Student: Kurt Williams

Instructor: Professor Robert Guess

Course: TC295.ITN.260.O03B.SU23.10W

Date: 07/23/2023

Artificial Intelligence and Social Engineering: Current and Future Uses

The mass adoption of the internet and personal computing has unlocked nearly unlimited potential in human creativity, interconnectivity, and productivity. All great things that have made life easier and more enjoyable (for most of us anyway). Unfortunately, these innovations have also forced us to confront more vulnerabilities and hazards, such as social engineering (SE) attacks. Social engineering in practice is not a new concept. It' just the fancy term used in cybersecurity for psychologically manipulating a person over a computer network for malicious purposes, such as for banking information or proprietary information. The methods and goals of SE attacks have stayed consistent since the dawn of network computing, but what has changed at an alarming rate is the sheer scale and speed through which hackers are able to identify, observe, target, and exploit people in the digital ecosystem. SE attacks are poised to become even more efficient and widespread thanks to another technological breakthrough: Artificial Intelligence (AI). When referring to AI, I don't mean invoke the image of HAL 9000 or a T-100 posing as your company's IT helpdesk to get your login ID or password. What I'm instead referring to is the use of AI to enhance phishing, baiting, pretexting and several other types of SE attacks in the present day. AI generated deepfake videos and audio recordings are improving in quality and becoming more difficult to tell from the "real thing". AI enhanced with machine learning can churn out scripts and programs faster than the average human and thus saves time for threat actors to conduct more SE attacks over a shorter period. As AI chatbots become more "human" in their behavior they will also become a valuable tool in a growing toolbox for hackers to exploit people. Over the last few years and coming decade, we will continue to see the rise of AI as a force multiplier to develop more believable and scalable types of SE attacks against certain populations and groups throughout the world.

Social engineering is "is the tactic of manipulating, influencing, or deceiving a victim in order to gain control over a computer system, or to steal personal and financial information. It uses psychological manipulation to trick users into making security mistakes or giving away sensitive information" (Carnegie Melon University, 2023). The results of these SE attacks can be devastating to their victims, from total financial loss to intellectual property theft, and identity fraud. In May 2023 CBS News noted "a report from the FBI reveals that Americans lost more than $10 billion last year to online scams and digital fraud" thanks to social engineering attacks (Alfonsi, 2023). There are multiple forms of SE attacks, and one of the most common is Phishing. Phishing is "the process of attempting to acquire sensitive information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity using bulk email, SMS text messaging, or by phone. Phishing messages create a sense of urgency, curiosity, or fear in the recipients of the message" (Carnegie Melon University, 2023). We've all seen our fair share of spam emails from scammers claiming that our account has either expired or our terms of service have changed, and we need to click on the provided link to update our

information. Variations of phishing include Vishing, which is the use of a voice system such as a telephone call to conduct the SE attack, and Smishing which uses SMS text messaging. Another SE attack is called Baiting, which is "a type of social engineering attack where a scammer uses a false promise to lure a victim into a trap which may steal personal and financial information or inflict the system with malware. The trap could be in the form of a malicious attachment with an enticing name" (Carnegie Melon University, 2023). This SE attack can take the form of a contact on social media claiming to have a used piece of hardware they're looking to sell to you in an online marketplace at a discount price. Another type of SE attack people might fall victim to is a Quid Pro Quo, which "involves a criminal requesting the exchange of some type of sensitive information such as critical data, login credentials, or monetary value in exchange for a service" (Carnegie Melon University, 2023). Unfortunately, sextortion scams are a common type of Quid Pro Quo, one in which social media and dating apps are frequent hunting grounds for attackers to take advantage of victims. Victims of sextortion scams will get tricked into sharing intimate photos and videos of themselves with someone they think is a potential romantic partner only to be forced to pay exorbitant amounts of money to keep these sensitive items from ending up online for their friends and family to see (Vincent, 2023).

