AI: The Future of Social Engineering!

Henry Collier, Ph.D.
Norwich University, Northfield, VT United States

hcollier@norwich.edu

Abstract: Artificial intelligence (AI) is at the forefront of computer science today. Everyone is talking about AI and how it is the way of the future. Companies are using machine learning (ML)algorithms to enhance their business offerings, which is showing promise in the realm of improved efficiency. The potential benefit of a fully developed AI is exceptional, but so are the threats that AI poses. While the developers of the various forms of AI are eager to be the first to create a fully functional, truly intelligent AI, they do not always consider the negative possibilities that AI creates. ChatGPT was recently used to hack itself and exposed a vulnerability in its open-source library. In addition to using AI to create hacks and exploits, AI is also being used to support social engineering efforts by creating more convincing social engineering attacks. Whether the attack is using AI to duplicate a person's voice to convince a loved one to send a gift card to get them out of jail or if it is being used to simply scrape a person's social media to develop a more precise method of attack, the concern that AI will be used for nefarious purposes is genuinely profound. This paper is a case study looking into how AI is and will be used to improve social engineering. A literature review was conducted to identify how researchers are already seeing how AI is being used and to project future threats. AI is here to stay, and the threats it brings are existential, and it is imperative that these threats are realized, and defensive measures are developed. This case study looks at how AI is and will be used to improve the efficacy of social engineering attacks.

Keywords: Social Engineering, Social Media, Artificial Intelligence, Information Security

1. Social Engineering

In the realm of information security, the human is the weakest link in the security chain (Schneier, 2004). The primary avenue of attack is Social Engineering. Social engineering is the art of tricking individuals into giving away sensitive information, that they otherwise would not give away (Nohlberg & Kowalski, 2008). It is a form of manipulation and influence used to deceive someone and it comes in many forms including the most common forms-phishing, vishing, and smishing. Phishing is a form of attack, whereby a threat actor sends someone an email trying to trick them into divulging their log-in credentials, and other sensitive information or clicking on a link that will load malware and give the attacker access to their system or lock it up with ransomware. Vishing is a form of attack whereby the threat actor calls an unsuspecting individual and tries to get them to give sensitive information, including bank account information or credit card information. The notorious vishing attack where the caller says they are from "Windows Computers and your system is infected'' is an example of this kind of attack. Smishing is a form of attack where the threat actor sends the end user a text over the SMS system. One known attack is where the text says a package is being delivered but has been incorrectly routed and asks the user to click on a link. The link will then either install malware or try to prompt the user for a credit card to cover the change in the cost of shipping due to the error. These are not the only forms of social engineering, but they are the most common and most successful. Social engineering boasts a success rate of 1 in 5 or 20% of people will fall for the tricks (Gundersen, n.d.). 74% of all cybersecurity breaches have a human element to them and are financially motivated (Verizon, 2023). Social engineers use psychological triggers like imposing a false sense of urgency, causing the victim to be afraid that if they don't act, they will become a victim, and of course offering the victim something for free, to dupe their victim.

Urgency and fear are only some of the psychological triggers used by social engineers, seven triggers have been previously identified: strong affect, overloading, reciprocation, deceptive relationships, diffusion of responsibility, authority, and integrity and consistency (Kancherla, 2020). A sense of urgency falls within a broader psychological trigger known as strong affect (Kancherla, 2020). Fear also falls within the category of strong affect because fear is a strong, emotional motivator (Andreasen, 2016). Strong affect is a situation where the social engineer uses emotion to trigger a response. In the case of urgency, the user would have a sense that if they didn't act, they would miss out, or perhaps be punished for their lack of action depending on the message of the attack.

In addition to urgency and fear, there is overloading. Overloading exists because people have very complex lives that are over-inundated with data. This can lead to too much information being sent and people becoming overburdened with the data and not evaluating the data correctly, leading to them becoming victims (Kancherla, 2020). Beyond urgency, fear, and overloading, we also see reciprocation, deceptive relationships, authority, integrity, consistency, and dispersion of responsibility (Kelly & Hadnagy, 2014) (Kancherla, 2020).

