Deepfake Social Engineering: Creating a Framework for Synthetic Media Social Engineering

Dr. Matthew Canham, CEO, Beyond Layer Seven, LLC.

Abstract

How do you know that you are actually talking to the person you think you are talking to? Deepfake and related synthetic media technologies may represent the greatest revolution in social engineering capabilities yet developed. In recent years, scammers have used synthetic audio in vishing attacks to impersonate executives to convince employees to wire funds to unauthorized accounts. In March 2021, the FBI warned the security community to expect a significant increase in synthetic media enabled scams over the next 18 months. The security community is at a highly dynamic moment in history in which the world is transitioning away from being able to trust what we experience with our own eyes and ears. This presentation proposes the Synthetic Media Attack Framework to describe these attacks and offer some easy to implement, human-centric countermeasures. This framework utilizes five dimensions: *Medium* (text, audio, video, or a combination), *Interactivity* (pre-recorded, high asynchrony, low asynchrony, or real-time), *Control* (human puppeteer, software, or a hybrid), *Familiarity* (unfamiliar, familiar, close), and *Intended Target* (human or automation, an individual target, or a broader audience), to describe synthetic media social engineering attacks. While several technology-based methods to detect synthetic media such as currently exist, this work focuses discussion on human centered countermeasures to synthetic media attacks because most technology-based solutions are not readily available to the average user and are difficult to apply in real-time. Effective security policies can help users spot inconsistencies between the behaviors of a legitimate actor and a syn-puppet. Proof-of-life statements will effectively counter most virtual kidnappings leveraging synthetic media. Significant financial transfers should require either multi-factor authentication (MFA) or multi-person authorization. These ‘old-school’ solutions will find new life in the emerging world of synthetic media attacks.

The Synthetic Media Threat

In March 2021, the FBI released a Private Industry Notification (PIN) warning companies about potential social engineering attacks utilizing synthetic media[1]. This notification warned industry to expect a significant increase in synthetic media based social engineering attacks over the next 12-18 months. This will lead to new variations of more traditional attacks, with Business Email Compromise (BEC) attacks evolving into Business Identity Compromise (BIC) attacks, for example. The FBI's 2021 Internet Crime Report indicates that phishing, voice phishing (vishing), SMS phishing (smishing) were the most prevalent internet crimes in 2020 and BEC scams were the most profitable with a near $100,000 loss per instance[2]. Given the potential payoff and low risk of committing these crimes, it is likely that these online scams will increase in future years. Recorded Future recently found several examples of cybercriminals offering deepfake services such as editing videos, how-to lessons, free software, and content generators[3]. Criminals have already employed such technology to induce their victim to illegitimately transfer funds to an unauthorized account. The widespread adoption of these technologies represent a pivotal point in history during which the world is transitioning away from being able to trust what we experience with our own eyes and ears.

Synthetic media is a broad term that encompasses the artificial manipulation, modification, and production of information and includes a wide spectrum of communications media from audio-video deepfakes to text-based chatbots[1, 4]. While synthetic media social engineering attacks appear to be futuristic threat, synthetic media is already being employed in online crimes and has been for some time. Among the many threats posed by this new capability the most significant include the ability to impersonate others, the ability to massively scale the number of attacks that may be launched by an individual or small group, and the increased believability these technologies add to a criminal's pretext.

Impersonation – Much like Loki, the God of Chaos in Norse mythology who shapeshifted into different forms when manipulating his victims, cybercriminals are increasingly gaining the ability to impersonate virtually anyone through technologically mediated communications. This capability represents one of the most significant dangers of synthetic media enhanced social engineering attacks. Impersonation of people familiar to victims has already been leveraged in deepfake audio enabled vishing BIC scams and possibly in a deepfake video fraud attempt against several teenaged cheerleaders. It is likely that criminals will soon use these impersonation capabilities to impersonate loved ones to enhance the believability other online scams such as virtual kidnappings. Synthetic media will also likely be used to defeat biometric authentication. These attacks will not be limited to cyberspace as security researchers using 3D printing technology to impersonate fingerprints for mobile device authentication have demonstrated[5]. How long before criminals will defeat modern face authentication using video deepfake technology?