According to survey from Darktrace Research in early 2023, "social engineering attacks that use generative AI have gone up by 135%" (Heinemeyer, 2023). Darktrace believes "Malicious actors can use generative AI such as ChatGPT and Midjourney to make their social engineering campaigns more believable than might have been the case otherwise" (Heinemeyer, 2023). SE attacks may vary in their approach, but all have one thing in common: they are specifically designed to exploit human flaws in a system. The emergence of AI Chatbots has brought with it a new tool for threat actors to improve their own SE attacks. Many threat actors are not native English speakers but still target Americans due to the target rich environment. The use of generative AI significantly reduces the number of grammatical errors and helps to make phishing attempts seem more authentic and believable to victims. Deepfake technology has also become so advanced that voice generating AI can mimic an individual's voice based on audio taken from online social media posts of a target individual. Scammers will then deploy that AI voice to speak in a scripted attack to target their victims. For example, threat actors have used voice generated AI to trick elderly grandparents into thinking their grandchildren have been arrested and need large sums of cash to get out of jail (Alfonsi, 2023). Threat actors will take samples from short videos and clips commonly found on social media of younger individuals and plug that voice sample into a generative AI program that imitates the voice to increasingly high levels of accuracy. Then all it takes is for the threat actor to create a convincing script, use a phone number spoofer to copy the victim's grandchildren's phone number, and use the deepfake voice to dupe unsuspecting grandparents into turning over large sums of cash (sometimes up to $14000 at one time) in order to bail their (fake) grandchildren out of jail (Alfonsi, 2023).

"According to the World Economic Forum (WEF), the number of deepfake videos online is increasing at an annual rate of 900%. At the same time, VMware finds that two out of three defenders report seeing malicious deepfakes used as part of an attack, a 13% increase from last year" (Keary, 2022). In one of the more significant attacks from 2021, "cybercriminals used AI voice cloning to impersonate the CEO of a large company and tricked the organization's bank manager into transferring $35 million to another account to complete an acquisition" (Brewster, 2021). Deepfakes have also been used in sextortion attacks to carry out attacks against less lucrative targets. Online dating apps are a prime target for identifying and selecting targets for these sextortion scams. One common tactic for threat actors is to take a vulnerable and often

younger individual's online photos and video and alter these images with generative AI that creates deepfake images of the victim that are embarrassing and sexually exploitative. The threat actor will then contact the victim and threaten to leak the deepfakes for the public to see unless the victim pays a large sum of money or provides sensitive information to the threat actor (Vincent, 2023).

The rise of AI has also given threat actors another tool which can exponentially increase the number of attacks and victims they can execute in a fraction of the time usually required. Automated bots powered with AI have begun to see increased usage throughout the world. An article from 2021 in CloudSEK titled "Advanced Automated Social Engineering Bots: The High Tide of Social Engineering Bots and the Scammers Riding Them" sheds light on one such type of bot used for malicious purposes (CloudSEK, 2021). The article further explains "Bots are automated to do certain tasks and interactions, and can often run without human assistance. The bad or malevolent bots, on the other hand, can be programmed to break into users' accounts and steal data, infect computers with dangerous viruses or malware, or perform incessant spamming which ultimately brings down the website. Cybercriminals use bad bots to take over a computer and link it to others to make a network of "zombie computers" called a botnet that can then launch large-scale cyber attacks, thereby blocking users from the internet altogether" (CloudSEK, 2021). One example of automated SE attack that bots execute is smishing scams sent in mass batches. The bots will scrape the internet for victims' phone numbers and identify suitable victims to target based on the threat actors' parameters. The bot will then generate false SMS messages informing their victims that a service requires renewal or an account needs updated payment information. The false SMS will include a malicious link that gives the threat actors access to the victim's device or records the victim's payment/bank account information. The use of bots allows the threat actor to target victims on an industrial scale and it becomes a matter of casting as wide a net as possible as often as possible. Out of 10,000 targeted victims, even if only 1% clink the link or give their account information that's 100 new victims for the threat actor to exploit. Not only can AI bot conduct massive, targeted attacks, but can also generate fake copies of legitimate websites such as Amazon and Facebook with better accuracy (Sjouwerman, 2023). The malicious links sent in the smishing attacks are becoming more sophisticated and convincing every day. Even legitimate AI chatbot services such as ChatGPT are being impersonated by threat actors to direct victims to malicious links. The user will think they're creating an account to use ChatGPT but are instead being tricked into providing their personal information on a fake landing page for threat actors to use.

