Bachelor's thesis

Bachelor's Programme in Computer Science

[DRAFT] AI-powered Social Engineering

Riku Talvisto

September 5, 2024

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

[DRAFT] 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	September 5, 2024	25 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 modern artificial intelligence (AI) technologies, cybercriminals are able to craft highly targeted and effective social engineering campaigns with novel, unexpected twists.

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 evaluates their effectiveness based on relevant literature. Actualized incidents are briefly examined where appropriate.

The results indicate that AI-powered social engineering attacks are more convincing and successful than traditional attacks and that contemporary countermeasures against them are becoming increasingly ineffective. This highlights the urgent need for cybersecurity professionals to update their strategies and tools for cyber defense against the emerging threat of AI, as social engineering is a threat to every organization and every individual.

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, AI, cybersecurity, security, hacking, deepfake

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Thesis includes a summary in Finnish and a declaration regarding the author's use of AI technologies.

Contents

Tekoälyavusteinen käyttäjän manipulointi

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

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

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.

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

Puolustuskeinot

Puolustautuminen tekoälyavusteisia käyttäjän manipulointihyökkäyksiä vastaan on pitkälti samankaltaista kuin tavallisiakin 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 syväväärennösten 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

The widespread adoption of interconnected IT devices and services has transformed every aspect of human life, from personal communication to business operations, and this reliance seems to be ever-expanding. The digital revolution has created numerous opportunities but has also introduced significant vulnerabilities. Among these, social engineering poses a particularly grave threat to security and privacy.

Social engineering is the art and science of manipulating victims into divulging 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).

Historically, social engineering relied heavily on human intuition and manual effort to deceive targets (Mitnick and Simon, 2003). With the advent of modern 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 explores the intersection of AI and social engineering and how contemporary AI technologies enhance the execution and impact of these attacks, and discusses the necessary actions to counter such advanced attacks. Several social engineering attack vectors and tools particularly relevant to the modern threat of AI were selected for in-depth analysis: spear phishing with the help of chatbots like ChatGPT, impersonation through deepfake-generated images, audio, and video, and the automated gathering and processing of intelligence data.

Based on a literature review and analysis of occurred social engineering incidents, the results indicate that contemporary countermeasures against social engineering attacks are ill-equipped to deal with the sophistication of AI-powered threats (Fakhouri et al., 2024; Blauth et al., 2022). Thus, the urgent need for cybersecurity professionals to update their tools and strategies is evident; however, AI can also play a valuable role in this area.

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 then discusses both technological and human-oriented countermeasures against these attacks. The effectiveness and viability of these measures are then assessed in Chapter 5. Lastly, Chapter 6 summarizes key findings and implications for the future of social engineering defense.

2 Social Engineering and AI

This chapter gives an overview of what social engineering constitutes, provides brief historical context and describes some key terminology that is necessary for further analysis, including about AI. After this, Chapter 3 examines AI-powered attack methods and tools.

The term *social engineering* dates back to 1842, when it was used to describe centralized planning in an attempt to manage the future development and behavior of a society (Hatfield, 2018). Since then, its use has shifted to the field of cybersecurity through the phone phreaking phase (late 1950s to early 1970s) and through to the contemporary hacker culture (Wang et al., 2020).

As one of the earliest hackers and social engineers, the phreakers, used impersonation to call the Bell Telephone company in order to gain insider information about the telephone networks in order to carry out further attacks without the need for social manipulation (Hatfield, 2018), modern hackers view social engineering not as something to be replaced but a key part of any hacker's toolkit, in fact perhaps the most important one (Mitnick and Simon, 2003).

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).

Some key concepts, namely open-source intelligence, pretexting, and generative AI, are explained next.

2.1 Open-source intelligence

In social engineering, publicly available information is referred to as **open-source intelligence**. Like the name implies, it involves gathering of intelligence data from publicly locatable sources, such as from the target company's website, or from the social networking profiles of an individual or from other public records.

Various online tools exist for the purposes of gathering intelligence on an individual or an

organization, the most famous of which in 2024 is perhaps Maltego. It offers automated forensic gathering and visualizes the found data, making it easier to identify patterns and connections.

Social engineering attacks typically begin with the gathering of open-source intelligence, which are subsequently used in conjunction with pretexting to attack an individual or an organization.

2.2 Pretexting

Pretexting involves fabricating a story or a scenario, a **pretext**, that is plausible but fraudulent, to engage the target with (Conteh and Schmick, 2016). With this story, the attacker hopes to gain the victim's trust by appearing legitimate. This type of attack relies heavily on the gathered open-source intelligence in assisting with the creation of the story (Hadnagy, 2018).

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.

