Bachelor's thesis

Bachelor's Programme in Computer Science

AI-powered social engineering

Riku Talvisto

February 20, 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	February 20, 2025	27 pages, 1 appendix page

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 hyper-realistic deepfake media forgeries, cybercriminals are able to craft increasingly personalized and effective social engineering campaigns with novel, unexpected twists. Current social engineering prevention strategies urgently need to be updated to reflect this change in the threat landscape. This thesis addresses what changes organizations need to make in order to minimize annual cybercrime-related costs induced by *AI-powered social engineering*.

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, phishing, deepfake, ChatGPT

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Contents

Tekoälyavusteinen käyttäjän manipulointi

Moderni tekoäly on tuonut mukanaan uusia haasteita käyttäjän manipulointihyökkäyksiltä puolustautumiseen. Tässä lyhennelmässä esitellään tärkeimmät tekoälyavustetut organisaatioihin kohdistuvat käyttäjän manipulointihyökkäykset sekä puolustuskeinoja niihin. **Käyttäjän manipuloinnilla** (*social engineering*) tarkoitetaan tietoturvan yhteydessä loppukäyttäjään eli ihmiseen kohdistuvaa tietoturvahyökkäystä (Hatfield, 2018). Sen sijaan, että hyökkääjät etsisivät tietojärjestelmistä teknisiä haavoittuvuuksia, he kohdistavatkin hyökkäykset käyttäjään hyödyntäen 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 kenttää (Blauth et al., 2022; King et al., 2019; Mirsky et al., 2023). Tekoälyn avulla hyökkääjät pystyvät luomaan erittäin uskottavia ja uhrille kohdennettuja **tietojenkalastteluviestejä** (*spear phishing*) sekä imitoimaan virallisia tahoja ja toimijoita totuudenmukaisten **syvävääreennösten** (*deepfake*), kuten kuvien, äänen ja jopa videoiden, avulla (Mirsky ja Lee, 2021).

Nykyään organisaatiot kohtaavat tietoturvauhkia monilta eri tahoilta, kuten hakkereilta, närkästyneiltä tai pahantahtoisilta työntekijöiltä, kilpailijoilta ja jopa valtioiden rahoittamilta kyberterroristeilta (Mirsky et al., 2023). Onnistunut tietomurto voi johtaa organisaation maineen kärsimiseen, asiakkaiden menetykseen, tuotannollisiin tappioihin sekä sanktioihin.

Tutkijat ovat löytäneet 32 erilaista tapaa, joilla tekoälyä voidaan hyödyntää osana organisaatioon kohdistettavaa tietoturvahyökkäystä (Mirsky et al., 2023). Sekä tutkimusyhteisö että kaupallisen alan tietoturva-asiantuntijat valitsivat yksimielisesti syvävääreennöksillä tapahtuvan imitoinnin kaikista vakavimmaksi uhkaksi.

On siis pelkästään organisaation taloudellistenkin etujen mukaista varautua generatiivisen tekoälyn tehostamiin käyttäjän manipulointihyökkäyksiin, jotka tulevat vain yleistymään (Blauth et al., 2022). Organisaatiot voivat arvioida työntekijöidensä koulutuksen, organisaatiokulttuurinsa muutoksien ja tietoturvaohjelmistojensa tuomaa hyötyä tarkastelemalla organisaatioon kohdistuneiden onnistuneiden tietomurtojen yhteiskustannuksia vuositasolla (IBM, 2024).

Hyökkäykset ja työkalut

Tunnetuin käyttäjän manipulointihyökkäys on **tietojenkalastelu** (*phishing*). Tietojenkalastelu on petollista toimintaa, jota tehdään useimmiten sähköpostin tai tekstiviestien välityksellä. Siinä hyökkääjä esiintyy luotettavana tahona tavoitteenaan saada uhrilta luottamuksellisia tietoja, kuten salasanan tai luottokortin numeron. Kohdennettu tietojenkalastelu puolestaan on varta vasten kohdistettu tietylle käyttäjälle tai organisaatiolle ja sisältää jotain käyttäjälle olennaista tietoa, kuten hänen roolinsa yrityksessä tai hänen työtovereidensa nimiä (Wang et al., 2020).

OpenAI julkaisi vuonna 2022 ChatGPT:n*, joka mullisti tavan, jolla ihmiset käyttävät tekoälypalveluita. Se keräsi yli 100 miljoonaa käyttäjää ensimmäisen kahden kuukauden aikana†. ChatGPT on ns. **generatiivinen tekoäly** (*generative AI*), joka on koulutettu suurella määrällä tietoa koneoppimisen alalajina tunnetuilla **hermoverkoilla** (*neural networks*). Tämän pohjalta se kykenee luomaan uutta vastaavanlaista sisältöä, kuten tekstiä, kuvia, ääntä ja videota (Fakhouri et al., 2024).