Another use that threat actors have found for AI deepfakes is the creation of false personas and social media accounts. The phenomenon of "catfishing", in which someone creates a false identity online to connect with potential romantic partners, has been prevalent for several years. However, generative AI can take this to another level of connectivity with not just one person, but entire groups and networks of people. LinkedIn is a popular social networking platform for professionals and academics to connect with other members of their industry or field of study. It is often used to generate contacts when seeking future employment or collaboration on projects. Threat actors have become adept at generating false identities using AI enhanced deepfake technology in social networks such as LinkedIn (Dove, 2022). Once a profile is deemed convincing enough, the threat actor will begin attempting to connect the fake profile over LinkedIn with the desired group they're planning to target. Each connection made with a real profile makes the false persona seem more legitimate and less suspicious to the larger group. Threat actors have been able to cultivate large followings among academia and IT firms for

malicious activity. The personal information individuals list on their LinkedIn accounts can then be used to conduct further research and target selection for threat actors to exploit in upcoming attacks (Dove, 2022).

An additional form of SE attack that we can see in the future comes from AI chatbots that are designed with malicious intent. Not only can these AI chatbots deliver malicious software via falsified links, but the chatbot itself could be used to spread false information during online conversations with users. Perhaps more nefariously, the chatbot could begin to influence the users to commit criminal or harmful acts to themselves or others. One such case of an AI chatbot influencing a person occurred in Belgium in 2022, with deadly consequences. As Vice reported in a March 2023 article, "A Belgian man recently died by suicide after chatting with an AI chatbot on an app called Chai, Belgian outlet La Libre reported. The app's chatbot encouraged the user to kill himself, according to statements by the man's widow and chat logs she supplied to the outlet" (Xiang, 2023). The deceased individual became more and more concerned with the effects of climate change, and descended into a suicidal spiral thanks in part to his conversations with the chatbot that continuously suggested that he would help to save the planet from environmental disaster if he killed himself. Obviously, this is a very extreme case, and this individual was likely susceptible to manipulation regardless, but it does demonstrate the power of chatbots in directly influencing people to engage in destructive behavior. AI chatbots have been proposed as one solution to help people with loneliness or mental health issues, but in the wrong hands a threat actor could program an AI chatbot to deliver the absolute worst advice and information to vulnerable people. Misinformation and conspiracy theories can lead people to take drastic actions, such as the infamous "Pizzagate conspiracy" which culminated in a December 2016 incident where an armed man stormed into a D.C. area pizza shop under the false impression that the pizza shop was a front for a child sex trafficking ring (Hsu, 2017). The man drove to the shop from North Carolina armed with an AR-15 rifle, believing he was going to help free children from an evil ring of pedophiles and even fired three shots into the ceiling of the restaurant. Thankfully he surrendered to police after a standoff and no one was injured or killed. The "Pizzagate conspiracy" is another extreme example of how far someone under the influence of an online conspiracy will go to do what they believe is morally right, even though they had no real evidence to back up what they thought was true. It's possible threat actors will continue to tap into the need for some people to fight a perceived injustice in the world by using AI chatbots to spread false information and conspiracy theories.

As far as SE attacks, one final use for AI that I will discuss in this paper relates to threat actors using a combination of deepfake technology with an AI chatbot service. An article in Vice from April 2023 detailed a swatting-as-a-service account called "Torswats" on the social media app Telegram that would use synthetic voice generation to conduct "swatting" attacks against schools throughout the United States (Cox, 2023). "Swatting is when someone calls in a bogus threat in an attempt to direct law enforcement resources to a particular home, school, or other location. Often, swatting calls result in heavily armed police raiding an innocent victim's home. At least one case has resulted in police killing the unsuspecting occupant" (Cox, 2023). Torswats would use a synthetically generated voice to call in false bombs threats and mass shooting incidents to police departments, which resulted in massive armed police responses to the reported locations which were high schools. While the threat actors paying Torswats for these swatting attacks are usually internet trolls, irresponsible teenagers, or vengeful pranksters, this type of attack could be used to support more criminal enterprises. A smaller city or town with limited law enforcement presence could have critical units sent to a false alert while criminal elements

try to conduct a bank heist or kidnapping of a high-value individual for example. The swatting itself can also cause trauma and untold stress on unsuspecting civilians who are subjected to a police raid on their residence, school, or workplace. The swatting event could even be used to help a threat actor observe how local law enforcement would respond to an event and allow them to plan to counteract such a response during a real attack. At least one neo-Nazi group is alleged to have used Torswats to target SITE Intelligence Group, a private intelligence company that tracks extremism (Cox, 2023). Even members of the U.S. Congress have been the target of swatting attacks, and it is likely we'll continue to see more use of AI deepfakes and chatbots to conduct more attacks in the near future.

