Bachelor’s thesis

Bachelor’s Programme in Computer Science

[DRAFT] AI-powered Social Engineering

Riku Talvisto

August 21, 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 Faculty of Science	Koulutusohjelma — Utbildningsprogram — Study programme 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 Bachelor's thesis	Aika — Datum — Month and year August 21, 2024	Sivumäärä — Sidoantal — Number of pages 17 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 AI technologies, cybercriminals are able to craft highly targeted and effective social engineering campaigns, with novel unexpected twists, that are sometimes entirely automated.

This thesis explores the evolving landscape of AI on social engineering, focusing on attacks such as spear phishing, automated intelligence gathering and impersonation with deepfake generated content. It also covers countermeasures against these attacks, such as those aided by AI, and evaluates their effectiveness based on relevant literature and case studies. The results indicate that contemporary countermeasures against social engineering attacks are becoming increasingly ineffective, highlighting the urgent need for cybersecurity professionals to update their strategies for cyber defense.

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 contains a Finnish language summary and information on the use of AI technologies.

Contents

Tekoälyavusteinen käyttäjän manipulointi

Käyttäjän manipuloinnilla (*social engineering*) tarkoitetaan tietoturvan kontekstissa tietojärjestelmän loppukäyttäjään, eli ihmiseen, kohdistuvaa tietoturvahyökkäystä. Sen sijaan että hyökkääjät etsisivät teknisiä haavoittuvuuksia, he kohdistavat hyökkäykset ihmiseen käyttäen hyväksi psykologisia menetelmiä.

OpenAI julkaisi vuonna 2022 ChatGPT:n joka mullisti tavan jolla ihmiset käyttävät tekoälyäpalveluita. Se keräsi yli 100 miljoonaa käyttäjään ensimmäisen 2 kuukauden aikana. ChatGPT on ns. generatiivinen tekoäly joka on koulutettu suurella määrällä dataa ja joka pystyy luomaan transformaatio-prosessin (*transformer*) uutta sisältöä, kuten tekstiä tai kuvia.

Hyökkäykset ja työkalut

Tässä luvussa käydään läpi hyökkäyksiä.

Hyökkääjät ovat onnistuneet valjastamaan ChatGPT:n kaltaiset suuret kielimallit (*large language model*) ohittamalla niiden kehittäjien niille asettamia rajoituksia.

Puolustuskeinot

Tässä luvussa käydään läpi puolustuskeinoja.

Puolustuskeinojen arviointia

Tässä luvussa käydään arvointia puolustuskeinojen tehokkuudesta.

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 adaptation of information technology (IT) devices and services has transformed every aspect of the modern man's life, from personal communication to business operations, and this reliance seems to be ever expanding. Although this digital revolution has opened up many opportunities, it has also brought about considerable vulnerabilities. One of the most sinister threats to security and privacy is social engineering, with FBI's Internet Crime Complaint Center (IC3) reporting year 2023's losses to organizations and individuals from business email compromise scams alone reaching a record $2.9 billion[†].

Social engineering (SE) 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, SE relies on human interaction and exploits weaknesses in human psychology (Wang et al., 2020).

With the advent of modern artificial intelligence (AI), the landscape of social engineering is undergoing significant transformation, augmenting the sophistication and effectiveness of such SE attacks. Based on a literature review, this thesis aims to explore both classical (pre-AI) and modern social engineering, with special focus on how contemporary AI technologies enhance the execution and impact of SE attacks, and discusses the necessary actions to counter such advanced attacks.

Historically, SE relied heavily on human intuition and manual effort to deceive targets. Today, AI systems are more and more capable of automating and amplifying these deceptive practices, often with shocking alarming precision. From sophisticated spear phishing schemes utilizing technologies such as ChatGPT to deepfake-technologies creating highly convincing fake identities via audio, pictures and even video, AI-powered attacks represent a truly profound shift within social engineering, which we have only begun to witness.

Certain SE attacks and tactics that are of particular interest when it comes to the modern threat of AI to SE were chosen for more in-depth analysis, such as spear phishing and deepfake-generated content, while leaving other attacks with less focus, such as dumpster diving, shoulder surfing, and tailgating.