Reciprocation is related to the social engineering attack known as Quid-Pro-Quo (give and take). Deceptive relationships are related to the concept of catfishing or dating manipulation. This is where a threat actor pretends to be something they are not, and they pretend to care about the victim, preying on their emotional state. Attacks using the integrity and consistency trigger are designed to take advantage of people's inherent belief that most people are honest. The authority trigger is commonly used by social engineers. The idea is that if they send an email that appears to come from someone's supervisor, they will be more likely to do what the email says to do. The diffusion of responsibility trigger is related to the group mindset. A threat actor will try to pretend to be a member of the group to trigger a response. This is most effective in spreading disinformation, especially during political events.

Because social engineers target psychological triggers that are built into each of us, defending against such attacks is difficult because of the nature of humans. Many people do not like conflict or the way conflict impacts their emotional state, and most are taught to trust authoritative figures and comply with their instructions (Lindner, 2006). Social engineers play against these emotions when they develop their attacks (Hadnagy, 2018). 128 behavioral traits influence a person to become a victim of social engineering (Collier & Collier, 2020) (Collier, 2021). These behavioral traits are further influenced by the emotions of the moment and can be effectively used against an unsuspecting user by a social engineer (Hadnagy, 2018) (Collier, 2021). From one perspective, it is possible to think of this type of influence as a form of manipulation that uses mind perception as its basis (Anderson, 2020).

At its most basic form, social engineering is simply manipulation. One of the reasons why social engineering is so successful is because manipulation occurs at the subconscious level (Anderson, 2020). Manipulation is a form of deceit and manipulators are people who use deceptive tactics to obtain their desired result, regardless of the damage it causes their target (Anderson, 2020). Deception is built around lies and lying is something that comes naturally to people (Pace, 2017). Anyone who would try and disagree with this only needs to look at a baby who fakes hunger cries to get attention, this is the beginning of deceit, and manipulation (Pace, 2017). It is our moral code that influences us to not lie. Unfortunately, it is also our moral code that makes many individuals think that people are generally good and don't lie. This belief in the average person's goodness is one of the key factors that makes social engineering so successful. Social engineering also takes advantage of the fact that biases exist like gender bias, age bias, racial bias, and status bias (Hadnagy, 2018).

Social media increases the effectiveness of social engineers by creating a wealth of information about the target (Collier, 2020)(Collier & Morton, 2024). The social engineering process starts with Open Source Intelligence (OSINT), whereby a threat actor begins to build a profile of their target (Kelly & Hadnagy, 2014) (Hadnagy, 2018). The more they get to know the target, the better the attack will be. After OSINT comes Pretext development, whereby the threat actor uses the information gathered during OSINT to develop pretexts. A pretext is the fabricated story used by the social engineer to gain the victim's trust. This could be something as simple as posting to the victim's social media account a congruent belief in something (Kelly & Hadnagy, 2014) (Hadnagy, 2018). Take the famous actor Betty White. Ms. White was a lover of animals and supported many organizations, so it is not entirely unreasonable to think that a social engineer might use this against her. First, the attacker would post something about their love of animals, and how they support organizations like People for the Ethical Treatment of Animals or the Humane Society. This would be an example of pretexting. Following this pretext, comes the attack plan.

The attack plan is where the threat actor develops the bones of the attack (Hadnagy, 2018)(Kelly & Hadnagy, 2014). How is the attack going to occur? What methods are going to be used? What is the right timing for the attack? After the attack plan is completed, then the attack is launched. The attack on Ms. White would occur when Ms. White "likes" this post or reposts it, the attacker would then generate another message about a poor animal that needs a home, and provide a link to support this animal. The link may be an avenue to simply steal Ms. White's credit card information, but it also could be a means of loading malware on her device, so the threat actor can obtain other financial information or load ransomware. The social engineer is counting on their message to pull on Ms. Whites's heartstrings, hopefully, strong enough to get her to click on the link. Emotion is one of the strongest parts of being a human being. Emotion directly influences our decision-making process (Bechara, Damasio, & Damasio, 2000) (Emotion and Decision Making, 2015). If emotions generate empathy, then the attack is more likely to be successful due to the way the human brain works. (Kelly & Hadnagy, 2014) Once a social engineer develops an emotional bond with their victim, the logic centers and rational will temporarily shut down and logic will not be used to make the decision that leads to the target becoming a victim (Kelly & Hadnagy, 2014). People are inherently insecure and this leads to breaches. As new tools, like Artificial

Intelligence, are being developed, threat actors are learning to use them to make better, more effective social engineering attacks.