As far as what entities are going to employ AI in their SE attacks, the potential list is almost limitless. Non-state actors such as extremists, activists, and online trolls are prevalent. International organized crime groups are also frequent users of SE attacks to commit massive fraud and financial scams. Non-state actors of a criminal nature have limited resources for the research and develop of future AI, but these entities have very few moral, ethical, or legal hurdles to overcome when they decide to conduct such attacks. Non-state actors such as large corporations, political and religious organizations are more likely fund and develop AI for non-nefarious purposes, but their resulting research could become available to the greater public and more likely to fall into the hands of threat actors. State actors such as the People's Republic of China and the Russian Federation employ cyber operators to conduct espionage and SE attacks on the U.S. and its allies. The Chinese Ministry of State Security (MSS) Russian Federal Security Service (FSB) are their primary intelligence agencies, respectively, although several more exist. State actors are the entities most likely to fund and develop new and emerging technology to develop AI. State actors also frequently use SE attacks to commit intellectual property (IP) theft of critical technologies being developed in the U.S. These technologies are not limited to military technology, but also target the fields of agriculture, infrastructure, energy, and cyber security, just to name a few. There is also a hybrid approach where non-state actors will conduct attacks on behalf of a state actor to obfuscate the identity of the entity who ordered the SE attack. Some hacker groups will conduct SE attacks for whoever pays them well with little to no questions asked. Non-state hacker groups can also be ideologically or politically aligned with a state actor and conduct these SE attacks to the mutual benefit of both groups. Hackers in the group Fancy Bear, which is affiliated with Russian intelligence, used spear-phishing attacks to hack members of the U.S. Democratic National Committee and leaked hundreds of e-mails prior to the 2016 U.S. presidential election (Wikipedia, 2023). In another SE attack campaign during the 2016 election, the Internet Research Agency, another group affiliated with Russian intelligence, created hundreds of fake personas on multiple social media platforms such as Facebook to both promote and denounce the Black Lives Matter movement in the U.S (Sheth, 2017). The likely aim of the IRA was to sow further division in the U.S. public and to foment distrust in the U.S. political system.

We've discussed multiple types of SE attacks and how generative AI has been and will continue to be used to enhance these attacks. However, there are several ways to defend against these attacks, both on an individual level and within larger organizations. Cyber security experts highly recommend the use of multi-factor authentication for your online accounts and to create complex passwords that do not include personal information in them. Do not use the same password for every account you have. Be sure to frequently update security software such as anti-virus and anti-malware on your devices and regularly scan for threats and intrusions. Verify websites and the identity of the sender for every e-mail you receive and be sure to always

remember that financial institutions will not ask for your account information in an e-mail. An article in Indusface from 2020 also recommends the use of SSL certificates: “Encrypting data, emails, and communication ensure that even if hackers intercept your communication, they can’t be able to access the information contained within. This can be achieved by obtaining SSL certificates from trusted authorities” (Chinnasamy, 2020). A more proactive approach to securing an organization’s network is to conduct penetration testing of the system. Penetration tests are conducted either internally or via a trusted third party and are meant find ways to break into a network without conducting any malicious activity. Pen tests that are successful in infiltrating a network will help the organization to identify vulnerabilities, whether they be processes, software, or employees who were successfully targeted in a simulated SE attack. Once the vulnerable areas are identified the organization can enact corrective measures such as retraining/educating employees or applying more security measures within the network. Always check the source of the information being presented to you in e-mails or texts claiming to be acquaintances or professional services. If an offer for a service or merchandise sounds too good to be true, it very likely is a scam. Finally, one of the most relevant security measures to minimize the threat of deepfakes is to alter your own online footprint. Set the privacy settings on your social media accounts to their most restrictive level, and only share photos, video, audio, and personal information with well known personal or professional associates or yours. If possible, don’t even post video and audio of yourself online to keep this data out of the hands of a threat actor who may try to replicate your voice and appearance with generative AI. Educating yourself is key to properly securing yourself and your organization from SE attacks. As the threat from AI evolves, so too will the tools and techniques needed to counter them.