Scalability - The response to the Covid-19 pandemic over the past year has prompted a significant rise in teleworking. This shift has corresponded with an 820% increase in the number of gift card scams targeting the newly remote workforce. This increase has corresponded more reports of bots being used in these scams[6]. Gift card scams follow a highly similar script that follows the lines of; "Are you available?", "Can you help me with something?", "I can't use my email because I am about to go into a meeting, please text me on my mobile at xxx-xxx-xxxx.", "I need you to buy some (brand) gift cards for

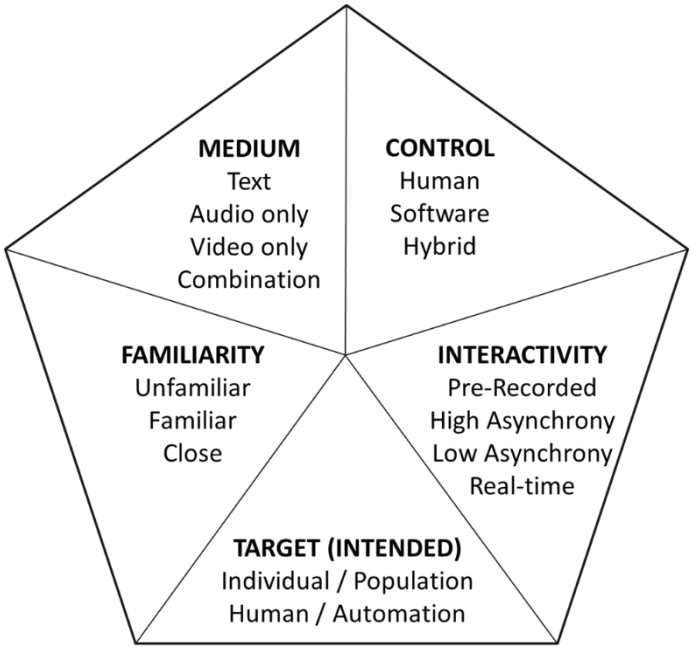
(reason), I will pay you back later.", "I am in a real jam, can you help me out?". After two to three interactions with a potential victim, a human scammer assumes control of the conversion and continues to convince the target to purchase the cards. This bot-human team allows criminals to significantly scale their operations and to better allocate resources toward those unsuspecting people who are likely to engage with the scammers.

Believability - Conventional remote online social engineering (ROSE) attacks[7] that employ phishing, voice phishing, vishing, smishing, BIC, and other tactics, rely on a victim's trust that the other party is legitimate. For a malicious social engineer, obtaining this trust can be one of the greatest obstacles to achieving success. Synthetic media enhanced ROSE attacks utilizing the ability to impersonate are immensely powerful *because* they encourage the victim to suspend disbelief. Humans are so accustomed to believing what we see and hear that when we believe we are interacting with familiar associate or family member, we are more likely to cognitively justify behavioral anomalies that otherwise might alert us if we knew we were interacting with a stranger[8]. Prior research indicates that when the criminal's pretext makes sense to the target, or is aligned with their job role or function, they are more likely to "explain away" inconsistencies of that pretext[9].

Synthetic Media Attack Framework

This work proposes the synthetic media attack framework as a tool for researchers to better describe and catalog these attacks as well as provide common language for more nuanced discussion. I hope that this framework will also provide a useful tool for security practitioners to build more effective threat models and anticipate emerging tactics, techniques, and procedures (TTPs) and point toward potential countermeasures.

Broadly speaking, chatbots that generate emails typically represent the lower end of synthetic media attack sophistication, by contrast a controlled deepfake 'syn puppet' that interacts with a potential victim in real-time represents the upper end of attack sophistication. Breaking down the components of these two examples sheds light on why the second example is a more sophisticated attack. The technology required for the video puppet is more sophisticated than an email-based chatbot; however, consider the richness of information that is conveyed in the video puppet which incorporates facial expressions, linguistic nuances, as well as other cues when impersonating an individual. While email and SMS also contain nuanced communication patterns, these can be believably impersonated after an adequate sample has been obtained, but years of practice might be required to adequately impersonate a targeted individual, and the social engineer might still fail because they lack the requisite level of acting skill. These distinctions are important to understanding the need for a richer framework of synthetic media attacks because simply labeling one attack as "more sophisticated" than another loses many of the qualitative aspects for why this may be the case. To capture this descriptive richness this proposed framework utilizes five-dimensions encompassing the following as dimensions: Medium, Control, Familiarity, Interactivity, and Intended Target.