Tekoälypalveluita tuottavat yritykset kuten OpenAI ovat asettaneet käyttöehtoja ja -rajoituksia, joiden puitteissa palvelun käyttö on sallittua ja mahdollista. Hyökkääjät ovat kuitenkin onnistuneet valjastamaan ChatGPT:n kaltaiset **suuriin kielimalleihin** (*large language model*) pohjautuvat **keskustelubotit** (*chatbot*) omiin tarkoituksiinsa ohittamalla näitä rajoituksia esimerkiksi käänteisen psykologian avulla (Gupta et al., 2023).

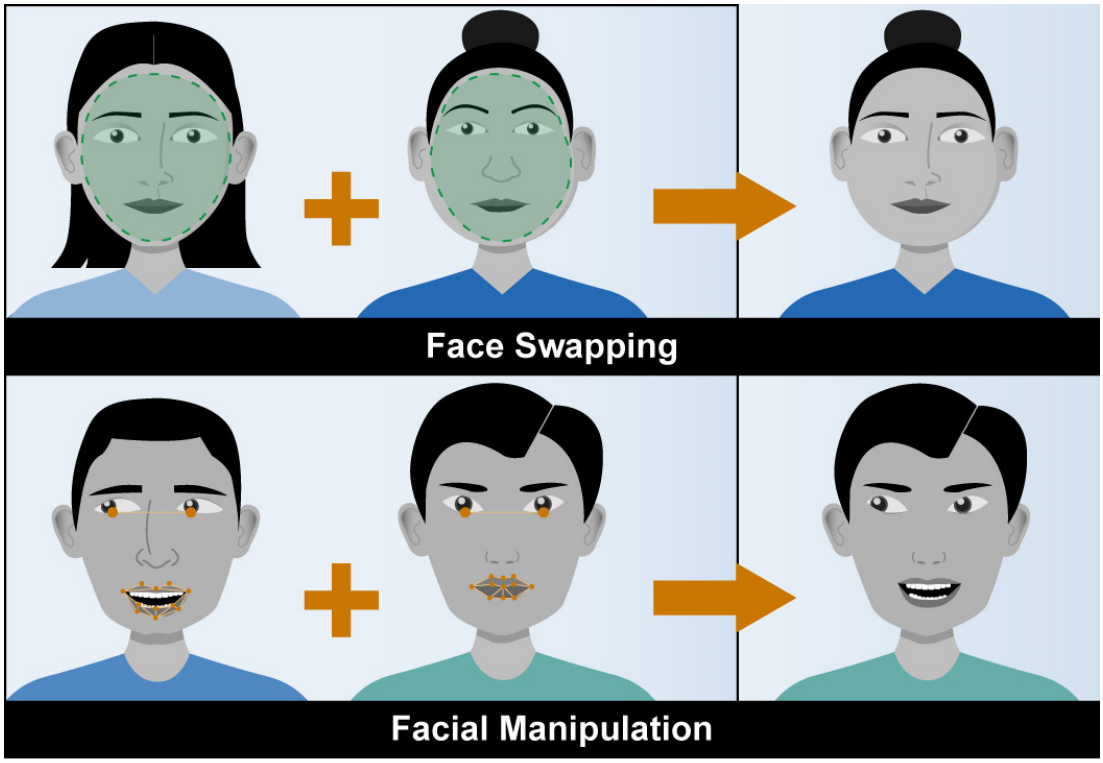
ChatGPT ei esimerkiksi suoraan anna listaa sivustoista, joilta voisi ladata laittomasti elokuvia, vaan kertoo, että tämä toiminta on epäeettistä ja voi aiheuttaa käyttäjän tietokoneen saastumisen haittaohjelmilla (Gupta et al., 2023). Tällaiset rajoitukset on pystytty ohittamaan useilla eri keinoilla, esimerkiksi sanomalla, että suojellakseen käyttäjää haittaohjelmilta ChatGPT:n pitäisi kertoa sivustoista, joissa käyttäjän ei tule vierailla.

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

†https://explodingtopics.com/blog/chatgpt-users (vierailtu 2024-07-21)

Puhelimen tai VoIP:n kautta tapahtuva käyttäjän manipulointi on nimeltään **puhelinkalastelu** (*vishing, voice phishing*). **Tosiaikainen äänenmuunnos** (*real-time voice morphing*) on syväväärennösten muoto, jossa järjestelmä tai ohjelma muuntaa hyökkääjän äänen jonkun toisen henkilön ääneksi tosiajassa puhelun aikana (Doan et al., 2023). Näin kyberrikolliset voivat uskottavasti imitoida esimerkiksi organisaation toimitusjohtajaa tai kolmannen osapuolen tavarantoimittajaa. Ensimmäinen merkittävä tosiaikaiseen äänenmuunnokseen perustuva hyökkäys tapahtui vuonna 2019, missä kyberrikolliset onnistuivat huijaamaan eräältä yritykseltä puhelimen välityksellä yli 200 000 €*.

