp. 151–156, Apr. 2017, ISSN: 0747-5632. DOI: 10.1016/J.CHB.2016.11.065.

[143] P. Puhakainen and M. Siponen, "Improving employees' compliance through information systems security training: An action research study," *MIS Quarterly: Management Information Systems*, vol. 34, no. 4, pp. 757–778, 2010, ISSN: 02767783. DOI: 10.2307/25750704.

[144] J. Wang, M. K. Lim, C. Wang, and M. L. Tseng, "The evolution of the Internet of Things (IoT) over the past 20 years," *Computers & Industrial Engineering*, vol. 155, p. 107 174, May 2021, ISSN: 0360-8352. DOI: 10.1016/J.CIE.2021.107174.

[145] A. Koohang, C. S. Sargent, J. H. Nord, and J. Paliszkiewicz, "Internet of Things (IoT): From awareness to continued use," *International Journal of Information Management*, vol. 62, p. 102 442, Feb. 2022, ISSN: 0268-4012. DOI: 10.1016/J.IJINFOMGT.2021.102442.

[146] N. Kaloudi and L. I. Jingyue, "The AI-Based Cyber Threat Landscape," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, Feb. 2020, ISSN: 15577341. DOI: 10.1145/3372823. [Online]. Available: https://dl.acm.org/doi/10.1145/3372823.

[147] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," *Proceedings - IEEE Symposium on Security and Privacy*, vol. 2020-May, pp. 1314–1331, May 2020, ISSN: 10816011. DOI: 10.1109/SP40000.2020.00095.

[148] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A New Era in LLM Security: Exploring Security Concerns in Real-World LLM-based Systems," Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.18649v1.

[149] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining Zero-Shot Vulnerability Repair with Large Language Models," *Proceedings - IEEE Symposium on Security and Privacy*, vol. 2023-May, pp. 2339–2356, 2023, ISSN: 10816011. DOI: 10.1109/SP46215.2023.10179324.

[150] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The Emerging Threat of Ai-driven Cyber Attacks: A Review," *Applied Artificial Intelligence*, vol. 36, no. 1, Dec. 2022, ISSN: 10876545. DOI: 10.1080/08839514.2022.2037254. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/08839514.2022.2037254.

# Appendix

## Chapter 6

# Included papers