Bachelor's thesis

Bachelor's Programme in Computer Science

AI-Powered Social Engineering: Attacks & Countermeasures

Riku Talvisto

June 17, 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

AI-Powered Social Engineering: Attacks & Countermeasures

Ohjaajat — Handledare — Supervisors

Dr. Lea Kutvonen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Bachelor's thesis	June 17, 2024	12 pages

Tiivistelmä — Referat — Abstract

This thesis explores the intersection of artificial intelligence (AI) and social engineering, examining the emerging threats as well as opportunities for defense in this space.

The findings reveal that AI-powered attacks are more convincing and successful than traditional attacks, but AI-driven defense mechanisms can improve detection rates, and AI can develop more engaging training material(?).

The study contributtes to our understanding of AI's impact on social engineering and highlights the need for cybersecurity professionals to adapt to these emerging threats.

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 grown even higher.

Technologies such as deepfake (deep learning, fake) videos and live voice morphing will change the landscape of SE attacks.

As the field is still emerging and the field of AI development has seen increased growth vastly, much of this the content of this thesis is speculative, based on attacks that have already been acted out succesfully and the developments on the AI field.

What, if anything, can the end-user trust anymore?

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, psychology

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction

The widespread adoption of information technology (IT) technologies and services has transformed every aspect of human life, from personal communication to business operations, and this reliance on devices and technologies is ever expanding. Although this digital revolution has opened up many opportunities, it has also brought about considerable vulnerabilities. One of the most dangerous threats to security and privacy is social engineering. Social engineering (SE) is the art of manipulating people into performing actions or divulging confidential information. Rather than looking for technical vulnerabilities, SE relies on human interaction and exploits weaknesses in human psychology (Wang et al., 2020).

This thesis is organized into five chapters, each focusing on a specific aspect of social engineering and AI-augmented attacks. Chapter 1 provides an overview of social engineering, including its definition and historical context. Chapter 2 delves into the various stages of a social engineering attack, highlighting the tactics and strategies employed by attackers. Chapter 3 examines common attack methods, such as pretexting, phishing, and tailgating, and explores how modern artificial intelligence (AI) capabilities amplify their effectiveness. Chapter 4 discusses AI-powered countermeasures and their potential to detect and prevent social engineering attacks. Finally, Chapter 5 concludes the thesis, summarizing the key findings and implications for the future of social engineering defense.

2 History and Definition of Social Engineering

The term *social engineering* is perhaps overused and is certainly misused, and for clarity, in this chapter a clearer definition is formed and some history reviewed. For the purpose of this thesis, we'll use the definition for SE given by Wang et al., 2020: "social engineeing is 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 techincal means and technical vulnerabilities."

The term *social engineering* seems to have first appeared in an article titled "More on Trashing", which was published in September, 1984 on one of the earliest hacker magazines, The Hacker's Quarterly (Wang et al., 2020), but the broader concepts of human exploitation date back much farther than this (Qin and Burgoon, 2007). For as long as humans have partaken in communication and trade, there have been those that have tried to exploit the system in their own favor, in an unethical and selfish way.

For as long as there has been social contracts, both written and oral or implied, there have been those that try to get away with cheating. Social engineering, thus, is not a new concept. In 19xx people who partook in actions akin to SE were called confidence men or con artists. They manipulated their victim to be complacent in the act.

As described by **abiteboul**, the term *social engineering* is perhaps overused and is certainly misused. What exactly constitutes social engineering? In this thesis, SE is defined as "social engineering is the deliberate act of convincing a victim, usually though the use of technology, to perform an action that may or may not be in their best interest". A definition of social engineering by Hadnagy has become quite popular "may or may not be in their best interest". HADNAGY

Some scholars have included acts like shoulder shurfing and dumpster diving as social engineering attacks (**abiteboul**), even though they do not rely on the manipulation of individuals. However, since they don't fall neatly into the typical category of "hacking attacks", they fall into a gray area in between. Since user education and awareness programs should include training the user against discarding important documents without shredding them first or by putting them into boxes designated "secure documents", and

people need to be on the lookout for people gazing over their shoulders, especially when entering sensitive information such as usernames, passwords and other access codes, they are included in this thesis.

It's also necessary to go over some basic terminology related to the field of SE.

Two common attack methods, namely **dumpster diving** and **shoulder surfing**, are often categorized as social engineering attacks, but do not necessarily fall under the SE category, as they do not include direct social interaction with the victim (Wang et al., 2020). In other words, the compliance of the victim is not necessary in these two types of attacks. However, since they are non-technical in that no devices or technologies need to be used, they fall in a "gray area". They require neither the use of devices or technologies nor the psychological manipulation of the user. Thus, it has been easy to refer to them as SE attacks. In this paper, they are still discussed because SE training and awareness programs should include mentioning about these.