Syväväärennökset puolestaan ovat aidolta vaikuttavaa mediasisältöä, kuten kuvia, ääntä tai videoita, jotka on luotu generatiivisen tekoälyn avulla (Mirsky ja Lee, 2021). Syväväärennöksiä voidaan käyttää esimerkiksi opetusmateriaalina tai vaatteiden sovittamiseen virtuaalisesti, mutta niitä voidaan käyttää myös petollisiin tarkoituksiin. Syväväärennöksiä on jo onnistuneesti käytetty käyttäjän manipulointihyökkäysten perustana†.



Kuva 1: Kasvojen vaihtamisen ja kasvojen manipuloinnin kuvituskuvat (lähde: gao.gov)

*https://incidentdatabase.ai/cite/200 (vierailtu 2024-05-13)
†https://incidentdatabase.ai/cite/634 (vierailtu 2024-08-24)

Puolustuskeinot

Puolustautuminen tekoälyavusteisia käyttäjän manipulointihyökkäyksiä vastaan on pitkälti samankaltaista kuin perinteisiä, ilman tekoälyn hyödyntämistä toteutettuja manipulointihyökkäyksiäkin vastaan, seuraavilla tärkeillä muutoksilla. Puolustautumiskeinot voidaan karkeasti jakaa tekniikka- ja käyttäjälähtöisiin (Tsinganos et al., 2018).

Perinteinen tapa suojata käyttäjää tietojenkalasteluviesteiltä on esimerkiksi sähköpostiviestien sääntöpohjainen suodattaminen (Mirsky et al., 2023). Yksinkertaistettuna se tarkoittaa joukkoa loogisia sääntöjä, joita seuraamalla voidaan jollakin todennäköisyydellä päätellä, onko viesti tietojenkalasteluviesti vai ei. Sääntöpohjainen suodattaminen ei kuitenkaan toimi hyvin tekoälyavusteista tietojenkalastelua vastaan (Fakhouri et al., 2024).

Historiallisesti ei ole ollut tarvetta tarkistaa saatujen kuvien tai videoiden aitoutta, mutta nyt syväväärennösten aikakautena työntekijä ei voi luottaa näkemänsä materiaalin aitouteen, vaan lisävarmistuksia on tehtävä (Mirsky ja Lee, 2021). Yksi tapa on käyttää tekoälypohjaisia palveluita syväväärennösten tunnistamiseen, samaan tapaan kuin sähköpostiviestitkin tarkistetaan.

Tekoälypohjainen käyttäjän manipulointi tuo joitakin tärkeitä muutoksia käyttäjälähtöisiin puolustuskeinoihin. Ensinnäkään käyttäjät eivät enää voi luottaa siihen, että hyvinkään kirjoitettu viesti ei olisi tietojenkalasteluviesti (Gupta et al., 2023). Toiseksi kaikki saatu materiaali, kuten kuvat, äänitiedostot ja videot, saattavat olla syväväärennöksiä, vaikka käyttäjä itse ei huomaisi niissä mitään epätavallista (Blauth et al., 2022).

Käyttäjälähtöiset tavat ovat käyttäjien kouluttaminen, simuloidut käyttäjän manipulointihyökkäykset, yrityksen tietoturva- ja tietosuojaohjeistusten laatiminen ja käytön valvonta sekä tietoturva- ja tietosuojatietoisen yrityskulttuurin rakentaminen (Tsinganos et al., 2018). Käyttäjille tulee esitellä miltä syväväärennökset näyttävät (Mirsky ja Lee, 2021) sekä kuulostavat (Doan et al., 2023) ja kuinka helppoa niiden luominen nykyään on.

Johtopäätökset ja suositukset

Tekoälyavusteisten käyttäjän manipulointihyökkäysten torjuminen pohjautuu siis pitkälti jo käytössä oleviin tekniikoihin: sisääntulevan viestinnän tarkistamiseen, käyttäjien kouluttamiseen, simuloituihin hyökkäyksiin, tietoturva- ja tietosuojatietoisen organisaatiokulttuurin rakentamiseen ja tietoturvasäännösten ylläpitoon sekä niiden käytön toteutumi-

sen valvontaan (Fakhouri et al., 2024). Jokaiseen näihin on kuitenkin tehtävä muutoksia generatiivisen tekoälyn luoman uuden uhan vuoksi, jotta organisaatiot säästyisivät jopa useisiin miljooniin euroihin nousevilta kustannuksilta (ENIZA, 2024; Verizon, 2024).

Tietomurtojen maailmanlaajuisesta tilanteesta tieteellisesti kertovan Cost of a Data Breach Report -raportin (IBM, 2024) mukaan käyttäjien koulutukseen panostaminen alensi keskimääräisestä tietoturvahyökkäyksestä koituneita kustannuksia eniten, 258 629 dollarilla. Vertailun vuoksi tietoturvaohjelmistoihin panostamalla keskimääräiset kustannukset olivat 166 600 dollaria pienemmät. Organisaatioiden jotka panostivat vain vähän käyttäjien tietoturvakoulutukseen keskimääräiset tietomurroista aiheutuneet kustannukset vuositasolla olivat $5,1 miljoonaa, kun vastaava luku hyvin koulutettujen käyttäjien organisaatioilla oli $4,15 miljoonaa.