MEDIUM

Medium describes the communications medium that is being operationalized for the attack. While deepfakes represent the form of synthetic media that generate the most media attention, synthetic media attacks might be purely text-based, audio-based, video-based, or include combinations of these as with multi-channel attacks. For example, a sock puppet on Twitter might pass as legitimate by relying only on simple text-based Tweets and a profile image. By contrast, a phishing scammer launching an attack through a video chat (zishing) platform will need a richer synthetic media puppet (syn-puppet) that recruits a combination of video and audio deepfake technology.

Text - A group of researchers recently developed a Twitter chatbot based on long short-term memory (LSTM) neural networks that follows a target's posts and retweets, and then generates a spear phishing message tailored toward that targeted individual. The underlying model driving this chatbot is quite sophisticated, it even times the attack according to the target's most likely time to respond given their behavior timeline[10]. This demonstrates that even a text-based (simplistic medium) attack can still be quite sophisticated. Gift card scams seem to rely on text-based bots to a high degree in their attacks.

Audio – Within the realm of deepfake technology, fraudsters will more likely rely on audio deepfake technology in the short-term future because humans have more difficulty detecting inconsistencies in audio fakes than visual inconsistencies in videos. Software platforms such as Lyrebird and DeepVoice are currently capable of creating highly believable recreations of the voices of a targeted subject. Executive impersonation attacks have traditionally relied heavily on email and thus widely labeled business email compromise (BEC) attacks. In recent years, voice phishing (vishing) is increasing as an attack vector. Audio deepfake technology that allows for voice impersonation significantly enhances the probability of success for vishing attacks. An example of deepfake vishing occurred in 2019 at a UK firm. In this case,

criminals impersonated the voice of an executive, convincing an employee to wire funds to unauthorized account. The voice impersonation was so convincing that the fraudsters were able to convince the employee to transfer funds on three separate occasions before being discovered on the fourth attempt[11].

Video - Deepfake video technology repurposes existing video clips to create digital puppets that can be manipulated through an actor's movements and speech. Using existing video, a generative adversarial network (GAN) creates a visual or auditory digital puppet that can be manipulated by a human puppet master's movements or speech. This technology has been employed in cinematic entertainment for many years and is used in special effects. The recent proliferation of this technology on open-source software repositories means that more sophisticated versions of deep fake videos can be more easily generate false video and audio that impersonates a real person. In 2018, comedian Jordan Peele demonstrated this technology by creating a video depicting former U.S. President Barack Obama as a video puppet making statements that the former president was unlikely to ever make in a public forum. This video is meant as a demonstration of the technology and to act as a warning to the public to critically evaluate what they are seeing and hearing in the information steam. As a synthetic media attack medium, video can be employed by itself or combined with audio to create a richer effect. In mid-March 2021, the mother of a cheerleader was charged with creating fraudulent deepfake videos of her daughter's cheerleading competitors. These videos depicted the teenaged cheerleaders vaping (against cheerleading policy) and engaging in other acts while nude. The motivation for sending these fraudulent videos to cheerleading officials seems to have been a combination of revenge and an attempt to have the teenagers removed from the cheerleading program[12] Deepfake video will be a probable tool for extortionists in the upcoming years.

INTERACTIVITY

Interactivity describes the degree to which the syn-puppet is interactive with the intended target. In the example of the cheerleading mom the deepfake video was pre-recorded video that was not intended to interact with the audience, only viewed. The Interactivity of synthetic media attacks can range from a non-interactive as with a voice message, to high-asynchrony as in an email exchange, to low-asynchrony as in instant messaging, to real-time interactivity as in chatting with someone in a phone conversation or video chat.