2. Artificial Intelligence

Artificial intelligence has been in the realm of science fiction for decades. Whether you think about the computer Hal in the movie 2001: A Space Odessy, SkyNet in the Terminator franchise, the Red Queen in the Resident Evil series, or any of the other movies that present AI as a theme, AI comes in many forms, many of them being malevolent. So as forms of AI like ChatGPT, and Bard are created, it is only natural to think about how these tools can be used by threat actors in their attack methodology.

When one takes a look at the power that AI brings, it is easy to see that this power can be used for both good and evil. AI can be used to reduce mistakes and human error, take over and automate repetitive tasks and processes, easily analyze and process large amounts of data, make data-driven decisions faster and more effectively, assist with predicting medical trends based on data, and solve complex problems faster than a human can. It is important to note that AI is not smarter than a human, but rather it processes data, especially large data sets, at a speed that the human mind cannot do. What AI cannot do is be creative, express emotions, make emotional decisions, or feel. For this reason, the current models of AI are still more machine learning than true AI. Because of this, AI cannot make decisions regarding whether something is or isn't moral or ethical, which is what threat actors rely on. A great example is when ChatGPT was asked to find a vulnerability within itself (Burgess, 2023). A human would not intentionally look for a weakness and then give it away, especially if the weakness could be used against them. Humans would use emotion, and feeling to determine if a request is a valid request and has merit.

AI comes with a variety of other cons. There are ethical concerns, privacy concerns, concerns that AI has the biases of the developers who created it, and then the concern that AI will take people's jobs away (Khanzode & Sarode, 2020) (Williams & Yampolskiy, 2021). Although AI is supposed to have safeguards put in place, there is also a concern that these safeguards are not fully developed, nor would they be followed if the AI did truly achieve a state of self-awareness.

If appropriate safeguards are not in place, or ignored by the AI, AI can do something counter to good order. There are several examples of this in the AI Failures Incident Database including where social media creators manipulated AI to provide inappropriate material or construct hate speech (Williams & Yampolskiy, 2021) (Burgess, 2023). In addition to this, there was an incident where AI was used to hack and spoof biometrics in the process of stealing Etherium cryptocurrency (Williams & Yampolskiy, 2021). Both of these examples demonstrate how AI can be used in a nefarious manner. These examples only brush the surface and open the door to threats coming from AI from multiple domains.

3. Social Engineering With AI

When you consider that AI has already been used to hack itself, steal cryptocurrency, and mold/manipulate social media, the jump to using AI in the development of Social Engineering attacks is small (Williams & Yampolskiy, 2021) (Burgess, 2023) (Arabo, 2023). As threat actors continue to look for new ways to target users, and users continue to make decisions based on the various psychological triggers, we will start to see rapid changes in how Social Engineering attacks are developed.

AI could be used to help reduce the number of phishing attacks that are successful, while at the same point, AI could be used to develop better phishing attacks that get through the traditional methods of detection, especially the human detection process. Think about the attack that most often impacts the elderly, the grandparent scam, whereby someone calls the grandparent in the middle of the night saying they are their grandchild and need money (Federal Communications Commission, 2023). If the person's voice on the other end isn't close to the grandchild's voice, then there is a probability that the elderly grandparent will realize it is a scam. Now, what happens if you take the same scam and you implement AI into it? We know that AI can take someone's voice and clone it pretty easily. If the grandchild is a frequent poster online on social media platforms, then a threat actor can get samples of their voice to use. They can then use the cloned voice to call the grandparent whom they identified using the same social media, which increases the likelihood of success. It has already been proven that social engineers use social media as a tromping ground (Abladi & Weir, 2018)(Collier, 2020)(Collier & Morton, 2024). Therefore everyone who has a social media account, where they don't know

personally everyone they are connected with, should assume there is at least one social engineering in their network of friends.