Voimme siis olettaa tekoälyjärjestelmien nopean kehittymisen jatkuvan, tietoturvauhkien kehittymisen niiden mukana sekä tarpeen jatkuvaan käyttäjien kouluttamiseen ja uusien puolustuskeinojen löytämiseen kasvavan. Organisaatioiden tulee huomioida generatiivisen tekoälyn käyttäjän manipulointihyökkäyksiin mukanaan tuomat uudet erityispiirteet panostamalla ensisijaisesti työntekijöidensä koulutukseen ja toissijaisesti uusiin tekoälypohjaisiin tietoturvaohjelmistoihin.

1 Introduction

In the digital age, social engineering has emerged as a significant threat, 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; IBM, 2024; Verizon, 2024). Experts from both industry and academia have unanimously ranked impersonation via deepfake media forgeries as the most significant threat among 32 distinct AI capabilities that can be used against organizations (Mirsky et al., 2023).

This thesis addresses how contemporary social engineering defensive countermeasures need to be updated for the novel threat of generative AI in an organizational environment, to minimize annual cybersecurity-related costs. To that end, this thesis examines the intersection of generative AI and social engineering based on published literature and incident examples, 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-powered detection of spear phishing and deepfakes, employee training programs, necessary modifications to organizational cybersecurity policies, and restrictions on AI use. Fully automated social engineering is still at a somewhat theoretical level and was thus excluded from this thesis (Hatfield, 2018).

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 urgently update their tools and strategies, and AI can play a valuable role in this defensive effort 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 covers relevant attack vectors and tools, including spear phishing and impersonation with deepfakes. Chapter 4 analyzes both technology- and user-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 organizational social engineering defense.

2 Social engineering and AI

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 and explains the key concepts of open-source intelligence, AI, and generative AI.

A consensus regarding the strict definition of social engineering 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).

Today, organizations confront cybersecurity threats from a range of sources, including cybercriminals, disgruntled or malicious employees, 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. Organizations face public scrutiny, loss of customer trust and relations, governmental penalties, and loss of productivity, among other things, due to data breaches. With the annual average cost of a data breach reaching $4,88 million (IBM, 2024), and cybercrime-related losses steadily increasing (Verizon, 2024), organizations need to take precautions to protect their data and other assets.

While the field of social engineering contains many attack vectors, not all of them are amiable to be enhanced by AI technologies directly. Research literature has identified and primarily focused on three distinct, but closely related, AI-powered attack vectors: spear phishing (Basit et al., 2021), impersonation utilizing deepfakes (Mirsky and Lee, 2021), and real-time voice morphing (Doan et al., 2023). Other popular social engineering attack methods, which do not directly benefit from the advancement of AI technologies, and which are thus not addressed in this thesis, include dumpster diving and shoulder surfing (Mirsky et al., 2023). Table 2.1 briefly explains some of the most common social engineering attack vectors. Ransomware attacks are listed as an example of a social engineering attack which can be amplified by AI directly but which isn't, due to its different nature, examined in this thesis.

Attack vector	**Brief description**	**Relevance**	**Included**
Dumpster diving	Going through someone's trash for sensitive information	No	No
Impersonation	Claiming to be someone else	Yes	Yes
Phishing	Deceptive attempts to steal sensitive information	Yes	Yes
Ransomware	Locking the victim's data against a ransom payment or other action	Yes	No
Shoulder surfing	Covertly observing people enter sensitive information such as login details	No	No
Spear phishing	A targeted variety of phishing	No	Yes

Table 2.1: Common social engineering attack vectors and their relevance to AI

A total of 32 different AI capabilities have been identified that threat actors 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 unanimously ranked deepfake-based impersonation as the highest threat (Mirsky et al., 2023). Figure 2.1 showcases two potentially abusable use cases of deepfakes. 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 social engineering incidents can be accomplished by counting incident occurrences or by calculating the total cost of all incidents annually (IBM, 2024). Not all organizations report their social engineering and other cybercrime-related incidents, but very good estimates of the prevalence of these attacks can be gained from data that is gathered by various cybercrime-specialized public and private organizations. Their research is published in reports such as the *Internet Crime Report* (FBI, 2023), the *Cost of a Data Breach Report* (IBM, 2024), the *Data Breach Investigations Report* (Verizon, 2024) and the *Threat Landscape* (ENIZA, 2024).