2.3 Generative AI

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 (Fakhouri et al., 2024).

A key 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 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.

*https://openai.com/index/chatgpt (accessed 2024-08-19)

†https://explodingtopics.com/blog/chatgpt-users (accessed 2024-08-11)

It utilizes natural language processing (NLP), a domain at the intersection of human language and computation. Another prominent form of generative AI is OpenAI's DALL-E project. It understands human written text and generates images based on the user's prompt.

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 explores the misuse of chatbots like ChatGPT for malicious content generation, followed by an investigation of automated intelligence-gathering processes. The discussion then covers deepfake-generated media that can be used for impersonation, concluding with how attackers could use all of this with spear phishing. After this, Chapter 4 goes over the countermeasures against these attacks.

3.1 ChatGPT and other chatbots

Malicious actors can use generative AI **chatbots** such as ChatGPT in their 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 ethic 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 doesn't 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.

3.2 Automated intelligence processing

Automated intelligence processing has emerged as a pivotal element in contemporary information dissemination processes and thus also in strategic decision-making. By utilizing sophisticated algorithms and machine learning techniques, threat actors can efficiently collect, process and analyze vast amounts of data from various source, including

[†]https://incidentdatabase.ai/cite/420 (accessed 2024-07-15)

social media, news outlets and public databases (Bilge et al., 2009). This process not only streamlines the acquisition of relevant data but also enhances the accuracy and timeliness of insights derived from the data.

These processes and tools can be used with open-source intelligence to gather exploitable data about the victim or the organization, or to analyze data that has been gathered via previous successful social engineering attacks. Even seemingly inconsequential bits of data could prove invaluable to the attacker (Mitnick and Simon, 2003), and AI can aid in the discovery of these insights (Blauth et al., 2022).

One of the most prominent examples of user's data being used without their consent was the Cambridge Analytica scandal (Blauth et al., 2022). The firm used the harvested data to create detailed voter profiles and targeted political advertisements, aiming to influence voter opinions and behaviors.

3.3 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 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 strategies (Mirsky and Lee, 2021). Complete facial reenactment (pose, gaze, blinking, mouth etc) was achieved with only one minute of training video, suggesting that if a malicious actor wants to reenact an individual, they don't need to gather a lot of video material for this. If video material isn't available, attackers can resort to filming the target person exiting the company's premises.

In 2024, deepfake technology was used in a video conference* to successfully scam an organization for $25 million.

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

3.4 Spear phishing

As the quintessential social engineering attack, **phishing** is characterized by malicious attempts to gain sensitive information from unaware users, usually 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 (Basit et al., 2021; 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 (Salahdine and Kaabouch, 2019). 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.

Phishing messages have traditionally been marked by noticeable spelling and grammatical errors. ChatGPT can effectively translate text from the attacker's native language to the victim's, maintaining fidelity and even enhancing the deceptive message, provided that the models' ethical restrictions have been bypassed successfully (Gupta et al., 2023).

Chatbots like ChatGPT can also integrate 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 phishing attempts.

By employing AI-powered techniques, attackers can automate the creation of deceptive spam messages, greatly enhancing the scale and precision of their spear phishing attacks.

3.5 Phishing with audio and video

Phishing that is done using voice is called **vishing** (Salahdine and Kaabouch, 2019). 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 may or may not be in their best interests (Hadnagy, 2018).

With real-time voice morphing, a type of 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.

Back in 2019, attackers successfully used deepfake-generated voice 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 is divided into two parts: technology-oriented countermeasures such as phishing and deepfake detection mechanisms, and human-oriented countermeasures such as training and awareness programs. Tech-oriented countermeasures are examined first since the human-oriented measures rely and build upon them. Chapter 5 then evaluates the effectiveness of these countermeasures in detecting and preventing social engineering attacks.

4.1 AI-based content detection

Traditional phishing detection systems are typically rule-based, which are ill-equipped to adapting to and idenitfying patterns in extensive data streams. AI-based phishing detection foster intricate data processing capabilities, predictive modeling and pattern detection (Fakhouri et al., 2024), ushering an anticipatory and adaptive defensive measure.

Using techniques such as natural language processing (NLP), 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 (Mirsky and Lee, 2021).

AI enhanced mechanisms significantly improve the detection and mitigation of social engineering attacks (Fakhouri et al., 2024)

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 betweeen authentic and malicious content. To counter these sophisticated phishing attacks, researches have developed various AI-enabled detection techniques, in-

cluding Machine Learning (ML), Deep Learning (DL), 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, ML algorithms can learn to distinguish between the two with some methods, such as Random Forest (RF), Support Vector Machines (SVM) and k-Nearest Neighor (k-NN) demonstrating over 95 % accuracy compared to traditional, non-AI based methods. However, care has to be taken when choosing the datasets.