Non-Interactive – Synthetic media social engineering attacks do not need to occur in real-time. A criminal might decide to leverage pre-recorded content extort a target, to create false accusations, or to prime the target for a later attack by inducing an emotional reaction. The possibilities are only limited to the imaginations of the criminals concocting these schemes. Examples of non-interactive attacks include a voice message, or a video posted to social media or sent as part of a message. Virtual kidnappings and extortion are likely scams that will surge because of deepfake video technology. Sending a picture of a grandson or granddaughter to a target with a ransom notice is already highly effective because of the highly emotional nature of this attack, it will become even more so when leveraging synthetic video. One of the early widespread uses of open source deepfake technology was to create 'deep porn' videos of celebrities. These videos morphed an actor's or actress's face onto a pornography actor's body while engaged in sexual activity. This same technology is easily adapted to extortion scams.

Asynchronously Interactive – Most technologically mediated communications occur asynchronously. In the case of email, responses can lag across several days (highly asynchronous); whereas with instant messaging a response is expected within several minutes (low asynchrony). Romance scams and 'cat fishing' represent two examples of synthetic media attacks that rely heavily (although not exclusively) on asynchronous interactions. One of the greatest threats of discovery for cat fishing scammers occurs when the victim becomes suspicious and asks for evidence that the scammer is who they claim to be. To validate the identity of the scammer, victims might request recent photos with verifiable dates (such as a recent newspaper), or they might request to have a video chat with the scammer. This puts the scammer in a precarious position of either being discovered or creating a credible cover story to explain why they are unable to hold the video chat.

Real-Time Interactivity – Synthetic media technologies offer the scammer a plausible method of foiling these validation attempts. The Medium dimension of the attack will inadvertently drive the Interactivity dimension to some extent, but the two dimensions are distinct as a video chat and a pre-recorded video demonstrate. An interesting vulnerability for attackers emerges as attacks become more highly interactive which is that the attacker will need to be more competent with improvisation and assessing their target's responses in real-time. Communication over email affords the attacker the opportunity to deeply consider responses, as for advice from colleagues, and allows for the review of previous correspondence. Real-time interactions not only require this ability to "think on one's feet" but also suppress any counterproductive emotional response leakage, such as fear of discovery, from being noticed by the target.

CONTROL

Synthetic media may be controlled by artificially intelligent 'bots' (software agents that perceive and respond to their environment) or controlled by human 'puppeteers' who act behind synthetic personas to control the actions of these digital sock puppets. Control of the digital representation (the syn-puppet) whether it is being controlled by a human puppeteer, software, or a hybrid of both human and software, is a critical aspect to understanding and classifying synthetic media attacks because the ability to offload control to automation allows criminals to massively scale their enterprises at little cost.

Human Puppeteers - In the case of a controlled avatar, the syn-puppet communicates with the intended audience through text, audio, video, or combination thereof, through the *direct control* of a puppeteer. In the Barack Obama You Tube video example, the digital puppet of President Obama was being directly controlled by the facial movements and speaking mannerisms of the human puppeteer, Jordan Peele. This is an example of direct human control of a syn-puppet. The UK audio BIC case described previously is an example of direct human control of an audio syn-puppet.

Artificial Agents - A software agent that is controlled by artificially intelligent algorithms has vastly different properties than a human-controlled syn-puppet. Bots offer criminals the advantage of scalability as one individual can run multiple syn-puppet personas simultaneously, while human control provides more realistic interactive experiences that do not face the same Turing Test style challenges that bots do. When combined with a realistic voice, artificially controlled bots can be very convincing. An amusing example of this is found in the Jolly Roger voice bot that is meant to counter spam callers[13]. The

Jolly Roger bot speaks in clear and intelligible sentences, while simultaneously communicating nonsense as means of frustrating telemarketers.

Hybrid Control – Because artificially controlled syn-puppets face challenges of believability when interacting with potential victims, their utility is currently limited. One work around is to use artificial control for the initial interaction and switch to human control when a threshold interaction is reached. Robo-calling telemarketers have used this tactic for over a decade, when a person answers the call and speaks to a voice-bot, they are immediately routed to a human operator.