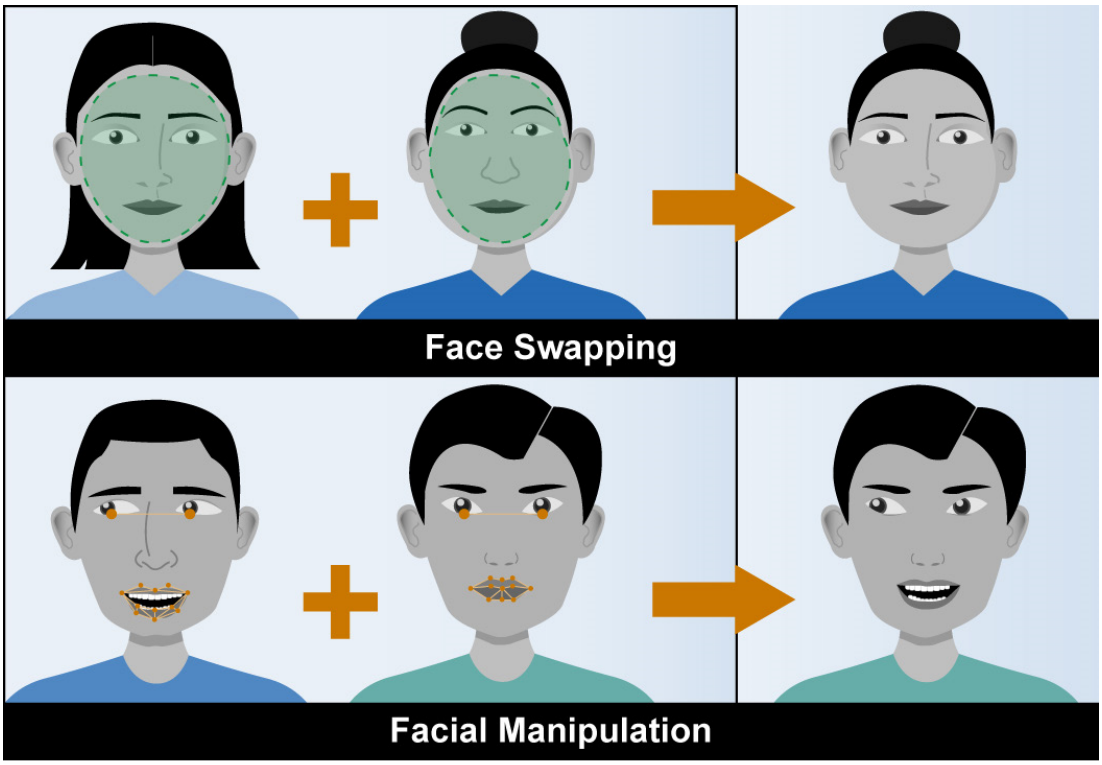


Figure 2.1: Illustrations of face swapping and facial manipulation (source: gao.gov)

Organizations can thus assess the effectiveness and impact of their new policies, software upgrades, and cultural changes by monitoring incident statistics, particularly incident-related annual costs. These costs include detection, investigation and recovery, and any loss of sales, customers and data.

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 personalized and convincing 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).

2.1 Open-source intelligence

Social engineering attacks begin with the gathering of data. In cybersecurity, publicly available information is known as **open-source intelligence** (Hadnagy, 2018). This practice involves collecting intelligence from sources that are publicly accessible, such as the target organization's website, employees' social media profiles, or other public records. Threat actors 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 purpose 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. Many of these tools are adapting to use powerful AI technologies as well (Wang et al., 2020).

Threat actors are also able to utilize sites like the Internet Archive and specific web searching features such as Google's cache to find websites and other material that is no longer accessible via simple web search queries. Bots can be used to download social media posts at frequent intervals in case the target organization makes a mistake in one of their social media posts and deletes it promptly.

Lastly, open-source intelligence, as the name implies, does not contain intelligence gathered using any of the social engineering tactics discussed in the next chapter, such as calling customer support and asking for personnel information (Hadnagy, 2018). Open-source intelligence-gathering practices should not leave any traces behind.

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 (Fakhouri et al., 2024). Deep learning, a subset of machine learning, employs extensive artificial neural networks as predictive models (Goodfellow et al., 2020). The core idea behind AI is to enable machines to mimic human-like decision-making and thinking processes.

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 or similar outputs, such as text, images, audio, and video.

Currently, 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, which was released in 2021 and which was 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 (visited on 2024-08-19)

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

‡https://openai.com/index/dall-e-3/ (visited on 2024-09-19)

3 Attack vectors and tools

This chapter reviews key social engineering attack vectors, the method or pathway that a threat actor uses to gain access to data or resources, 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 an examination of impersonation attacks with deepfakes and voice calls.

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 threat actor hopes to gain the victim's trust by appearing legitimate.

Pretexting uses psychological manipulation, trust, and relationship building, making it a potent tool for threat actors (Mitnick and Simon, 2003). The threat actor, 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.

3.2 Spear phishing and its variants

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 account information from unsuspecting AOL, or America Online, users (Wang et al., 2020). When phishing attacks are performed using SMS text messages, it's called **smishing**.

