Bachelor's thesis

Bachelor's Programme in Computer Science

AI-powered social engineering

Riku Talvisto

January 30, 2025

Faculty of Science
University of Helsinki

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Faculty of Science	Bachelor's Programme in Computer Science

Tekijä — Författare — Author

Riku Talvisto

Työn nimi — Arbetets titel — Title

AI-powered social engineering

Ohjaajat — Handledare — Supervisors

Docent Lea Kutvonen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Bachelor's thesis	January 30, 2025	26 pages

Tiivistelmä — Referat — Abstract

Social engineering, a subdomain of cybersecurity, is the art and science of manipulating people into divulging confidential information or taking actions that may or may not be in their best interests. Traditionally, social engineering relied heavily on manual labor and human intuition, but with the advent of generative artificial intelligence (AI) technologies such as ChatGPT and deepfake media forgeries, cybercriminals are able to craft highly targeted and effective social engineering campaigns with novel, unexpected twists.

This thesis addresses how to protect organizations from social engineering attacks that are enhanced by generative AI technologies. To that end, this thesis explores the evolving landscape of AI in social engineering, focusing on attacks such as spear phishing aided by chatbots like ChatGPT and impersonation with hyper-realistic deepfake-generated forgeries. In contrast, the thesis also covers countermeasures against these attacks and discusses issues related to them based on relevant literature. Actualized incidents are briefly examined where appropriate.

The findings show that generative AI -powered social engineering attacks are more persuasive and effective than traditional methods, while current defenses are increasingly inadequate. This underscores the urgent need for cybersecurity professionals to revise their strategies and tools, with AI contributing to this defensive effort as well.

ACM Computing Classification System (CCS)
Social and professional topics → Computing / technology policy → Computer crime
→ **Social engineering attacks**
Security and privacy → Intrusion/anomaly detection and malware mitigation
→ **Social engineering attacks**

Avainsanat — Nyckelord — Keywords

social engineering, artificial intelligence, generative AI, cybersecurity, security, deepfake, deepfakes

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Contents

Tekoälyavusteinen käyttäjän manipulointi

Käyttäjän manipuloinnilla (*social engineering*) tarkoitetaan tietoturvan yhteydessä tietojärjestelmän loppukäyttäjään eli ihmiseen kohdistuvaa tietoturvahyökkäystä (Mitnick and Simon, 2003). Sen sijaan, että hyökkääjät etsisivät tietojärjestelmistä teknisiä haavoittuvuuksia, he kohdistavatkin hyökkäykset ihmiseen käyttäen hyväksi psykologisia menetelmiä (Wang et al., 2020).

Historiallisesti käyttäjän manipulointi on ollut riippuvainen ihmisen intuitiosta ja manuaalisesta työstä, mutta nyt moderni tekoäly (*artificial intelligence, AI*) on muuttamassa koko kenttää uusiksi. Tekoälyn avulla hyökkääjät pystyvät luomaan uskottavia tietojenkalasteluviestejä (*phishing*) sekä imitoimaan virallisia tahoja ja toimijoita realististen syväväärennösten (*deepfake*), kuten kuvien, äänen ja jopa videoiden, avulla.

Tässä kandidaatintutkielman suomenkielisessä lyhennelmässä käydään läpi tärkeimmät käyttäjän manipulointihyökkäykset sekä puolustuskeinot niihin.

Hyökkäykset ja työkalut

Tunnetuin käyttäjän manipulointihyökkäys on tietojenkalastelu (*phishing).* Tietojenkalastelu on petollista toimintaa missä hyökkääjä esiintyy luotettavana tahona tavoitteenaan saada käyttäjältä luottamuksellisia tietoja kuten hänen salasanansa tai luottokorttinsa numeron. Kohdennettu tietojenkalastelu (*spear phishing*) on kohdennettu tietylle käyttäjälle tai yritykselle, sisältäen jotain relevanttia tietoa kuten esimerkiksi käyttäjän nimen tai roolin yrityksessä.

OpenAI julkaisi vuonna 2022 ChatGPT:n joka mullisti tavan, jolla ihmiset käyttävät tekoälyäpalveluita. Se keräsi yli 100 miljoonaa käyttäjää ensimmäisen kahden kuukauden aikana*. ChatGPT on ns. generatiivinen tekoäly (*generative AI*) joka on koulutettu suurella määrällä tietoa ja joka pystyy tämän pohjalta luomaan uutta sisältöä, kuten tekstiä tai kuvia.

*https://explodingtopics.com/blog/chatgpt-users (luettu 2024-07-21)

OpenAI ja muut tekoälypalveluita varmistavat yritykset ovat asettaneet käyttöehtoja, joiden puitteissa palvelun käyttö on sallittua ja mahdollista. Rajoituksia käytölle on asetettu myös tekoälytoiminnallisuuksien sisälle. Hyökkääjät ovat kuitenkin onnistuneet valjastamaan ChatGPT:n kaltaiset suuret kielimallit (*large language model*) omiin tarkoituksiinsa ohittamalla nämä rajoitukset käyttäen esimerkiksi käänteistä psykologiaa.

ChatGPT ei esimerkiksi suoraan anna listaa sivustoista joilta voisi ladata laittomasti elokuvia, vaan sanoo että tämä toiminta on epäeettistä ja voi aiheuttaa käyttäjän tietokoneen saastumisen haittaohjelmilla (*malware*). Tällaiset rajoitukset on pystynyt ohittamaan useilla eri keinoilla, esimerkiksi sanomalla että suojellakseen käyttäjää haittaohjelmilta ChatGPT:n pitäisi kertoa sivustoista joille käyttäjän ei tule mennä. Näin käyttäjä saa haluamansa tiedot käyttäen käänteistä psykologiaa.

Näin hyökkääjät ovat pystyneet käyttämään suurten kielimallien tekoälytyökaluja tietojenkalasteluviestien laatimisessa, mikä on huomattaavasti parantanut niiden uskottavuutta.