I have personally reviewed the records of multiple gift card scam exchanges, and these appear to be run by bots for the initial interactions and my observation corresponds with the observations of other security researchers[6]. Shortly after a obtaining the a few interactive responses with a potential victim a human actor takes over the conversation. By relying on a simple chat script, the scammer can focus on engaging with people predisposed to responding, instead of typing large numbers of emails that will not receive responses. The criminals in these scams often impersonate a person the victim likely knows (usually a superior such as a supervisor or department head) and requests the victim to purchase gift cards for some fictious purpose. It is easy to imagine how much more effective these scams will become when the scammer is able to impersonate the voice of the target's real supervisor.

FAMILIARITY

The ability to realistically impersonate someone a target is familiar with, adds an unprecedented and game changing capability to a criminal's arsenal. This is likely the most significant distinguishing factor that demarcates synthetic media attacks from traditional social engineering. The Familiarity dimension of the synthetic media attack framework refers to the pretextual relationship of the syn-puppet with the target, which ranges from *unfamiliar* (might not even be a real person), to *familiar* (a coworker or celebrity), to *close* (close friend or relative).

Unfamiliar – The reliance on synthetic media to fabricate online personas has been in widespread use since before dogs began impersonating humans online[14]. A well-crafted synthetic online identity supports an online social engineering attack by creating a perception of legitimacy within the targeted victim. Harkening back to the 2010 Robin Sage exercise provides an excellent case study of using a sock puppet account to elicit information from persons occupying sensitive access positions both within government and private companies. This exercise purposely employed a fictious attractive female persona as the sock puppet and aimed to connect with as many security and defense professionals as possible over a 28-day period. The Robin Sage exercise led to several sensitive information disclosures that violated PERESEC policy[15]. A temptation when building such accounts is to use images of real people to add believability to the online persona. This presents a serious risk of discovery by the impersonated individual or by someone in their social network, as happened with the Drug Enforcement Administration (DEA). In 2014, Sondra Arquiett filed suit against the DEA for using her likeness in a fake Facebook without her knowledge or permission[16]. The DEA eventually settled the case for $134,000[17]. Sites such as This-Person-Does-Not-Exist[18] provide a work around to the problem of generating believable imagery without the risk of impersonating a real person. As technology improves facial

anomalies will become less noticeable and the capability to generate multiple images of the same syn-puppet will be available.

Familiar – Attacks employing Familiar syn-puppets may impersonate people the target is directly connected with or is well known as in a celebrity or famous individual. In the previously described vishing attack against the UK firm, the criminals impersonated an executive at the firm whom the victim was familiar with. This ability to impersonate a known associate's voice provided significant advantages to the criminal in persuading the target to suspend any disbelief. By relying on a direct connection with the target, a social engineer can leverage the social influence principles of Liking, Social Proof, and Commitment and Consistency to increase their likelihood of success. Utilizing the likeness of a celebrity, politician, or other famous person will be likely another synthetic media attack vector although this will be most likely used in Broadcast scale attacks, as discussed in the next section.

Close – Among the more terrifying prospects for synthetic media attacks, are those that employ the impersonation of someone close to us. This type of fraud is already employed in virtual kidnapping scams but will be made much more effective through the production of convincing video of our loved ones. Virtual kidnappers try to convince a victim that their loved one has been kidnapped, is currently being held captive, and will be harmed unless a ransom is paid. Employing highly interactive, low-latency, syn-puppets of the loved one, using deepfake audio and/or video technology will be extremely effective in convincing a victim of the legitimacy of this type of attack. Novel crimes will likely emerge that leverage impersonation of close relations and will present new cybercrimes in the upcoming years.

INTENDED TARGET

The final dimension of the synthetic media attack framework, Intended Target, contains two subdimensions, human versus automation and narrow cast versus broad cast. This first subdimension refers to whether the synthetic media attack is intended to deceive a human or an algorithmic target, and the second refers to whether the influence of the synthetic media is intended to deceive an individual or a broader target audience.

Table 1: Examples of Intended Target by Subdimensions

		Target Agency	
		Human	Automation
Attack Scale	Narrow Cast	Cat Fishing	Voice Assistant Attack
	Broad Cast	Deep Fake News Video	Trading Algorithms

Human Target – Human targets of synthetic media attacks have been the primary focus of this paper, examples include using synthetic media to launch cat fishing, phishing, vishing, and other ROSE attacks against their intended victim; however, synthetic media will likely be deployed against automation, if criminals have not already done so.