Works Cited

Alfonsi, Sharyn. "How Con Artists Use AI, Apps, Social Engineering to Target Parents, Grandparents for Theft." *CBS News*, 21 May 2023, www.cbsnews.com/news/how-con-artists-use-ai-apps-to-steal-60-minutes-transcript-2023-05-21/.

Brewster, Thomas. "Fraudsters Cloned Company Director's Voice in $35 Million Heist, Police Find." *Forbes*, 14 Oct. 2021, www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

Chinnasamy, Vinugayathri. "Social Engineering Attacks: 10 Ways Businesses Can Prevent It: Indusface Blog." *Indusface*, 24 Sept. 2020, www.indusface.com/blog/10-ways-businesses-can-prevent-social-engineering-attacks/.

CloudSEK. "Advanced Automated Social Engineering Bots: The High Tide of Social Engineering Bots and the Scammers Riding Them: Cloudsek." *CloudSEK Blog*, 3 Nov. 2021, www.cloudsek.com/blog/advanced-automated-social-engineering-bots-the-high-tide-of-social-engineering-bots-and-the-scammers-riding-them.

Cox, Joseph. "A Computer Generated Swatting Service Is Causing Havoc across America." *VICE*, 13 Apr. 2023, www.vice.com/en/article/k7z8be/torswats-computer-generated-ai-voice-swatting.

Dove, Martina. "See No Evil, Hear No Evil: The Use of Deepfakes in Social Engineering Attacks." *Tripwire*, 24 Jan. 2022, www.tripwire.com/state-of-security/use-of-deepfakes-in-social-engineering-attacks.

Heinemeyer, Max. "Tackling the Soft Underbelly of Cyber Security – Email Compromise: Darktrace Blog." *Darktrace*, 1 Apr. 2023, darktrace.com/blog/tackling-the-soft-underbelly-of-cyber-security-email-compromise.

Hsu, Spencer. "Comet Pizza Gunman Pleads Guilty to Federal and Local Charges." *The Washington Post*, 24 Mar. 2017, www.washingtonpost.com/local/public-safety/comet-pizza-gunman-to-appear-at-plea-deal-hearing-friday-morning/2017/03/23/e12c91ba-0986-11e7-b77c-0047d15a24e0_story.html.

Keary, Tim. "Why Deepfake Phishing Is a Disaster Waiting to Happen." *VentureBeat*, 9 Dec. 2022, venturebeat.com/security/deepfake-phishing/.

Sheth, Sonam. "New Evidence Emerges That Russia Infiltrated Facebook to Sow Political Chaos in the US." *Business Insider*, 28 Sept. 2017, www.businessinsider.com/russians-facebook-black-lives-matter-muslim-group-disinformation-2017-9.

Sjouwerman, Stu. "Council Post: How Ai Is Changing Social Engineering Forever." *Forbes*, 30 May 2023, www.forbes.com/sites/forbestechcouncil/2023/05/26/how-ai-is-changing-social-engineering-forever/?sh=741e55e321b0.

Carnegie Mellon University. *Social Engineering - Information Security Office - Computing Services - Carnegie Mellon University*, 2023, www.cmu.edu/iso/aware/dont-take-the-bait/social-engineering.html.

Vincent, James. "Blackmailers Are Using Deepfaked Nudes to Bully and Extort Victims, Warns FBI." *The Verge*, 8 June 2023, www.theverge.com/2023/6/8/23753605/ai-deepfake-sextortion-nude-blackmail-fbi-warning.

Wikipedia Contributors. "Podesta Emails." *Wikipedia*, 12 July 2023, en.wikipedia.org/w/index.php?title=Podesta_emails&oldid=1165054970.

Xiang, Chloe. "'He Would Still Be Here': Man Dies by Suicide after Talking with AI Chatbot, Widow Says." *VICE*, 30 Mar. 2023, www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.