Syvääärennökset (*deepfake*) ovat aidolta vaikuttavaa sisältöä kuten kuvia, ääntä tai videota, joka on luotu generatiivisen tekoälyn avulla. Syvävväärennöksiä voidaan käyttää esimerkiksi opetusmateriaalina, mutta niitä voidaan käyttää myös petollisiin tarkoituksiin. Syväväärennöksiä on jo onnistuneesti käytetty käyttäjän manipulointihyökkäysten perustana.

Puolustuskeinot

Puolustautuminen tekoälyavusteisia käyttäjän manipulointihyökkäyksiä vastaan on pitkälti samankaltaista kuin tavallisiakin (ei-tekoäly) hyökkäyksiä vastaan, muutamilla muutoksilla. Puolustautumiskeinot voidaan jakaa karkeasti kahteen, ihmislähtöisiin ja tekniikkalähtöisiin.

Perinteinen tapa suojata käyttäjää tietojenkalasteluviesteitlä on ollut sääntöpohjainen suodattaminen (*rule-based filtering*). Yksinkertaistettuna se vain tarkoittaa joukkoa loogisia sääntöjä joita seuraamalla voidaan päätellö onko viesti jollain todennäköisyydellä tietojenkalasteluviesti vai ei.

Sääntöpohjainen suodattaminen ei kuitenkaan toimi kovin hyvin tekoälyavusteista tietojenkalastelua vastaan. Tässä kohtaa tekoäly on valjastettu myös suojaamaan käyttäjää, eli siinä missä hyökkääjät käyttävät tekoälyä luodakseen kalasteluviestejä, puolustajat käyttävät sitä tunnistamaan näitä viestejä.

Historiallisesti ei ole ollut tarvetta tarkistaa saatujen kuvien tai videoiden aitoutta, mutta nyt syvävväärennösten aikakautena käyttäjä ei voi luottaa näkemänsä materiaalin aitouteen vaan lisävarmistuksia on tehtävä. Yksi tapa on käyttää tekoälypohjaisia palveluita syväväärennösten tunnistamiseen, samaan tapaan kuin sähköpostiviestienkin tarkistaminen.

Käyttäjälähtöiset tavat ovat tyypillisesti olleet käyttäjien kouluttaminen, simuloidut käyttäjän manipulointihyökkäykset, yrityksen tietoturva- ja tietosuojaohjeistusten laatiminen ja käytön valvonta, sekä tietoturva- ja tietosuojatietoisen yrityskulttuurin rakentaminen.

Tekoälypohjainen käyttäjän manipulointi tuo joitain muutoksia käyttäjälähtöisille puolustuskeinoille. Ensinnäkin käyttäjät eivät enää voi luottaa siihen, että hyvinkään kirjoitettu viesti ei olisi tietojenkalasteluviesti. Toiseksi kaikki saatu materiaali, kuten kuvat, äänitiedostot ja videot, saattavat olla syväväärennöksiä, vaikka käyttäjä ei itse pystyisi huomaamaan niissä mitään epätavallista.

Koska tekoälyohjelmat pystyvät automaattisesti etsimään Internetistä tietoa joita voisi käyttää osana käyttäjän manipulointihyökkäyksiä, myöskään viestit joissa on maininta joistain käyttäjälle relevanteista, ehkä jopa henkilökohtaisista asioista, ei voida enää varmuudella sanoa olevan aitoja.

Puolustuskeinojen arviointia

Tässä luvussa käydään arvointia puolustuskeinojen tehokkuudesta.

Tekoälyavusteisten hyökkäysten torjuminen pohjautuu hyvin paljon samoille tekniikoille jotka ovat jo käytössä: sisääntulevan viestinnän tarkistaminen, käyttäjien kouluttaminen, simuloidut hyökkäykset, yrityskulttuurin rakentaminen ja tietoturvaohjeistusten ylläito. Jokaiseen näihin kuitenkin on tehtävä muutoksia generatiivisen tekoälyn luoman uuden uhan vuoksi.

Yhteenveto

Vaikuttaa siis siltä että voimme olettaa tekoälyjärjestelmien nopean kehittymisen jatkuvan, tietoturvauhkien kehittymisen niiden mukana, ja tarpeen jatkuvalle käyttäjien kouluttamiselle ja uusien puolustuskeinojen löytämiselle kasvavan.

1 Introduction

Social engineering has emerged as a significant threat in the digital age, impacting individuals and organizations worldwide. As a subdomain of cybersecurity, social engineering is the art and science of manipulating people into revealing confidential information or performing actions that may or may not be in their best interests (Hadnagy, 2018). Rather than looking for technical vulnerabilities, social engineering relies on human interaction and exploits weaknesses in human psychology (Wang et al., 2020).

Traditionally, social engineering depended heavily on human intuition and manual effort to deceive its targets (Mitnick and Simon, 2003; Mirsky et al., 2023). However, with the advent of generative artificial intelligence (AI), the landscape of social engineering is undergoing a significant transformation, augmenting the sophistication and effectiveness of current and emerging attack methods (Fakhouri et al., 2024).

This thesis addresses how contemporary social engineering defensive countermeasures must be updated for the novel threats of generative AI in an organizational environment. To achieve this, the thesis examines the intersection of generative AI and social engineering, detailing how advanced AI tools amplify the execution and impact of these attacks while discussing and evaluating the necessary countermeasures.

Relevant social engineering attack vectors and tools are examined, including spear phishing with the help of chatbots like ChatGPT and impersonation using deepfake-generated content. Countermeasures discussed include AI-driven detection of spear phishing and deepfakes, user training, and relevant company policies, laws, and AI development and usage guidelines.

Contemporary countermeasures against social engineering attacks are ill-equipped to deal with the sophistication of AI-powered threats (Blauth et al., 2022; King et al., 2019). Cybersecurity professionals must thus update their tools and strategies, and AI can play a valuable role in this area as well. (Fakhouri et al., 2024; Tsinganos et al., 2018).