Building on the foundations of machine learning and other AI technologies discussed above, deepfake detection via AI methods is likewise very resource intensive.

Where once experts in the field could recommended 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.

Deepfakes often contain subtle anomalies called artifacts, just as image forgeries of the past did (Mirsky and Lee, 2021). These artifacts can be subtle, such as a strange blob of pixels, or overt such as a person having clearly warped eyes. 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.

Deepfake detection is based on machine learning and forensic analysis, attempting to identify specific artifacts in the multimedia content (Mirsky and Lee, 2021). Seven different types of artifacts are identified in two categories. Spatial-type artifacts cover blending, environment and forensics, while temporal-type artifacts cover behavior, physiology, sychnorization 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. Sychnorization artifacts can be observed in mismatched audio-visual elements, and coherence artifacts relate to inconsistencies in logical sequences happening over time.

4.2 User training and company policy

Human-oriented countermeasures usually fall into four categories: 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 (Tsinganos et al., 2018; Mitnick and Simon, 2003).

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 could help protect users against future threats.

Conducting AI-assisted simulated social engineering and phishing attack campaigns, via numerous channels such as email, SMS and even phone/VoIP, allows organizations to assess the suspectibility of their employees to social engineering tactics. 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.

Feedback from these simulations can be a powerful tool for personnel development, but employees who fall victim to these simulated attacks should never be punished but re-educated. Along the same lines, it is important that employees should be informed beforehand that such campaigns may be intermittently run, which has the double benefit of keeping them on their guard and also not causing unnecessary bad emotions from "being tricked" by their own company (Hadnagy, 2018; Mitnick and Simon, 2003).

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 profesionnal's. They can't do all of the work.

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.

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 [*].

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

5 Evaluation of countermeasures

This chapter evaluates current countermeasures and their effectiveness at detecting and preventing social engineering attacks, particularly those enhanced by AI technologies. The landscape of cybersecurity is continously evolving, and traditional countermeasures such as email filtering and user awareness programs, altought still crucial, are increasingly insffucient 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 AI-powered attacks. After this, Chapter 6 concludes the thesis.

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 and especially in the near future this will not be enough due to the prevalence of deepfake-generated content. User training and awareness programs need to be updated for novel threat of AI in social engineering.

If the attacker manages to get a hold of hashed passwords, AI can be used in the brute force attack, with significantly higher success chance (Blauth et al., 2022).

The uncanny valley is a phenomenal feeling that something is not quite right, deepfake video looks almost real but not quite

Just as spam filters are inclined to report false positives (Fakhouri et al., 2024), so too are deepfake detection systems citepmirskyTheCreationAndDetectionOfDeepfakes2021, filtering legitimate communications causing operational hiccups and perhaps lost engagements.

Technological solutions like phishing detection systems that utilize Natural Language Processing (NLP) and Machine Learning (ML) 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 hihgher 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.

5.2 User-centric

Human-oriented measures remain pivolta in the defense against social engineering. Regular training programs are crucial for equipping end-users with the knowledge to recognize potential threats (Hadnagy, 2018), and 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. Thus comprehensive, regular and innovative user training and awareness programs can never be outlooked. The human is the weakest link in the cybersecurity chain (Mitnick and 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 NLP 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.

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). Certain other demographics, like people on the autism spectrum, a hallmark of which is difficulties in social interaction, may, perhaps contrary to expectations, be more adept 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 these potential differences in demographics. AI can help in designing tailored training content.

5.3 Law and ethics

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

Spreading information about deepfakes to the public faces the hurdle of the "liar's divi-

dend", 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).

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, as discussed earlier in this thesis, 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 on national and global levels. For instance, 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 GAN's and Generative AI(Goodfellow et al., 2020), and thus was not speficially 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 regultaion, 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.

Naturally, restrictions set on AI systems can only be effective if the system stays within the control of its developer. While today, the feasibility of running one's own version of LLM tools such as ChatGPT, due to prohibily high computational costs and other factors, this might not always be the case.

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

*https://openai.com/policies/usage-policies (accessed 2024-08-22)

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 dev in AI and the inherent human factor, being the weakest link (Mitnick and Simon, 2003), within cybersecurity. Therefore, continous 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 (Salahdine and Kaabouch, 2019), is truly imperative for an organization to adapt to and counteract the advancing 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).

*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 modern AI (Fakhouri et al., 2024). This thesis explored how AI empowers malicious actors and also how current countermeasures need to be updated to reflect this evolving threat landscape.

Modern 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 ever more ineffective, requiring a re-evaluation of current strategies and tactics.