2.1 Open-Source Intelligence, OSINT

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.

As people have adopted to using social media as sometimes their primary means of communication, a lot of exploitable data is shared on these platforms as well.

3 Stages of Social Engineering Attacks

Social engineering attacks are usually deployed in multiple stages. Various scholars define these stages in different ways (Mouton et al., 2016), but they generally follow a 4-stage process.

The 4-stage process is:

1. Collect information about the target
2. Develop a relationship with the target
3. Exploit the gathered information and the built relationship by executing the attack
4. Exit having cleaned any traces

An attack against an individual or a corporation can happen in multiple iterations, with subsequent iterations building on from the information and resources (such as passwords, access badges) gathered from the previous.

Next, each of these four stages is examined in detail. Special emphasis is placed on how the emergence of modern AI technologies could impact, or has already impacted, these stages. A subsequent chapter will go over the different attack types, followed by a chapter that examines countermeasures against them.

3.1 Intelligence Gathering & OSINT

First stage is research or information collection. This relies on what's known as OSINT, or open-source intelligence. OSINT is any information that is publicly available via the Internet or other means and that doesn't require any breach of security to be accessed.

This intelligence is then used to formulate an attack plan against the target individual or corporation.

3.2 Developing the Relationship

Once the attacker has gained information and formulated an attack plan based on it, he

3.3 Exploiting the Victim

3.4 The Exit

Hadnagy et al. emphasize that in whatever you do, leave your target better off for having met with. This is in stark contrast to typical SE attacks.

Any bridges should not be burned but the relationship should be exploitable in the future as well

The SE is an iterative process, use what was found from previous attacks with the next one

4 Attack Methods

Awareness of various social engineering (SE) attacks is crucial for professionals across all industries, not just those in cybersecurity. This chapter provides an overview of the most common SE attack methods.

Some attacks that are discussed here are not always considered a type of SE attack, specifically shoulder surfing and dumpster diving, as these do not require the co-operation of the victim (Wang et al., 2020). However, since they are often used in conjunction with other SE attacks, and since training for against them is often included in SE training and awareness programs, they are explained here

To better understand the threat posed by SE, it is essential to examine the diverse strategies employed by attackers. The following are some of the most common and effective SE attack methods. We'll also analyze how the emergence of modern AI technologies might, or already has, powered up these types of attacks.

Countermeasures against these attacks are examined in a later chapter.

4.1 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. Phishing has been around since 1996, when cybercriminals began using deceptive emails and websites to steal AOL (America Online) account information from gullible users (Wang et al., 2020).

Spear phishing is a more targeted version of phishing, where attackers customize their deceptive emails to a target individual or organization. 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 scam.

Last on the list of phishing attacks is **whaling**. Whaling, also known as CEO fraud, is a highly targeted phishing attack aimed at high-profile individuals within an organization,

such as executives or senior management, "the big whales" (Abraham and Chengalur-Smith, 2010). The attackers carefully research their targets to create convincing and typically urgent messages that appear to come from trusted sources, often impersonating colleagues, business partners, or government agencies. The goal is often to authorize large financial transactions or to leverage the target's authority and access within the company.

Two additional types of phishing need to be addressed, and they are **vishing** or voice phishing and **smishing** or SMS phishing. Despite having complicated names, the idea behind them are quite simple.

Regular phishing doesn't usually require OSINT but spear phishing does.

4.2 Baiting

Baiting is a similar attack method to phishing, discussed above, but emphasizes luring the victim via enticement strategies (Conteh and Schmick, 2016; Salahdine and Kaabouch, 2019). This technique exploits the target's curiosity or greed to gain unauthorized access to resources or premises or to obtain sensitive information.

4.3 Pretexting

Pretexting involves fabricating a story or a scenario, a **pretext**, that is plausible but fraudulent, to engage the target and extract information with (Conteh and Schmick, 2016). This type of attack relies heavily on OSINT, or the gathered open-source intelligence, in assisting with the creation of the story.

Modern AI can assist with the deployment of the pretext by being a sparring partner to the attacker, giving a safe "sandbox" to try out potential attacks and find ways in which the target might react.

4.4 Tailgating

Tailgating, also known as **piggybacking**, is a social engineering tactic that involves following an authorized person through an access-controlled passage, such as a security gate. This type of attack exploits individuals with temporary access rights, such as de-

livery personnel or maintenance workers (Conteh and Schmick, 2016). The attacker may use manipulation to gain access, for instance, by carrying a heavy object and asking for assistance or pretending to be a delivery person with a forged pretext.