The rest of the thesis is structured as follows: Chapter 2 introduces social engineering, generative AI, and other essential concepts for further analysis. Chapter 3 examines relevant attack vectors and tools, including spear phishing and deepfake impersonation. Chapter 4 discusses both technological and human-oriented countermeasures against these attacks. The effectiveness and viability of these measures are assessed in Chapter 5. Chapter 6 summarizes key findings and implications for the future of social engineering defense.

2 Background

In recent years, the integration of generative artificial intelligence (AI) into social engineering offensive practices has emerged as a significant concern within the field of cybersecurity (Blauth et al., 2022; King et al., 2019; Mirsky et al., 2023). This chapter provides an overview of the role of generative AI in social engineering, explaining key concepts and terminologies essential for understanding the evolving threat landscape. After this, Chapter 3 examines generative AI-powered attack vectors and tools.

A strict consensus regarding the definition of a social engineering attack is lacking in the field (Hatfield, 2018). For the purposes of this thesis, social engineering is defined as "*a type of attack wherein the attacker(s) exploit human vulnerabilities by means of social interaction to breach cybersecurity, with or without the use of technical means and technical vulnerabilities*" (Wang et al., 2020).

Organizations today face cybersecurity threats from a range of sources, including cybercriminals, disgruntled employees, "script kiddies" (amateur hackers), hacktivists, competitors, and even state-sponsored cyber terrorists (Mirsky et al., 2023). These threat actors may be driven by motives such as financial gain, intellectual property theft, sabotage, fame, or revenge.

A total of 32 different AI capabilities have been identified that attackers could use against an organization (Mirsky et al., 2023). The top three most threatening categories are (1) social engineering, (2) information gathering, and (3) exploit development. Experts from both academia and industry ranked deepfake-based impersonation as the highest threat (Mirsky et al., 2023). Social engineering attacks are ranked the most threatening because these types of attacks are outside of the defender's control, are relatively easy to achieve, have high payoffs, are hard to prevent and cause the most harm.

Tracking incidents can be accomplished by counting occurrences or calculating the total cost of incidents annually. Not all organizations report their occurred social engineering and other cybercrime-related incidents, but some estimates of the prevalence of these attacks can be gained from data that is gathered by various public and private organizations and released in reports such as the FBI's Internet Crime Complaint Center's Internet Crime Reports and IBM's Cost of a Data Breach Reports. Organizations can thus assess the effectiveness and impact of their new policies, software upgrades, and cultural changes by monitoring changes in incident numbers and especially incident-related costs.

The dynamic nature of AI-driven social engineering poses a significant challenge for traditional cybersecurity frameworks, which often rely on static defenses and predefined patterns of attack (Fakhouri et al., 2024). As generative AI technologies advance, their application in crafting more convincing and personalized social engineering attacks becomes increasingly evident (Blauth et al., 2022). This new capability not only enhances the likelihood of success but also complicates the detection and mitigation of such threats (Mirsky et al., 2023).

Defense against AI-enhanced social engineering will thus require a multifaceted approach that combines technological innovation, user education, and a proactive stance and strict enforcement of cybersecurity policy (Blauth et al., 2022). As the landscape continues to evolve, staying ahead of these threats will necessitate ongoing research and collaboration across the cybersecurity community to develop effective countermeasures and best practices (Fakhouri et al., 2024).

The remainder of this chapter elaborates on essential concepts, specifically open-source intelligence and generative AI, which are essential for further analysis.

2.1 Open-source intelligence

In social engineering, publicly available information is known as **open-source intelligence** (Hadnagy, 2018). This involves collecting intelligence from sources that are publicly accessible, such as a target company's website, individuals' social media profiles, or other public records. Attackers are increasingly utilizing platforms such as LinkedIn, Facebook, and X (formerly Twitter) to gather information about their victims (Fakhouri et al., 2024).

Various online tools have been created for the purposes of gathering intelligence on an individual or an organization (Mirsky et al., 2023). They often offer automated forensic gathering and are able to visualize the found data, making it easier to identify exploitable patterns and connections.

2.2 Generative AI

Artificial intelligence (AI) encompasses the development of algorithms designed to automate complex tasks (Mirsky et al., 2023). Currently, the most prevalent type of AI is machine learning, which enables systems to enhance their performance as they gain experience. Deep learning, a subset of machine learning, employs extensive artificial neural networks as predictive models. The core idea behind AI is to enable machines to mimic human-like decision-making and thinking processes (Fakhouri et al., 2024).

When AI is used to generate content, it is called **generative AI** (Goodfellow et al., 2020). Unlike traditional AI, which follows programmed rules, generative AI utilizes machine learning to learn patterns from large training datasets to produce new outputs, such as text, images, audio and video.

Perhaps the most prominent example of generative AI is ChatGPT[*], a chatbot released by OpenAI in 2022. While far from being the first (Weizenbaum, 1966), this chatbot revolutionized how people use and interact with generative AI systems, reaching over 100 million users in just two months[†]. Built on the GPT (Generative Pre-trained Transformer) architecture, ChatGPT is designed to understand and generate human-like text by predicting the next word in a sequence.

Another relevant generative AI technology for social engineering is DALL-E[‡], also developed by OpenAI. This system generates images from textual descriptions, facilitating digital manipulation and the creation of misleading visuals. It enables the production of hyper-realistic images that can distort or shape public perception.

[*] https://openai.com/index/chatgpt (accessed 2024-08-19)
[†] https://explodingtopics.com/blog/chatgpt-users (accessed 2024-08-11)
[‡] https://openai.com/index/dall-e-3/ (accessed 2024-09-19)

3 Attack vectors and tools

This chapter reviews key social engineering attack vectors and tools relevant to the modern threat of generative AI. It first introduces pretexting and spear phishing, then explains how chatbots like ChatGPT could be manipulated, leading to impersonation attacks with deepfakes. After this, Chapter 4 goes over the countermeasures against these attacks.