This thesis is organized as follows. The first chapters build the ground work for fur-

[†] https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf (accessed 2024-07-13)

ther analysis, beginning with Chapter 2 which provides an overview of social engineering, including its definition and some background knowledge. Chapter 3 examines common attack methods and vectors, such as spear phishing and deepfake impersonations, and explores how modern AI capabilities amplify their effectiveness.

Finally, with the groundwork being built, Chapter 4 discusses how existing countermeasures need to be augmented for the novel threat of AI. Finally, Chapter 6 concludes the thesis, summarizing the key findings and implications for the future of social engineering defense with the modern threat of AI.

2 Social Engineering and AI

This chapter gives an overview of what social engineering (SE) constitutes, provides brief historical context and describes some key terminology that is necessary for further analysis, including about AI. After this, Chapter 3 a typical SE attack.

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).

There isn't a strict consensus on the field about what actually constitutes an SE attack (Hatfield, 2018). For the purposes of this thesis, SE 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, such as open-source intelligence and pretexting, are explained next.

2.1 Open-Source Intelligence

OSINT, sometimes written as OS-INT, means 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.

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 OSINT, or the gathered open-source intelligence, in assisting with the creation of the story. Pretexting is discussed here and not in the chapter which discusses attack vectors because it is usually the precursor to other attacks and is heavily tied with the concept of OSINT (Hadnagy, 2018).

Unlike other SE attacks that rely on technical methods or overt threats or fabricated time pressures, pretexting uses psychological manipulation, trust and relationship-building, making it a potent tool for attackers. 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 (Mitnick and Simon, 2003).

2.3 Artificial Intelligence

Artificial intelligence (AI) refers to the various levels of simulation of human, and sometimes animal, intelligence in machines that are programmed to think and learn like humans or other animals. AI can be categorized in two ways, narrow AI which is designed for specific tasks such as playing StarCraft, and general AI which aims to perform any intellectual task that a human could do.

Generative AI (GenAI) refers to a subset of AI that focusses on creating content, whether that be text, images, music, or more. Unlike with traditional AI, which generally follows programmed rules to make decisions, GenGenAI learns patterns from data and can generate new outputs based on that training. The most prominent example of GenAI is ChatGPT, a type of narrow AI, released by OpenAI to the public in 2022*.

ChatGPT, built on the GPT (Generative Pre-trained Transformer) architecture, is de-

*https://openai.com/index/chatgpt (accessed 2024-08-19) signed to understand and generate human-like text by predicting the next word in a sequence and when to stop generating that sequence (stop symbol). Its widespread availability has demonstrated the potential of AI tools to assist in numerous tasks, from customer service to creating content and aiding in complex information retrieval and processing.

3 Attack Vectors and Tools

This chapter provides an overview of some of the most common SE attack methods, paying attention to how modern AI technologies are or could be augmenting them. After that, Chapter 4 goes over the countermeasures against these attacks.

3.1 ChatGPT and LLM's

Generative AI's (GenAI) can be used by malicious actors in their schemes, but due to the manufacturer's set limits, some workarounds need to be used (Gupta et al., 2023). Asking ChatGPT to provide links to websites which provide pirated content such as movies results in ChatGPT denying the request, stating that downloading pirated content is unethical and may also lead the user's computer to be infected with malware.

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 (Gupta et al., 2023). 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 Phishing & spear phishing

As the quintessential SE 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).

Since email users are used to seeing URL's of different types, some where the "login" text is used, some where it is omitted, and some where it's used as a subdomain (login.ibm.com) and some where it's used as a subfolder (paypal.com/us/signin), attackers hope they are able to deceive their targets with their fabricated URL's which are designed to look like their authentic counterparts, sometimes replacing an "i" with an "l" or using the domain name of an URL as a subdomain (paypal.com.login.example.com).

These links are then placed in email messages which redirect the user to a website that looks authentic and tries to gather sensitive data from the user, such as usernames, passwords or credit card details.