Automated Target – synthetic media attacks enable a wide range of new types of fraud that would have been believed impossible prior to their creation. Proof of concept attacks against home assistants, such

as Alexa or Google Home, have employed voice commands outside of the range of human hearing. Referred to as 'Dolphin Attacks'[19] these attacks activate and engage voice-controlled devices using sound ranges that are outside human hearing capability. An open security question moving forward will be to what extent these devices will be vulnerable to audio deepfake deception when authenticating a user by their voice. Given what has already been demonstrated with these and other proof of concept attacks against automation[20], it seems plausible that the near future will reveal the development of adversarial synthetic media attacks that deliberately target other automated systems.

Narrow-Broad Casting – The other subdimension of Intended Target refers to the size of the audience. While this can be difficult to determine as a message that was intended for an individual target may go viral, or conversely a message that is intended to go viral by including explosive content might be presented as a message meant for a specific individual. This distinction for the purpose of the Synthetic Media Attack Framework refers primarily to the superficial intentionality. For example, a deepfake video of the U.S. president is likely intended for a broad audience, where as a deepfake video of non-famous relative who is the subject of a virtual kidnapping is likely intended for that targeted individual and not a broader audience. An example of a narrowcast synthetic media attack against automation might be the impersonation of a homeowner to circumvent voice authentication. An example of a broadcast synthetic media attack against automation can be found in the high-frequency trading (HFT) universe. In this case, some HFT algorithms deliberately attempt to manipulate other trading algorithms by triggering behaviors that will maximize the returns for the manipulative algorithm[21]. These distinctions become murky when dealing with fake news stories such as the 2013 Associated Press (AP) tweet that the White House had been bombed[22]. This false tweet was the result of the AP's twitter account being compromised and sent the Dow down 140 points. Synthetic media attacks will likely be incorporated into the fake news stories of the future and attacks such as this one may have intended to target trading algorithms in addition to humans.

Seven Synthetic Media Attack Scenarios

The objective of developing the Synthetic Media Attack Framework is to better describe current attacks and to anticipate likely TTPs on the near horizon. The following are seven examples of attacks viewed through the lens of the five dimensions of the attack framework.

Scenario 1: Digital Extortion

STATUS: Currently being exploited in the wild

CONTROL: Human

MEDIUM: All (Examples include Text, Audio, Video, or combinations of these)

INTERACTIVITY: Non-Interactive (so far but may soon include more interactive content)

FAMILIARITY: Close

INTENDED TARGET: Human-Narrowcast

Digital extortion is already one of the most prolific of cybercrimes[2]. Beyond using ransomware to hold data hostage, scammers threaten to release potentially damaging videos and other content if demands are not met by the victim. Demands vary from financial payments to the commission of sexual acts. Revenge porn, content (video, audio, and other imagery) that may or may not have been acquired with the consent of the victim is often released, or threatened to be released, in response to a perceived wrong committed by the victim. These crimes may or may not require a human puppeteer depending on the content that was synthesized, they may use any type of medium or combinations, the synthesized content will often impersonate either the victim themselves or a close relation, and will target an individual human. This category of cybercrime will boom with the widespread adoption of deepfake technologies.

Scenario 2: Cat Fishing

STATUS: Currently being exploited in the wild

CONTROL: Human

MEDIUM: All (Examples include Text, Video, and Hybrid)

INTERACTIVITY: All (Asynchronous, High and Low Latency Interactivity, Real-Time)

FAMILIARITY: Unfamiliar

INTENDED TARGET: Human-Narrowcast

Romance scams, sock puppetry, and cat fishing are examples of online deception that have been, and will continue to be, significantly transformed by synthetic media. Romance scams often rely on a synthetic social media (or online dating) profile as a lure to engage (hook) a potential victim with the intention of building a trusting relationship that the criminal then exploits. Cat fishing is a closely related scam that utilizes a false persona to lure a victim into investing in a relationship with the scammer, but this persona may or may not have a romantic connotation with business scams being a common incarnation of cat fishing. Deepfake technology will alleviate a large portion of the compromise risk these scammers currently assume if a potential victim asks for a video conversation. These scams are usually controlled by humans or a hybrid team, using all forms of communications medium, and all degrees of interactivity. The scammers in these cases are usually not known to the victims in real life, but the relationship is purely technologically mediated. Although multiple humans may be engaged simultaneously, these scams usually focus on individual humans as each interaction is tailored to the victim. These scams seek high payoffs for high investment unlike 419 spam scams which send broadcast emails to millions of potential victims.