3.1 Pretexting

Social engineering attacks typically begin with the gathering of open-source intelligence, which is subsequently used in conjunction with pretexting to attack an individual or an organization (Hadnagy, 2018). Pretexting involves fabricating a story or a scenario, a **pretext**, that is plausible but fraudulent, to engage the target with (Wang et al., 2020). With this story, the attacker hopes to gain the victim's trust by appearing legitimate.

Pretexting uses psychological manipulation, trust and relationship building, making it a potent tool for attackers (Mitnick and Simon, 2003). The attacker, often assuming the likeness and character of a legitimate entity such as a trusted colleague, an IT service worker, a government official, or a 3rd party service provider, creates a believable narrative story tailored to the target victim's context.

Humans possess advanced perceptual and decision-making capabilities shaped by lifelong experiences. Attackers can exploit these mental models by presenting deceptive information via pretexting (Mirsky et al., 2023).

3.2 Spear phishing

As the quintessential social engineering attack, **phishing** is characterized by malicious attempts to gain sensitive information from unaware users, traditionally via email and by using spoofed websites that look like their authentic counterparts (Basit et al., 2021). Phishing has been around since 1996, when cybercriminals began using deceptive emails and websites to steal AOL (America Online) account information from unsuspecting users (Wang et al., 2020).

Spear phishing, on the other hand, is a more targeted version of phishing, where attackers customize their deceptive messages to a target individual or organization (Fakhouri et al., 2024). Spear phishing that is targeted at high-profile individuals is called **whaling**.

Unlike generic phishing attempts, spear phishing involves gathering detailed information about the victim, via open-source intelligence or otherwise, such as their name, position, and contacts to craft a convincing and personalized message (Hadnagy, 2018). This tailored approach increases the likelihood of the victim falling for the phishing attempt, but has traditionally been a lot more time and energy consuming (Mirsky et al., 2023).

3.3 Chatbots like ChatGPT

Malicious actors can use generative AI **chatbots** such as ChatGPT in their social engineering schemes, but due to the manufacturer's set limits, some workarounds may need to be used (Gupta et al., 2023). For instance, when asking ChatGPT to provide links to websites that provide pirated content such as movies results in the chatbot denying the request, stating that downloading pirated content is unethical and may also lead to the user's computer being infected with malware.

However, regular users and scholars have found a number of ways to bypass ChatGPT's inherent ethical and behavioral guidelines, such as by using reverse psychology*. In the above example, instead of directly asking for links to the pirate websites, the user can say that because he does not want his computer to be infected by malware, ChatGPT should provide links to sites the user should avoid visiting, thus causing ChatGPT to reveal the content the user originally wanted.

ChatGPT can effectively translate text from the attacker's native language to the victim's, maintaining fidelity and correcting any spelling or grammatical errors. It can even enhance the deceptive message, provided that the models' ethical restrictions have been bypassed successfully (Gupta et al., 2023). Phishing messages have historically been marked by noticeable spelling and grammatical errors (Herley, 2009), and people have traditionally been advised to look out for these errors as a hallmark of a phishing message.

Chatbots like ChatGPT can also integrate any gathered intelligence into spam messages, enhancing their relevance. Additionally, incorporating deepfake content, such as a video of the company's CEO issuing demands, can further increase the effectiveness of spear

*https://incidentdatabase.ai/cite/420 (accessed 2024-07-15)

phishing attempts.

3.4 Impersonation with deepfakes

Deepfake, a portmanteau of "deep learning", a type of machine learning, and "fake", is technology which uses artificial neural networks to create highly convincing fake media, either by altering existing content or creating them from scratch (Mirsky and Lee, 2021). When existing content is being altered, it's called reenactment or replacement, and when entirely new content is created, it's called synthesis.

Deepfake content can be images, audio, and even full-resolution video. These hyper-realistic forgeries can depict a person saying or doing things that didn't actually happen, making it difficult for people and even AI systems to discern what is real and what is fake (Blauth et al., 2022).

By utilizing deepfake-generated content, deepfakes, attackers can convincingly impersonate trusted individuals or organizations, enhancing the credibility and even the emotional impact of their deceptive social engineering strategies (Mirsky and Lee, 2021). In 2021, complete facial reenactment, such as pose, gaze, blinking, mouth, and movements, was achieved with only a minute of training video, suggesting that if a malicious actor wants to reenact an individual, they do not need to gather a lot of video material for this. If video material is not available, attackers might be able to resort to filming the target person exiting the company's premises.

Deepfake technology has advanced within just two years to the point where reenactment can be done in real-time with training requiring only a few images or seconds of audio from the victim (Mirsky et al., 2023), while higher quality deepfakes still require more audio/video data. This was evident in a 2024 incident where deepfake technology was used in a live video conference to successfully scam an organization for $25 million*.

3.5 Phishing with voice, vishing

Phishing that is done using voice is called **vishing** (Doan et al., 2023). By utilizing traditional phone systems or VoIP (Voice-over-IP), the attacker calls the victim with a pretext to manipulate them into revealing sensitive information or performing actions that

*https://incidentdatabase.ai/cite/634 (accessed 2024-08-24)

may or may not be in their best interests (Hadnagy, 2018).

With real-time voice morphing, a type of deepfake natural speech synthesis, the attacker can effectively and realistically impersonate someone else (Doan et al., 2023). This technology converts the attacker's own voice (as input) to the chosen person's voice (as output) automatically during the call. It's hard for the human auditory system to distinguish between real and fake voice samples, especially through voice calls.

The deepfake model has to be trained before it can be used. This is done using audio, which can be sourced from places like YouTube, a company website, or by calling the person the attacker wants to mimic the voice of and recording the conversation.

Social engineering with real-time voice morphing of employees' voices has been found to be one of the top threats posed by AI to organizations (Mirsky et al., 2023). The first significant incident occurred back in 2019, where attackers successfully used deepfake-generated voice during a call to impersonate an authentic entity for monetary gains exceeding 200,000 €*.