Threat actors are already using AI in an adversarial way to generate cyber attacks (Thota & Menaka, 2023). AI is now being used to develop more effective phishing schemes, vishing schemes, and phishing websites. AI is further being used to implement many of these attacks, reducing the operational overhead of the threat actor. SlashNext's report *The State of Phishing 2023,* reveals that phishing emails increased by 1265% due to ChatGPT, which signals a new era of cybercrime fed by Artificial Intelligence (SLASHNEXT, 2023). The report goes on to show that there has been a 967% rise in credential phishing (SLASHNEXT, 2023). Generative AI creates an environment where threat actors can generate new, more insidious phishing campaigns that have a higher success rate than in previous years (SLASHNEXT, 2023). In 2023 the phishing market saw a growth in generative AI in business email compromise (BEC) attacks (SLASHNEXT, 2023) (Violino, 2023). It was found that Worm GPT, which is an AI chatbot with a large customized language model (LLM), helped cybercriminals with numerous activities, many related to business email compromise (BEC) attacks (SLASHNEXT, 2023). Add to this problem, it is evident, ChatGPT can leak training data and violate privacy (Ray, 2023). Threat actors are using jailbreak attacks to bypass the ethics safeguards in AI, and develop harmful responses, which can be used to develop new forms of attack on end users (Xie, et al., 2023). With ChatGPT, and other forms of AI/ML in the wild, it is clear that the cyber threat landscape is rapidly changing.

On a positive note, AI is also being used to detect social engineering attacks (Basit, et al., 2020) (Dhake, et al., 2023). AI is being used in phishing attack detection strategies (Basit, et al., 2020) (Dhake, et al., 2023). In addition to phishing attack detection, AI is being used to enhance intrusion detection in the Internet of Things, which is an area that has seen significant attacks over the last few years, including the attack on Dyn, an ISP Domain Name Service provider (Wang, 2018) (Saied, et al., 2024).

The trends that are being seen with AI and cyber attacks are concerning. Some of the questions that need to be answered by researchers are as follows: If AI is responsible for these attacks, how can we defend against them? Will we use AI to respond to an AI-based attack? If so, will AI against AI create an environment on the Internet whereby it becomes useless? It is doubtful that there will be easy answers to these questions. Cybersecurity researchers must begin to develop tools and techniques that can be used to identify AI-generated attacks.

4. Conclusion

The threat of AI being used to enhance social engineering attacks is real and needs to be addressed. As AI continues to grow, both the benefits and threats will continue to grow. Cybersecurity researchers need to continue to understand how AI is being used to target users, and how new defenses can be developed to successfully defend against such attacks. As part of this process, cybersecurity researchers must do more to understand why people are susceptible, and how to strengthen their security mindsets and modify their behavioral reactions. Combining a better understanding of the user and a better understanding of how AI can be used to attack end users, is the only way we can defend against this new threat.

Acknowledgments

I would like to acknowledge and thank my friend and soon-to-be Ph.D. holder, Charlotte Morton for her exceptional advice and proofreading ability. I would further like to acknowledge and thank my wife Heidi for her support, especially when I am writing a paper one of these papers.

References

Abladi, S. M. & Weir, G. R., 2018. User characteristics that influence judgment of social engineering attacks in social networks.. *Hum. Cent. Comput. Inf. Sci,* 8(9).

Anderson, E., 2020. *The Art of Manipulation.* Middletown : s.n.

Andreasen, S., 2016. Fear: The social motivator-the only thinkg you have to fear is everything.. *Journal of Multidisciplinary Scientfic Research,* 4(2), pp. 13-18.

Anon., 2015. Emotion and Decision Making. *Annual review of psychology,* pp. 799-823.

Arabo, A., 2023. *The use AI (ChatGPT) for Offensive.* Stanford, The 2nd International Conference on Computing Innovation and Applied Physics.

Basit, A. et al., 2020. A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunications Systems,* Volume 76, pp. 139-154.

Bechara, A., Damasio , H. & Damasio, A. R., 2000. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex,* pp. 295-307.

Burgess, M., 2023. The Hacking of ChatGPT Is Just Getting Started. *WIRED*, 13 April.

Collier, H., 2021. *Enhancing Information Secuity by Identifying and Embracing Executive Functioning and the Human Behaviors Related to Susceptibility.* Colorado Springs: ProQuest.

Collier, H., 2022. *Including Human Behviors into IA Training Assessment: A Better Way Forward.* Reading, Academic Conferences International Limited.