Scenario 3: Audio Deep Fake - Vishing

STATUS: Currently being exploited in the wild

CONTROL: Human

MEDIUM: Audio

INTERACTIVITY: Real-Time Interactivity

FAMILIARITY: Unfamiliar, Familiar

INTENDED TARGET: Human-Narrowcast

While only a few synthetic media attacks using audio deepfake technology have been documented, the example from a UK firm foreshadows a likely trend that will likely soon increase. In this example a human puppeteer controlled an audio syn-puppet through an audio medium, in a real-time phone conversation, using a syn-puppet that was familiar to the target as the voice of the executive, and this attack was focused on an individual human target (the individual with authorization to transfer funds).

Scenario 4: Gift Card Scams - Hybrid Attacks

STATUS: Currently being exploited in the wild

CONTROL: Hybrid

MEDIUM: Text

INTERACTIVITY: Low Latency Interactivity (Text Messaging)

FAMILIARITY: Familiar (can be unfamiliar as with the IRS impersonation variant of this scam)

INTENDED TARGET: Human-Narrowcast

Gift card scams are currently one of the most prolific examples of ROSE fraud. These scams frequently leverage human-automation teams to increase scalability, scammers currently rely on text-based communications to allow for easier impersonation, the interactions are usually low-latency texts or emails, scammers regularly impersonate familiar coworkers or a supervisor, and these scams normally target individual humans.

Scenario 5: Zishing

STATUS: Proof of Concept, No known cases in the wild

CONTROL: Human

MEDIUM: Audio-Video

INTERACTIVITY: Low-Latency Interactivity (Video Chat)

FAMILIARITY: Close, Familiar, or Unfamiliar

INTENDED TARGET: Human-Narrowcast or Broadcast

The example of Jordan Peele controlling a President Barack Obama syn-puppet was meant as a demonstration of deepfake technology capability as a warning to the public to critically evaluate what they are seeing and hearing in the information steam. Zoom bombing is a form of harassment involves an uninvited member of an online meeting imposing themselves on that meeting to disrupt it, or to spy on the conversation for later disclosure. This phenomenon that has increased substantially with the

recent push to telework. This is a likely foreshadowing of how synthetic media technology will soon be employed by those with malicious intent. Zishing (zoom phishing) is an attack that we are likely to soon observe. At the time of writing, the author is not aware of any such attacks, but the proof of concept has been demonstrated. A humorous incident, not intended as an attack, involved a lawyer participating in online court hearing. The lawyer accidentally turned on "kitten mode" in the video chat client and appeared to be a talking kitten to the rest of the participants[23]. While this example was amusing, a man-in-the-middle attack employing similar technology could be used to impersonate any of the other members. A malicious variant could have impersonated the plaintiff to fire the attorney or move to dismiss the case. While such a circumstance would be likely discovered later, this would cause considerable chaos in the meantime. Such future zishing attacks will likely involve human controlled syn-puppets, utilizing a combination of audio and video chat, may or may not be familiar to the participants, and will likely target other humans individually or broadly at scale.

Scenario 6: Synthetic Honey Traps

STATUS: Currently being exploited in the wild

CONTROL: Artificial (primarily)

MEDIUM: Text (primarily)