*https://incidentdatabase.ai/cite/200 (accessed 2024-05-13)

4 Countermeasures

In this chapter, countermeasures against the attacks covered in the previous chapter are examined. This chapter focuses on two parts: technology-oriented countermeasures such as phishing and deepfake detection mechanisms, and user-oriented countermeasures such as personnel training programs, company policies, and laws and guidelines. Technology-oriented countermeasures are examined first since the human-oriented measures rely and build upon them. Chapter 5 discusses and evaluates these countermeasures in detecting and preventing social engineering attacks.

Traditionally, defense against social engineering relied on human education and awareness campaigns (Fakhouri et al., 2024). This reliance, despite its many merits, has revealed its fragility, as even the best trained user can fail to detect a social engineering attack and fall victim amidst the myriad digital threats. Defense against generative AI -based social engineering thus requires a multifaceted approach, incorporating both technical and human-oriented measures.

4.1 AI-based phishing detection

AI systems learn, evolve and adapt based on the datasets that they are processing, and thus continuously refining their operational methods and predictions (Fakhouri et al., 2024), rather than relying on pre-defined and rigid algorithms. This presents a paradigm shift in how computers perceive, then process and finally respond to data.

Machine learning models are trained with vast datasets containing both safe and malicious samples of e.g. phishing attempts and phishing URL's. Given time and further training, these models learn to identify patterns, behaviors and anomalies, which means they are very capable of detecting threats, including the novel and unseen (Fakhouri et al., 2024). Including AI in cybersecurity measures thus doesn't mean just adding another tool for cybersecurity, but fundamentally defining anew the foundations of our digital defenses. Thus, AI systems are more attuned to the constantly evolving threat landscape.

Traditional phishing message detection systems, i.e. those not based on machine learning and AI, are typically rule-based and signature-based, which often falter when faced with novel or evolved threats like those enhanced by AI.(Fakhouri et al., 2024). These defenses often leave systems vulnerable to novel, uncharted attacks.

Using techniques such as natural language processing, AI systems can be trained to recognize common patterns and especially anomalies in communications to and from the network that are indicative of phishing attempts (Basit et al., 2021). These systems can flag suspicious emails or messages by analyzing factors such as unusual use of language, unexpected requests for private data, or other inconsistencies.

Just as incoming and outgoing email messages are analyzed for phishing attacks, and the attachments are scanned for malware such as viruses or Trojan horses, images, audio and videos need to be scanned as well to aid the user in detecting if they are genuine or deepfakes, or given the possibility for a scan (Mirsky and Lee, 2021).

Modern phishing attacks leverage advanced AI techniques to create highly convincing fake websites and emails that mimic legitimate entities, making it increasingly difficult for users to distinguish between authentic and malicious content. To counter these sophisticated phishing attacks, researches have developed various AI-enabled detection techniques, including machine learning, deep learning, hybrid learning and scenario-based approaches (Basit et al., 2021). These methods have shown great promise in identifying phishing attempts with high accuracy, often surpassing traditional detection methods.

Machine learning, for instance, combats phishing by analyzing massive amounts of data to identify patterns and features typical of phishing attempts. By training models on datasets containing both legitimate and phishing emails or websites, machine learning algorithms can learn to distinguish between the two demonstrating over 95 % accuracy compared to traditional, non-AI based methods. However, care has to be taken when choosing the datasets.

Where once experts in the field could recommend that a caller be authenticated by recognizing their voice, accent, and intonations (Mitnick and Simon, 2003), with the advent of generative AI, and especially deepfakes, this no longer holds true (Doan et al., 2023). Technologies such as the BTS-E encoder have been proposed for detecting idiosyncrasies in speech that might not or even could not be consciously registered by human observers. BTS-E detects correlations between breathing, talking, and silence to detect spoofed audio.

4.2 AI-based deepfake detection

Deepfakes often contain subtle anomalies called artifacts, just as image forgeries of the past did. Deepfake detection procedures are primarily based on machine learning and forensic analysis, attempting to identify these specific artifacts in the multimedia content (Mirsky and Lee, 2021). The artifacts can be subtle, such as a strange blob of pixels, or overt such as a person having clearly warped eyes.

Seven different types of artifacts have been identified in two main categories (Mirsky and Lee, 2021): spatial-type artifacts cover blending, environment and forensics, while temporal-type artifacts cover behavior, physiology, synchronization and coherence.

Blending artifacts occur when the generated content is integrated back into a frame (the background), which is detectable with techniques such as edge detection and frequency analysis. Environment artifacts can appear when fake facial content seems inconsistent with the surrounding background frame, often due to mismatches in warping, lighting or fidelity. Forensic-type artifacts are residues from the generative models, such as generative adversarial network fingerprints or sensor noise.

Behavior-type artifacts involve monitoring anomalies in the target's mannerisms, while physiological artifacts focus on inconsistencies in natural biological cues like blinking of the eyes or head movements. Synchronization artifacts can be observed in mismatched audio-visual elements, and coherence artifacts relate to inconsistencies in logical sequences happening over time.

In regards to cybersecurity, machine learning models are trained with vast datasets that encompass both benign and malicious activities (Fakhouri et al., 2024). Over time, the models "learn" to notice patterns, anomalies and behavior which makes them exceptionally good at detecting threats, even completely novel ones.

4.3 Laws and usage guidelines

Building and maintaining guidelines for the ethical use of AI systems has been at the forefront of their development. OpenAI, the organization behind the GPT architecture and its publicly accessible front-end ChatGPT, has made strides in an attempt to prevent the misuse of their AI systems.

The ethical use of AI contributes significantly to mitigating risks (Gupta et al., 2023), with AI developers such as OpenAI implementing guidelines* to limit the misuse of their AI systems, such as ChatGPT. Despite these efforts the complete prevention of AI system misuse remains elusive, particularly since older versions without the latest restrictions might still be accessible.