Collier, H. & Collier, A., 2020. *The Port z3R0 Effect!: Human Behaviors Related to Susceptibility.* Copenhagen, AIRCC Publishing Corporation, p. 5.

Collier, H. D., 2020. *Social Media: A Social Engineer's Goldmine.* Larnaca, s.n.

Collier, H. & Morton, C., 2024. *Teenagers: A Social Media Threat Vector.* Johannesburg, Academic Conferences International.

Dhake, B. et al., 2023. The threat of AI being used to enahance social engineering attacks is real and needs to be addressed.. *SSR.*

Federal Communications Commission, 2023. *FCC.gov.* [Online]
Available at: https://www.fcc.gov/grandparent-scams-get-more-sophisticated
[Accessed 14 01 2024].

Gundersen, G. M., n.d. *cyberpilot.io.* [Online]
Available at: https://www.cyberpilot.io/cyberpilot-blog/does-phishing-training-work-yes-heres-proof#:~:text=The%20results%20showed%20that%20a,the%20employees%20submitted%20their%20credentials.

Hadnagy, C., 2018. *Social Engineering: The Science of Human Hacking.* Indianapolis: Wiley.

Kancherla, J., 2020. Motivational and Psychological Triggers in Social Engineering. *SSRN.*

Kelly, P. F. & Hadnagy, C., 2014. *Unmasking the Social Engineer: The Human Element of Security.* Indianapolis: John Wiley & sons inc..

Khanzode, K. C. A. & Sarode, R. D., 2020. Advantages and Disadvanteges of Artificial Intelligence and Machiner Learning: A Literature Review. *International Journal of Librry & Information Science (IJLLIS),* 9(1), pp. 30-36.

Lindner, E. G., 2006. Emotion and Conflict: Why It Is Important to Understand how Emotions Affect Conflict and How Conficlt Affects Emotions.. In: *The Handbook of Conflict Resolution.* s.l.:Jossey-Bass, pp. 286-.

Nohlberg, M. & Kowalski, S., 2008. *The Cycle of Deception - A Model of Social Engineering.* Plymouth, HAISA.org.

Pace, M., 2017. *Dark Psychology 202.* Coppell: Make Profits Easy LLC.

Ray, T., 2023. *zdnet.com.* [Online]
Available at: https://www.zdnet.com/article/chatgpt-can-leak-source-data-violate-privacy-says-googles-deepmind/
[Accessed 15 01 2024].

Saied, M., Guirguis, S. & Madbouly, M., 2024. Review of artificial intelligence for enhancing intrusion detection in the internet of things. *Engineering Applications of Artifical Intelligence,* 127(Part A).

Schneier, B., 2004. *Secrets & Lies: Digtiatl Security in a Networked World.* Indianapolis: John Wiley & Sons, Inc. .

SLASHNEXT, 2023. *slashnext.com.* [Online]
Available at: https://slashnext.com/wp-content/uploads/2023/10/SlashNext-The-State-of-Phishing-Report-2023.pdf
[Accessed 05 01 2024].

Thota, S. & Menaka, D., 2023. Botnet detection in the internet-of-things networks using convolutional neural network with pelican optimization algorithm. *Journal for Control, Measurement, Electronics, Computing and Communications,* 65(1), pp. 250-260.

Verizon, 2023. *DBIR 2023 Data Breach Investigations Report.* [Online]
Available at: https://www.verizon.com/business/resources/T616/reports/2023-data-breach-investigations-report-dbir.pdf

Violino, B., 2023. *cnbc.com.* [Online]
Available at: https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html#:~:text=Technology%20Executive%20Council-,AI%20tools%20such%20as%20ChatGPT%20are%20generating,increase%20in%20malicious%20phishing%20emails&text=Sin
[Accessed 20 01 2024].

Wang, C., 2018. *The 2016 Dyn Attack and its Lessons for IoT Security,* Stanford: Stanford Management Science and Engineering.

Williams, R. & Yampolskiy, R., 2021. Understanding and Avoiding AI Failures: A Practical Guide. *Philosophies,* 6(53).

Xie, Y. et al., 2023. Defending ChatGPT against jailbreak attack via self-reminders. *Nature maching Intelligence,* Volume 5, pp. 1486-1496.