In another scenario, the attasacker may use psychological manipulation to evoke sympathy, claiming to have forgotten their security ID and worrying about losing their job. This tactic preys on people's natural instinct to help and be polite. The mundane routine of passing through security gates can also lead to a false sense of security, making individuals less vigilant.

4.5 Shoulder Surfing

Observing people enter sensitive information, such as login details or financial data, without their knowledge or approval is called **shoulder surfing**. The name implies a person watching over the shoulder of someone when they are typing or viewing sensitive data, but the attack surface is actually far larger. The attacker could use cameras, either hacked, or those placed there by the attacker for this purpose, or even everyday objects such as binoculars.

The proliferation of high-resolution cameras, such as those with HD or 4K resolution, has exacerbated the issue of unauthorized information gathering since security cameras of the past didn't have the resolution necessary to show what's being viewed on a screen. Going through tens or even hundreds of hours of video material where sensitive information might be visible on an individual's or an employee's screen is time-consuming and tedious, but with the help of AI this job can be carried out with more ease. AI can not only turn words in a video into text, but it can also summarize the found text and search for anything that could be exploited.

4.6 Dumpster Diving

As the name implies, **dumpster diving** refers to the practice of going through an organization's or an individual's trash in order find sensitive information that should have been disposed of properly (Syafitri et al., 2022). Rummaged content may yield interesting results, typically in the form of documents such as papers but also media devices such as optical discs, hard drives or USB thumbdrives, to be used as-is or as part of a future SE

attack.

Since dumpster diving does not include manipulation of people in its purest form, rather relying on the improper care of documents and other material, it is sometimes not classified as a social engineering attack (Wang et al., 2020). However, it is often considered a precursor or supplementary tactic in broader social engineering schemes, as the information gathered can be used to craft more convincing and targeted attacks.

5 Countermeasures

5.1 Company Policy

Regarding bating attacks, company policy should enforce antivirus scanning of attached media, such as USB thumbdrives, before they can be accessed. Personnel should be instructed to never pick up and plug in found media (Salahdine and Kaabouch, 2019). However, if media is found that is suspected of being used as part of a baiting attack, the employee could notify the front desk or the IT services that such a media device was found within the premises, rather than just ignoring it

6 Conclusions

As we've seen, SE is still as much a threat as it has ever been, despite major efforts to the contrary.

What's certain is that we can count on AI developing, AI-based social engineering attacks evolving with it, 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.

X in Y references that training users effects will wear off in 3 weeks, necessiting continous retraining approaches.

I'll end with the question that I started with; what, if anything, can the end-user trust anymore? And perhaps, with the advances in AI technology, the answer is "no-one".

Bibliography

Abraham, S. and Chengalur-Smith, I. (Aug. 1, 2010). "An overview of social engineering malware: Trends, tactics, and implications". In: *Technology in Society*, 32(3), pp. 183–196. ISSN: 0160-791X. DOI: 10.1016/j.techsoc.2010.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0160791X10000497.

Conteh, N. and Schmick, P. (Feb. 12, 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.

Mouton, F., Leenen, L., and Venter, H. S. (June 1, 2016). "Social engineering attack examples, templates and scenarios". In: *Computers & Security*, 59, pp. 186–209. ISSN: 0167-4048. DOI: 10.1016/j.cose.2016.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0167404816300268.

Qin, T. and Burgoon, J. K. (May 2007). "An Investigation of Heuristics of Human Judgment in Detecting Deception and Potential Implications in Countering Social Engineering". In: *2007 IEEE Intelligence and Security Informatics*. 2007 IEEE Intelligence and Security Informatics, pp. 152–159. DOI: 10.1109/ISI.2007.379548. URL: https://ieeexplore.ieee.org/document/4258689.

Salahdine, F. and Kaabouch, N. (Apr. 2019). "Social Engineering Attacks: A Survey". In: *Future Internet*, 11(4). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 89. ISSN: 1999-5903. DOI: 10.3390/fi11040089. URL: https://www.mdpi.com/1999-5903/11/4/89.

Syafitri, W., Shukur, Z., Mokhtar, U. A., Sulaiman, R., and Ibrahim, M. A. (2022). "Social Engineering Attacks Prevention: A Systematic Literature Review". In: *IEEE Access*, 10. Conference Name: IEEE Access, pp. 39325–39343. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3162594. URL: https://ieeexplore.ieee.org/document/9743471.

Wang, Z., Sun, L., and Zhu, H. (2020). "Defining Social Engineering in Cybersecurity". In: *IEEE Access*, 8. Conference Name: IEEE Access, pp. 85094–85115. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2992807. URL: https://ieeexplore.ieee.org/document/9087851.