Guidelines are also being developed at national and global levels, where they can take the form of a law. For instance, the European Union's General Data Protection Regulation (GDPR), and its relationship with AI, including AI-powered social engineering, is a complex and evolving topic. Introduced in 2018, GDPR and its development predates the widespread emergence of technologies such as generative adversarial networks and generative AI (Goodfellow et al., 2020), and thus was not specially designed to address these issues.

EU's European Parliamentary Research Service (EPRS) released a study (European Parliamentary Research Service, 2020) which eventually lead to the formal approvement by the European Parliament on Feb 13, 2024. As of the writing of this thesis, the AI Act is not yet in effect, and once it is published, there will be a transition period before the act is fully enabled. This act is considered a landmark regulation, as it is the first comprehensive AI law in any major jurisdiction around the world, paving the wave for other jurisdictions, such as the US, to follow suite.

4.4 User training and company policy

User-oriented countermeasures against social engineering attacks usually fall into four broader categories (Tsinganos et al., 2018; Mitnick and Simon, 2003). These categories are simulated penetration tests with social engineering techniques, employee security awareness training programs, creation and application of corporate cybersecurity policies, and the development of a security-conscious company culture.

Regular and comprehensive training programs are vital to educate employees about social engineering tactics. Regularity is stressed by experts in the field as users tend to forget what they have learned (Hadnagy, 2018; Mitnick and Simon, 2003). It is thus suggested that training against social engineering attacks is not something that is done annually, or even bi-annually, but rather that it's something that is baked into the company's culture. The inoculation theory (Blauth et al., 2022) suggests that prior exposure to social

*https://openai.com/policies/usage-policies (accessed 2024-08-22) engineering attacks could help protect users against future threats.

Conducting simulated social engineering and phishing attack campaigns, via numerous channels such as email, SMS and even phone/VoIP, allows organizations to assess the susceptibility of their employees to social engineering tactics (Hadnagy, 2018). These exercises help identify vulnerabilities in the workforce, enabling further targeted training and reinforcing the importance of scrutinizing unsolicited communication. With the advent of generative AI and deepfakes, this needs to be extended to cover any and all communication (Mirsky and Lee, 2021).

Employees should be shown what different varieties of deepfake content look like, as well as how easy it is to doctor them.

Feedback from these simulations can significantly aid personnel development. However, employees who fall victim to these simulated attacks should be re-educated rather than punished (Mitnick and Simon, 2003). Furthermore, it is essential to inform employees in advance that such campaigns may be run occasionally. This approach not only keeps them vigilant but also mitigates negative feelings associated with "being tricked" by their own company (Hadnagy, 2018).

A company culture that is open about sharing if any of its members fall victim to social engineering attacks is more robust due to employees not having to feel shame or hide the fact that they got tricked (Hadnagy, 2018). This can be reinforced by executives talking openly about times when they fell victim, to what kind of an attack and why, and what they did about the incident. It's always better that employees report suspected or actualized social engineering attacks rather than trying to hide them for fear of ridicule or punishment (Mitnick and Simon, 2003).

It's imperative that every user understands that they are the weakest link in the cybersecurity chain (Mitnick and Simon, 2003) and that the responsibility of the organization's cybersecurity is in everyone's hands, not just the cybersecurity professional's.

Finally, because AI can source social media sites and the Internet automatically for open-source intelligence, it's imperative for people to know to be careful of what they share, with whom and when (Mitnick and Simon, 2003). Even seemingly private or coincidental information, such as photos indicative that the employee is now on a company picnic, could be used against them and their employer in a social engineering attack.

5 Discussion

This chapter evaluates current countermeasures and their effectiveness at detecting and preventing social engineering attacks, particularly those enhanced by generative AI technologies. The landscape of cybersecurity is continuously evolving, and traditional countermeasures such as email filtering and user awareness programs, although still crucial, are increasingly insufficient against the sophistication of AI-powered threats (Fakhouri et al., 2024). While current countermeasures provide a baseline defense against social engineering attacks, this evaluation reveals a critical gap between existing strategies and the rapidly evolving sophistication of generative AI -powered attacks. After this, Chapter 6 concludes the thesis.

According to IBM's 2024 Cost of a Data Breach Report, organizations are increasingly leveraging AI and automation in their security operations, with 31% of studied organizations deploying these technologies extensively, 36% reporting limited use and the remaining 33% reporting no use (IBM, 2024). Notably, when AI was extensively deployed in prevention workflows, organizations saw an average breach cost reduction of 45% ($2.2 million compared to the average of $4,88 million). The key finding of this year's report is a striking correlation: the more an organization relied on AI, the lower its average breach costs were.

5.1 Generative AI and deepfakes

Where previously an employee could authenticate a caller by recognizing their voice, intonations, and accents (Mitnick and Simon, 2003), today this is not enough due to the prevalence of deepfake-generated content.

Just as spam filters are inclined to report false positives (Fakhouri et al., 2024), so too are deepfake detection systems (Mirsky and Lee, 2021), filtering legitimate communications causing operational hiccups and perhaps lost business engagements.

Technological solutions like phishing detection systems that utilize natural language processing and machine learning show potential in identifying anomalous communications (Basit et al., 2021). However, these systems are being challenged by the ever-improving quality of AI-generated content such as spear phishing messages, which often mimic human interaction and presentation with higher and higher fidelity. Similarly, tools designed to detect deepfakes are in their early stages (Mirsky and Lee, 2021), and face significant hurdles in keeping up with the rapid advancements in AI technologies that create such content.

Just as people have differing propensities for detecting phishing attempts and noticing subtle anomalies in spelling and grammar (Nicholson et al., 2020; Neupane et al., 2018), so too are people variously adept at spotting these anomalies in deepfakes. Part of the solution regarding deepfake content is to raise population awareness about such technology use (Blauth et al., 2022). In 2019, the Democratic Party (USA) presented a deepfake video of their own chairman to highlight their concerns about deepfake content*.

virus detection signatures are developed by their respective companies, and cybersecurity personnel must be trained regularly. However, AI makes a difference here because AI systems can learn from other AI systems. Where one network is the target of a novel type of cybersecurity threat, and once its detected, this AI system can inform other systems in the same "network", thus bolstering defenses on a possibly global scale?