Spear phishing is a more targeted version of phishing, where attackers customize their deceptive emails to a target individual or organization (Basit et al., 2021). Unlike with generic phishing attempts, this type of phishing involves gathering detailed information about the victim, via OSINT 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 is a lot more time and energy consuming.

By employing AI-powered techniques, attackers can automate the creation of deceptive spam messages, greatly enhancing the scale and precision of phishing attacks. An advanced level of personalization is reached automatically through data mining and analysis where AI processes through vast amounts of publicly available information on social media platforms such as Facebook, X (Twitter) and Instagram, on forums and other digital resources to extract insights about potential victims. These insights, such as by expressing in an email message the hope that the victim enjoyed the private company picnic last month and the caffeine-free sodas that were on offer, is used in the generation of the spear phishing messages to increase their seeming authenticity.

3.3 Deepfake-generated media

Deepfake, a portmanteau of "deep learning", a type of machine learning, and "fake", is technology which uses AI to create highly convincing fake media, either by altering existing content or creating them from scratch (Mirsky and Lee, 2021). Deepfake content can be images, audio, and even full-resolution video.

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. For example, a deepfake video of the victim's company's CEO making an urgent request for sensitive information can exploit the employee's natural tendencies to comply with authority, thus bypassing any skepticism that could've risen from a simple email message.

These deepfakes are then delivered to the victim via a number of different channels, such as email, instant messaging, SMS messages or phone/VoIP (Voice over IP) along with any other relevant information pertaining to the attacker's attempt at influence and manipulation.

4 Countermeasures

In this chapter, countermeasures against both classic social engineering and AI-powered social engineering are examined. This chapter is divided into two parts: tech-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. The division is not always clear cut and is made only to simplify the reading experience.

Chapter 5 then evaluates the effectiveness of these countermeasures in detecting and preventing social engineering attacks.

4.1 Technology oriented

Technology-oriented countermeasures can be implemented on a per end-user device or on a network level, and usually combination of the two is always required (e.g. with firewalls and antivirus software).

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).

Since there are a myriad of ways to get multimedia content to the target victim, such as via USB thumbdrives dropped at the company's premises, technological methods alone will never be enough to counter SE attacks and user training and awareness programs need to be continuously updated (Hadnagy, 2018). These methods are discussed next.

Generative AI (defense)

ChatGPT can also be a force for good.

4.2 Human oriented

Regular and comprehensive training programs are vital to educate employees about SE 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 SE 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.

Conducting AI-assisted simulated SE and phishing attack campaigns, via numerous channels such as email, SMS and even phone/VoIP, allows organizations to assess the susceptibility of their employees to SE 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 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 SE 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 SE attacks rather than trying to hide them for fear of ridicule or punishment (Mitnick and Simon, 2003).

Finally, because AI can source social media sites and the Internet automatically for OSINT, it's imperative for people to know to be careful of what they share, with whom and when. 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.

5 Evaluation of Countermeasures

This chapter evaluates current countermeasures and their effectiveness at detecting and preventing social engineering attacks. Next, Chapter 6 concludes the thesis.

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.

6 Conclusions

The subfield of social engineering (SE) within cybersecurity is undergoing a significant transformation with the advent of modern artificial intelligence (AI). 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 automated spear phishing and voice phishing (vishing) with real-time voice morphing. 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.

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 SE.

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.

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.

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.

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.

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 Computing Surveys*, 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.

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

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.

Utilization of AI Technologies

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

During the research and writing of this thesis the use of AI when working on my thesis. How I used AI tools and which tools were used.

Most of my AI use has been done via Sider.ai, which is a browser extension sidebar app that allows direct interaction with the current webpage, or any specific parts of it, with various LLM tools. Sider.ai provides Sider Fusion, which dynamically selects an appropriate model to be used based on the given query.

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 model from the list of used LLM's based on the query.
GPT-3.5, GPT-4, GPT-4o, GPT-4 mini, Claude 3.5 Sonnet, Gemini 1.5 Flash, Llama-3	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. Convert 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.

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