One of the most notable contributions of AI is its ability to automate and enhance deceptive practices. Machine learning facilitates the crafting of personalized phishing messages that closely mimic legitimate communications, while deepfake technologies alter or produce synthetic media that convincingly impersonate authentic images, audio and videos. Such advancements enable attackers to deceive targets more efficiently into disclosing sensitive information or taking actions that compromise security.

AI has potential to increase the scale and reach of social engineering (Blauth et al., 2022)

Where previously an employee could authenticate a caller by recognizing their voice, intonations, and accents (Mitnick and Simon, 2003), today and especially in the near future this will not be enough. User training and awareness programs need to be updated for novel threat of AI in social engineering.

In their article, Gupta et al., 2023, claim that "*through continued efforts and cooperation among various stakeholders, it's possible to prevent the misuse of AI systems and ensure their continued benefit to society*", but this can only be true if advanced AI systems remain in the hands of their developers and that they retract older versions of their AI systems from use, since the older versions have already been used by malicious actors. And since with social engineering an attacker can ask ChatGPT to roleplay a certain scenario that the attacker will later enact in a live call, misuse of AI systems can never be fully prevented. AI is a tool, and like any tool it can be used for its intended purpose or in ways the original manufacturer did not intend or would not want.

According to IBM's 2024 Cost of a Data Breach* report, organizations are increasingly leveraging AI and automation in their security operations, with 2/3's of studied organizations deploying these technologies. This presents a 10% increase from last year. Notably, when AI was extensively deployed in prevention workflows, including attack surface management (ASM), red-teaming and posture management, these organizations saw an average reduction of $2.2 million breach costs compared to the average of $4,88 million, a 45% reduction. IBM found a striking correlation, that the more an organization relied on AI, the lower their average breach costs were.

While AI can help detect social engineering attacks, it does not mitigate the need for user training and awareness programs. Quite the contrary, with AI-powered attacks, the need for awareness and vigilance will likely grow even higher. Chatbots like ChatGPT can be used to develop stronger security guidelines and design more engaging social engineering awareness programs.

Dual use property of AI, software created for defense can also be utilized for offensive purposes (Blauth et al., 2022)

What seems certain is that we can count on the rapid development of AI technologies, AI-based social engineering attacks evolving with them, and the need for continuous, innovative user training growing in the future. Attackers and defenders are playing a never-ending game of "cat & mouse" where nobody can rest.

*https://www.ibm.com/reports/data-breach (accessed 2024-08-11)

Bibliography

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., and Kifayat, K. (Jan. 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.

Bilge, L., Strufe, T., Balzarotti, D., and Kirda, E. (Apr. 20, 2009). "All your contacts are belong to us: automated identity theft attacks on social networks". In: *Proceedings of the 18th international conference on World wide web*. WWW '09. New York, NY, USA: Association for Computing Machinery, pp. 551–560. DOI: 10.1145/1526709.1526784.

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. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3191790.

Conteh, N. and Schmick, P. (Feb. 2016). "Cybersecurity:Risks, Vulnerabilities and Countermeasures to Prevent Social Engineering Attacks". In: *International Journal of Advanced Computer Research*, 6, pp. 31–38. DOI: 10.19101/IJACR.2016.623006.

Doan, T.-P., Nguyen-Vu, L., Jung, S., and Hong, K. (June 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)*. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 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*. 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. (Feb. 2024). "AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response". In: *2024 2nd International Conference on Cyber Resilience (ICCR)*. 2024 2nd International Conference on Cyber Resilience (ICCR), 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. (Oct. 22, 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. (Mar. 1, 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.

Mirsky, Y. and Lee, W. (Jan. 2, 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. (Oct. 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. (Dec. 3, 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*. ACSAC '18. 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. (Sept. 2020). "Investigating Teenagers' Ability to Detect Phishing Messages". In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 140–149. DOI: 10.1109/EuroSPW51379.2020.00027.

Salahdine, F. and Kaabouch, N. (Apr. 2019). "Social Engineering Attacks: A Survey". In: *Future Internet*, 11(4), p. 89. DOI: 10.3390/fi11040089.

Tsinganos, N., Sakellariou, G., Fouliras, P., and Mavridis, I. (Aug. 27, 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*. ARES '18. 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. (Jan. 1, 1966). "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Commun. 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**
Sider Fusion	Automatically choosing the best LLM based on the query.
Large Language Models	Finding synonyms for words. Generating LaTeX code for tables and images. Brainstorming ideas about my thesis. Finding related keywords. Highlighting my abstract text and asking how many words it contains. Converting human-written text that I had difficulty reading.
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

I've trialed multiple tools and compared their outputs to find the best ones for my current purposes, which is why the list of LLMs is so extensive.