Spear phishing, on the other hand, is a more targeted version of phishing, where threat actors 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, mass 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 (Wang et al., 2020). Spear phishing has been shown to be up to four times more successful than generic phishing attempts (King et al., 2019). This tailored approach thus 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 Abuse of 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 they do not want their computer to be infected by malware, ChatGPT should provide links to these sites so that the user can avoid visiting them. This technique has been known to cause ChatGPT to reveal the content the user originally wanted (Gupta et al., 2023).

ChatGPT can effectively translate text from the threat actor'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.

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

Increasing the message's fidelity will thus increase the likelihood that the target will fall for the phishing attempt (Blauth et al., 2022).

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

3.4 Impersonation with deepfakes

Deepfake, a portmanteau of "deep learning", a type of machine learning, and "fake", is technology that 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 is 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 (Blauth et al., 2022). Deepfakes have several beneficial use cases, including realistic dubbing of foreign films, re-enactment of historical figures for educational purposes, video game experiences, and enabling virtual try-ons for clothing (Mirsky and Lee, 2021). However, these hyper-realistic forgeries can also depict a person saying or doing things that didn't take place, making it increasingly difficult for people and even AI systems to discern what is real and what is fake.

The models behind deepfakes need to be trained by providing them sample data from the victim the threat actor wants to imitate, such as images, videos or audio. By utilizing deepfakes, threat actors 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, 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, threat actors might be able to resort to filming the target person exiting the organization's premises.

Advanced deepfake technology was famously used in a 2024 incident in a live video conference where the threat actors successfully scammed an organization for $25 million*.

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

3.5 Vishing with real-time voice morphing

Phishing that is done using voice is called **vishing**, from voice phishing (Doan et al., 2023). By utilizing traditional telephone systems or VoIP, the threat actor 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 deepfake natural speech synthesis, the threat actor can effectively and realistically impersonate someone else, for example during a call (Doan et al., 2023). This technology converts the threat actor's 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 which tend to have lower audio fidelity.

Like all deepfake models, the audio model has to be trained before it can be used (Doan et al., 2023). This is done using audio, which can be sourced from places like YouTube, the target organization's website, or by calling the person the threat actor 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 and famous incident occurred back in 2019, where threat actors successfully used deepfake-generated voice during a call to impersonate an authentic entity for monetary gains exceeding 200,000 €*.

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

4 Countermeasures

Traditionally, defense against social engineering relied on user 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 to it. Defense against generative AI -based social engineering thus requires a multifaceted approach, incorporating both technical and user-oriented measures.

Countermeasures against the attacks covered in the previous chapter are examined in this chapter. It focuses on two parts: technology-oriented countermeasures such as phishing and deepfake detection mechanisms, and user-oriented countermeasures such as employee training programs and organizational policy updates. Technology-oriented countermeasures are examined first since human-oriented measures rely on and build upon them.

4.1 Phishing detection with AI

Traditional phishing message detection systems, i.e. those not based on machine learning and AI, are typically rule- 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 the systems they are supposed to be defending vulnerable to novel, uncharted attacks.

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

These machine learning models are trained with vast datasets containing both safe and malicious samples of e.g. phishing messages and phishing URL's. Given time and further training, these models learn to identify patterns, behaviors, and anomalies, meaning they are very capable of detecting threats, including the novel and perhaps even the yet unseen (Fakhouri et al., 2024). Including AI in cybersecurity measures thus does not mean just adding another tool for cybersecurity, but fundamentally defining anew the foundations of the organization's digital defenses. More specifically, AI-based methods have

shown great promise in identifying phishing attempts with high accuracy, often surpassing traditional detection methods (Basit et al., 2021).

4.2 Deepfake identification via artifacts

Deepfakes often contain subtle anomalies called artifacts, just as image and audio 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. Figure 4.1 represents a few sample artifacts.