Spreading information about deepfakes to the public faces the hurdle of the "liar's dividend", a situation where a "liar" discredits a real video claiming it to be a deepfake. The more users are aware of deepfake content and the ability of AI to doctor and create videos, the more skeptical they will be, causing them to question images and videos that are real (Blauth et al., 2022). Deepfakes may thus erode the public's very trust in multimedia content.

5.2 User-centric

Human-oriented measures remain pivotal in the defense against social engineering. Regular training programs are crucial for equipping end-users with the knowledge to recognize potential threats (Hadnagy, 2018). This holds true especially because AI technologies are evolving rapidly on both the offensive and defensive sides, leading to a situation where the attackers are one step ahead of the defenders and automated AI-based social engineering detection and prevention systems fail to protect the user (Fakhouri et al., 2024). Thus comprehensive, regular and innovative user training and awareness programs can never be outlooked, as the human is the weakest link in the cybersecurity chain (Mitnick and

*https://edition.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon/index.html (accessed 2024-08-25)

Simon, 2003).

The deployment of simulated social engineering campaigns offers substantial insights into employee vulnerability, yet these must be meticulously crafted to avoid adverse impacts on workplace morale (Mitnick and Simon, 2003). Utilizing natural language processing to craft highly convincing but simulated phishing messages to be sent to the employees can further aid in the detection of the need for further training, with open-source intelligence being incorporated also.

Certain parts of the population, such as teenagers and young people who haven't yet gained enough experience on the Internet, may be more susceptible to social engineering attacks (Nicholson et al., 2020). People on the autism spectrum, often facing challenges in social interaction, may unexpectedly excel at detecting social engineering attacks (Neupane et al., 2018). It is thus suggested that training efforts, while they must be targeted at everyone, would take into account any potential differences in demographics. Chatbots like ChatGPT can help in designing tailored and engaging training content.

5.3 Law and ethics

AI excels in detecting subtle patterns and anomalies which might elude more conventional systems (Fakhouri et al., 2024). This capability exceeds mere threat recognition and covers concepts such as anticipation of future potential vulnerabilities based on real-time and also historical data, which helps ensure defensive measures are not just reactive but predictive (proactive).

It seems evident that the highly dynamic nature of AI technologies fuel a continuous arms race between attackers and defenders, causing many countermeasures to become obsolete quickly (Fakhouri et al., 2024). Thus, protecting against AI-powered attacks requires not a single solution but an integrated approach that is baked in the company culture, that combines technological defenses, comprehensive and continuous user education, and robust organizational policies.

As AI is developed further and the more its availability increases, the risk or malicious or criminal use increases as well, and these risks if not properly addressed may lead to the excessive strict regulation of AI technologies (King et al., 2019).

To summarize the evaluation of countermeasures against AI-powered social engineering, while they currently provide a fundamental level of defense, they struggle to keep up with the rapidly evolving AI-powered social engineering tactics. The limited effectiveness of these measures is attributable to both the fast-paced development in AI and the inherent human factor, being the weakest link within cybersecurity (Mitnick and Simon, 2003). Therefore, continuous innovation in both technological solutions, such as AI-based phishing and deepfake detection algorithms (Mirsky and Lee, 2021), and human-centric strategies, such as awareness programs and simulated spear phishing campaigns (Hadnagy, 2018), is truly imperative for an organization to adapt to and counteract the advancing generative AI -powered threat landscape.

In 2024, the state of Tennessee enacted the ELVIS (Ensuring Likeness Voice and Image Security) Act*, protecting artists from the use of their voice via deepfake technologies. Further legislation need to address the use of deepfakes in other ways, such as in social engineering. EU's AI Act explicitly prohibits the use of AI for human manipulation and social engineering.

Because regulatory frameworks and other governance mechanisms might not be developed at the same pace of technological advancements, proactivity is vital to reduce the risks (Blauth et al., 2022). The faster the potential for AI misuse is understood, the earlier potential preventive, mitigative, disincentivizing and redressing policies may be applied (King et al., 2019).

*https://aibusiness.com/responsible-ai/tennessee-enacts-elvis-act-to-protect-artist-voices-from-ai-misuse (accessed 2024-08-24)

6 Conclusions

The subfield of social engineering within cybersecurity is undergoing a significant transformation with the advent of generative AI (Fakhouri et al., 2024). This thesis explored how generative AI empowers malicious actors in this space and how current countermeasures in an organizational environment need to be updated to reflect this evolving threat landscape.

Generative AI is revolutionizing social engineering attacks, enabling attackers to use sophisticated tactics like spear phishing (Basit et al., 2021), impersonation with deepfake content (Mirsky and Lee, 2021) and voice phishing, vishing, with real-time voice morphing (Doan et al., 2023). These advancements reveal that traditional countermeasures are becoming increasingly ineffective, requiring a comprehensive re-evaluation of current strategies and tactics.

Where previously, an employee could authenticate a caller by recognizing their voice, intonations, and accent (Mitnick and Simon, 2003), today, this is not enough. User training and awareness programs must be updated to address the novel threat of AI in social engineering.

AI can help detect social engineering attacks, but it does not eliminate the necessity for user training and awareness programs. On the contrary, as AI-powered attacks proliferate, the need for awareness and vigilance will grow even higher (Fakhouri et al., 2024). Chatbots like ChatGPT can help develop more robust security guidelines and design more engaging social engineering awareness programs.

One area not addressed in this thesis, but deserving of future research, is the potential for AI to automate social engineering attacks (Mirsky et al., 2023). Currently, AI technology lacks the sophistication needed to develop fully autonomous agents capable of executing such attacks without human oversight.