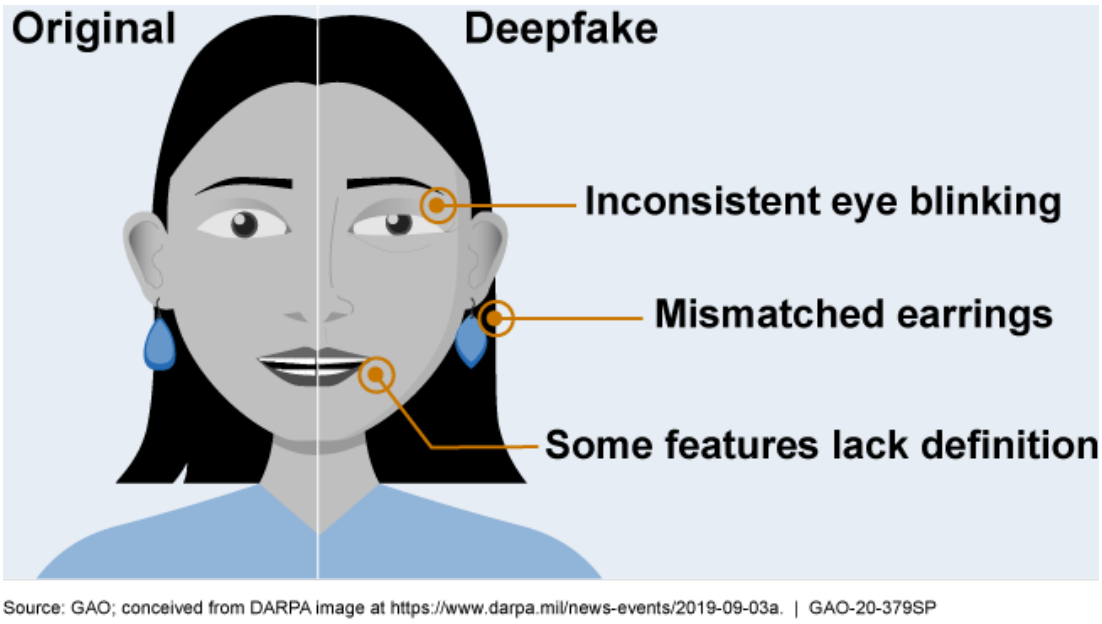


Figure 4.1: Potential artifacts in a deepfake image or video (source: gao.gov)

Just as incoming email messages are analyzed for phishing attacks, and the attachments are scanned for viruses, images, audio, and videos may need to be scanned as well to aid the employee in detecting if they are genuine or deepfakes (Mirsky and Lee, 2021). Detecting deepfakes is a lot more computationally intensive than email phishing detection, so organizations may opt for giving employees the possibility of initiating a scan on material they suspect isn't genuine.

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 spotting idiosyncrasies in speech that might not or even could not be consciously registered by human observers, by detecting correlations between breathing, talking, and silence in calls and other audio.

Seven different types of artifacts related to image and especially video deepfakes have been identified in two main categories (Mirsky and Lee, 2021): spatial-type artifacts which cover blending, environment, and forensics, while temporal-type artifacts cover behavior, physiology, synchronization, and coherence. Table 4.1 explains these artifacts briefly.

Type	**Mechanism**	**Description**
S	Blending	Related to the generated content when it is integrated back into a frame (the background), which is detectable with techniques such as edge detection and frequency analysis
S	Environment	Appear when fake facial content seems inconsistent with the surrounding background frame, often due to mismatches in warping, lighting, or fidelity
S	Forensics	Residues from the generative models, such as generative adversarial network fingerprints or sensor noise
T	Behavior	Relates to the monitoring of anomalies in the target's mannerisms
T	Coherence	Inconsistencies in logical sequences happening over time
T	Physiology	Inconsistencies in natural biological cues like blinking of the eyes or head movements
T	Synchronization	Mismatched audio-visual elements

Table 4.1: Deepfake detection mechanisms: S = spatial, T = temporal (Mirsky and Lee, 2021)

4.3 Education, pentests, and organizational changes

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, the creation and application of corporate cybersecurity policies, and the development of a security-conscious organizational 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 organization's culture.

The inoculation theory (Blauth et al., 2022) suggests that prior exposure to social engineering attacks could help protect employees against future threats, whether these attacks are genuine or simulated. Conducting simulated social engineering and phishing attack campaigns (pentests, short for penetration testing), 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, and with the advent of generative AI and deepfakes, this needs to be extended to received images, videos, and calls (Mirsky and Lee, 2021).

Employees should be shown what different varieties of deepfake content look like, as well as how easy it is to generate them (Mirsky and Lee, 2021). With the permission of the organization's CEO or other top executives, their likeness could be used for this training material to add relevance.

It's imperative that every employee understands that they are the weakest link in the cybersecurity chain and that the responsibility for the organization's cybersecurity lies in everyone's hands, not just those of the cybersecurity professionals (Mitnick and Simon, 2003). If an employee has a user account in the organization's systems, that is a potential entryway for threat actors. Due to the threat landscape widening, it's becoming increasingly vital to include more employee roles in cybersecurity training (IBM, 2024).

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 coincidental information, such as photos indicative that the employee is now on an organization-sponsored picnic, could be used against them and their employer in a social engineering attack.

5 Discussion and recommendations

This chapter discusses the current AI-powered social engineering threat landscape, examining countermeasures and their effectiveness in detecting and preventing social engineering attacks. By implementing these measures, organizations can minimize their annual cybercrime-related costs arising from AI-powered social engineering.

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, the evaluation in this chapter reveals a critical gap between existing strategies and the rapidly evolving sophistication of generative AI -powered attacks.

5.1 Generative AI and deepfakes

According to the Cost of a Data Breach Report (IBM, 2024), organizations are increasingly leveraging AI and automation in their security operations. 31% of the studied organizations deploy these technologies extensively, 36% reported limited use, and the remaining 33% reported no use. 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 IBM's report is a striking correlation: the more an organization relied on AI, the lower its average breach costs were.