Experts from both academia and industry should concentrate their efforts on deterring the top threats organizations face from AI, namely social engineering powered by AI and impersonation with deepfakes (Mirsky et al., 2023).

IBM's 2023 Cost of a Data Breach report revealed that organizations using AI to address cybersecurity threats experienced an average of 45% reduction in annual incident-related costs compared to those that did not. Further, they found that a correlation exists, indicating that increased reliance on AI corresponded with lower incident costs.

What seems certain is that we can count on the rapid development of AI technologies continuing, generative AI -powered social engineering attacks evolving with them, and the need for continuous, innovative user training growing in the future as well as the need for the development of AI-based mitigation and prevention technologies. Organizations need to utilize AI to combat generative AI -powered social engineering.

Bibliography

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., and Kifayat, K. (2021). "A comprehensive survey of AI-enabled phishing attacks detection techniques". In: *Telecommunication Systems*, 76(1), pp. 139–154. DOI: 10.1007/s11235-020-00733-2.

Blauth, T. F., Gstrein, O. J., and Zwitter, A. (2022). "Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI". In: *IEEE Access*, 10, pp. 77110–77122. DOI: 10.1109/ACCESS.2022.3191790.

Doan, T.-P., Nguyen-Vu, L., Jung, S., and Hong, K. (2023). "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095927.

European Parliamentary Research Service (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*. Tech. rep. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf (visited on 08/22/2024).

Fakhouri, H. N., Alhadidi, B., Omar, K., Makhadmeh, S. N., Hamad, F., and Halalsheh, N. Z. (2024). "AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response". In: *2024 2nd International Conference on Cyber Resilience (ICCR)*. Dubai, United Arab Emirates, pp. 1–8. DOI: 10.1109/ICCR61006.2024.10533010.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). "Generative adversarial networks". In: *Communications of the ACM*, 63(11), pp. 139–144. DOI: 10.1145/3422622.

Gupta, M., Akiri, C., Aryal, K., Parker, E., and Praharaj, L. (2023). "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy". In: *IEEE Access*, 11, pp. 80218–80245. DOI: 10.1109/ACCESS.2023.3300381.

Hadnagy, C. (2018). *Social Engineering: The Science of Human Hacking*. John Wiley & Sons. ISBN: 978-1-119-43338-5.

Hatfield, J. M. (2018). "Social engineering in cybersecurity: The evolution of a concept". In: *Computers & Security*, 73, pp. 102–113. DOI: 10.1016/j.cose.2017.10.008.

Herley, C. (2009). "So long, and no thanks for the externalities: the rational rejection of security advice by users". In: *Proceedings of the 2009 workshop on New security* *paradigms workshop*. NSPW '09. New York, NY, USA: Association for Computing Machinery, pp. 133–144. DOI: 10.1145/1719030.1719050.

IBM (2024). *Cost of a Data Breach Report 2024*. URL: https://www.ibm.com/reports/data-breach (visited on 08/07/2024).

King, T. C., Aggarwal, N., Taddeo, M., and Floridi, L. (2019). "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". In: *Science and Engineering Ethics*, 26(1), pp. 89–120. DOI: 10.1007/s11948-018-00081-0.

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., and Biggio, B. (2023). "The Threat of Offensive AI to Organizations". In: *Computers & Security*, 124, p. 103006. DOI: 10.1016/j.cose.2022.103006.

Mirsky, Y. and Lee, W. (2021). "The Creation and Detection of Deepfakes: A Survey". In: *ACM Comput. Surv.*, 54(1), 7:1–7:41. DOI: 10.1145/3425780.

Mitnick, K. D. and Simon, W. L. (2003). *The Art of Deception: Controlling the Human Element of Security*. John Wiley & Sons. ISBN: 978-0-7645-4280-0.

Neupane, A., Satvat, K., Saxena, N., Stavrinos, D., and Bishop, H. J. (2018). "Do Social Disorders Facilitate Social Engineering? A Case Study of Autism and Phishing Attacks". In: *Proceedings of the 34th Annual Computer Security Applications Conference*. New York, NY, USA: Association for Computing Machinery, pp. 467–477. DOI: 10.1145/3274694.3274730.

Nicholson, J., Javed, Y., Dixon, M., Coventry, L., Ajayi, O. D., and Anderson, P. (2020). "Investigating Teenagers' Ability to Detect Phishing Messages". In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. Genoa, Italy, pp. 140–149. DOI: 10.1109/EuroSPW51379.2020.00027.

Tsinganos, N., Sakellariou, G., Fouliras, P., and Mavridis, I. (2018). "Towards an Automated Recognition System for Chat-based Social Engineering Attacks in Enterprise Environments". In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*. New York, NY, USA: Association for Computing Machinery, pp. 1–10. DOI: 10.1145/3230833.3233277.

Wang, Z., Sun, L., and Zhu, H. (2020). "Defining Social Engineering in Cybersecurity". In: *IEEE Access*, 8, pp. 85094–85115. DOI: 10.1109/ACCESS.2020.2992807.

Weizenbaum, J. (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM*, 9(1), pp. 36–45. DOI: 10.1145/365153.365168.

AI Declaration

I hereby state all of the use cases where I have utilized advanced AI technologies during the research and writing processes of this thesis.

Table 6.1 lists all of the AI tools that I have used and their use scenarios.

Tool	**Use cases**
Large language models	Finding synonyms for words. Generating LaTeX code for tables and images. Brainstorming what the topic for my thesis could be. Finding related keywords. OCR from handwritten text.
Writefull	Correcting spelling errors on Overleaf when prompted.
Keenious	Finding relevant research articles based on released literature and also my own, unfinished work.

Table 6.1: AI tools used during the writing of this thesis.

Large language models used: GPT-3.5, GPT-4 (4o & mini), Claude 3.5 Haiku & Sonnet, Gemini 1.5 Flash & Pro, Llama-3