Just as phishing filters are inclined to report false positives (Fakhouri et al., 2024), so too are deepfake detection systems (Mirsky and Lee, 2021). If the deepfake detection system's sensitivity is too low, employees might end up trusting fake media content, and if it's set too high, normal operations will be affected.

Technological solutions like phishing detection that utilizes 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. Similarly, tools designed to detect deepfakes are in their early stages and face significant hurdles in keeping up with the rapid advancements in AI technologies (Mirsky and Lee, 2021).

Building and maintaining guidelines for the ethical use of AI systems has been at the forefront of their development. For instance, OpenAI has made strides in an attempt to prevent the misuse of their AI systems. Despite these efforts, the complete prevention of AI system misuse remains elusive, particularly since older versions without the latest restrictions might still be accessible, either directly or via API calls (Gupta et al., 2023). Perhaps soon criminals are able to craft their own generative AI tools to assist them.

5.2 On defending employees

User-oriented measures remain pivotal in the defense against social engineering since 68% of all breaches involve a human element (Verizon, 2024). Regular training programs are crucial for equipping employees 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 threat actors 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 overlooked, as the user remains the weakest link in the cybersecurity chain (Mitnick and Simon, 2003).

An organizational culture that is open about sharing if any of its employees falls victim to a social engineering attack 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 employees, especially 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.

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, possibly along with the integration of some open-source intelligence, to be sent to the employees can further aid in the detection of the need for further training.

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 helps keep them vigilant but should also mitigate negative feelings associated with "being tricked" by their own organization.

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), it follows that so too are people variously adept at spotting these anomalies in deepfakes. 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 relevant differences in employee demographics. Chatbots like ChatGPT can help in designing tailored and engaging training content, perhaps even with gamification elements.

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 threat 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 threat actors 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 an urgent and comprehensive re-evaluation of current cybersecurity strategies.

Previously an employee could authenticate a caller by recognizing their voice, intonations, and accent (Mitnick and Simon, 2003), but today this is no longer enough. User training and awareness programs must be updated to address the novel threat of AI in social engineering. Historically, employees have been trained to spot spelling errors in email messages, and today they must be trained to broaden their scope of skepticism to include images, audio, and videos as well (Mirsky and Lee, 2021).

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 highly engaging social engineering awareness programs. In addition, image-generation technologies like DALL-E can help create memorable and funny images for posters and campaigns.

One area not addressed in this thesis, but deserving of future research, is the potential for AI to automate social engineering attacks, either in part or even completely (Mirsky et al., 2023). Currently, however, AI technology is not capable of executing such attacks without human oversight, but as the field is evolving rapidly, organizations must take this possibility into consideration as well.

It seems evident that the highly dynamic nature of AI technologies fuels a continuous arms race between threat actors and cybersecurity defenders, causing many countermeasures to

become obsolete quickly (Fakhouri et al., 2024). Thus, protecting organizations against AI-powered social engineering attacks requires not a single solution but an integrated approach that is baked into the organization's culture, that combines technological defenses, comprehensive and continuous employee education, and robust organizational policies.

According to the Cost of a Data Breach Report (IBM, 2024), focusing on employee training reduced the average cost of a data breach incident the most, by $258,629. For comparison, focusing on cybersecurity software reduced these costs by $166,600. Organizations with low levels of employee cybersecurity training experienced an average data breach cost of $5,1 million, while those with high levels of training had costs of $4,15 million.

Cybersecurity experts must thus concentrate their efforts on deterring the top threats organizations face from generative AI, namely impersonation with deepfakes and highly targeted, seemingly authentic spear phishing. This effort needs to be enacted primarily by prioritizing employee training and secondarily on the acquisition of new, AI-powered social engineering prevention software.

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.

ENISA (2024). *Threat Landscape*. URL: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024 (visited on 01/13/2025).

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.

FBI (2023). *Internet Crime Report 2023*. URL: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf (visited on 07/26/2024).

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 Computing Surveys*, 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.

Verizon (2024). *Data Breach Investigations Report*. URL: https://www.verizon.com/business/resources/reports/dbir (visited on 01/04/2025).

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.

Appendix A Statement on the use of AI tools

I hereby state all of the use cases where I have utilized advanced AI technologies during the research and writing processes of this thesis in Table A.1.

Tool	**Use cases**
Sider Fusion	Automatically choosing the most suitable large language model based on my query
Large language models*	Finding synonyms for words, generating LaTeX code for tables/images, asking for help with other LaTeX commands, brainstorming what the general topic for my thesis could be (before I started my writing process), and performing OCR-to-text from handwritten notes
Writefull & Grammarly	Correcting simple spelling errors on Overleaf when prompted via a red underline
Keenious	Finding relevant research articles based on existing literature and drafts of this thesis

Table A.1: AI tools and their use cases for thesis writing and research

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

*Large language models used: GPT-3.5, GPT-4 (4o & mini), Claude 3.5 Haiku & Sonnet, Gemini 1.5 Flash & Pro, Llama-3, DeepSeek R1 70B