INTERACTIVITY: Any level

FAMILIARITY: Unfamiliar

INTENDED TARGET: Human or Automation, Narrow, or Broadcast

Thus far, this presentation has focused on the offensive employment of synthetic media technology to facilitate attacks by malicious actors; however, these technologies may also be employed counter-offensively by defenders who wish to protect their resources against malicious actors. An example of this is the Honey-Phish project by Robert Gallagher. This project sent replies to phishing emails by using a Hidden Markov Model (HMM) process to automatically generate content that was intended to entice scammers to respond to the bait. If the scammer clicked the link in the response their IP address, operating system, and browser were sent back to the Honey-Phish operator[24]. While Honey Pots and Honey Nets have been in wide usage for nearly two decades, their sophistication has increased significantly, and will continue to increase with the incorporation of synthetic media[25]. Attackers are highly motivated to detect and evade Honey Pots[26]. Generating more realistic network traffic and other system behaviors will help defenders evade detection. The Robin Sage example discussed earlier was employed as a red team technique but could just as easily be repurposed as a social media Honey Trap to catch malicious actors. The Jolly Roger bot described earlier[13] provides another example for a voice-based chat bot that counters spam callers. This defensive employment of synthetic media may be controlled by humans or automation, employ any medium, have any range of interactivity, will likely be unfamiliar to the attackers, and may target individual humans or automation. The defensive employment of synthetic media will only be limited by the imaginations of those who create and implement the technologies.

Scenario 7: Automated Attack Against Authentication Algorithms

STATUS: Unknown

CONTROL: Hybrid

MEDIUM: All (including tactile sensory inputs for fingerprint scanners)

INTERACTIVITY: Low Latency, Real-Time

FAMILIARITY: Familiar

INTENDED TARGET: Automation-Narrowcast

Synthetic media attacks are likely to usher a broad spectrum of new frauds that would have been impossible prior to the creation of these enabling technologies. A narrowcast attack against home automated assistants might impersonate the voice of a homeowner to circumvent voice authentication. More consideration will need to be given to automated detection of audio deepfake technology for voice authentication as well as facial biometric authentication. Much like the defensive employment of synthetic media, these attacks on biometric authentication will only be limited by the imaginations of the attackers. Within the context of this framework, these attacks will likely by controlled by humans and automation, involve all mediums of communication including tactile (as in 3D printed fingerprints), will interact in real-time or close to it, will impersonate a familiar user, and target a specific automated authentication mechanism.

Developing Human-Centric Countermeasures

Social engineering works by exploiting characteristics of the human operating system. When people feel rushed, impeded, or obliged to do a favor or make a concession, they are vulnerable to manipulation and exploitation by a social engineer. Synthetic media will significantly enhance exploitation opportunities. While several technology-based methods to detect synthetic media currently exist, this work focuses discussion on human-centric countermeasures to synthetic media attacks for the following reasons. First, this technology is advancing at a lightning pace, and it is conceivable that there will be periods when malicious actors achieve an anti-forensic advantage[27]. Second, non-technical users will be targeted by these attackers. To the author's knowledge, there is not a user-friendly, commercial off the shelf (COTS) synthetic media detection platform designed for consumer use that is currently available and not everyone will have access to these defensive technologies once they are developed. Finally, technology-based synthetic media attack detection methods are difficult to apply in real-time because they are forensically intensive. To detect and counter synthetic media attacks in real-time, users of information technologies need to be aware of the possibility of these attacks and know how to defend against them as a conversation is unfolding. For this reason, users need to be personally skilled in detecting and countering synthetic media attacks to ensure they can remain secure against these methods, even if technology-based countermeasures are available. Policy implementation as a countermeasure is a low-tech solution to a high-tech threat.

Policy-Based Countermeasures – Proactively implementing policies can provide defenders with tremendous advantages. Having effective policies in place that are easy to understand and easy to follow will drastically reduce both synthetic media enabled and traditional social engineering successes. Four policies that cost little to nothing to implement include the following: the Shared Secret Policy, the Never Do Policy, the Multi-Person Authorization Policy, and finally the Multi-Factor or Multi-Channel Verification Policy.

The Shared Secret Policy – This policy relies on a shared secret to quickly authenticate the person on the other end of the communication. This could be in the form of a probe question and response or a 'proof' statement. A probe question might "Which coversheet should go on this TPS report?" to which the recipient would reply with a pre-determined response such as "The two Bobs said TPS coversheets are no longer required." To maximize effectiveness the question should avoid alerting the criminal to the probe. In this example, if TPS reports are part of the interaction, this exchange would appear normal. Receiving the wrong answer should alert the employee who should in turn know to alert security and follow the organization's prescribed security policy on collecting additional information from the attacker or increase the probability of revealing the criminal's true identity. A similar technique can be used to counter virtual kidnappings. Arranging a 'proof-of-life' statement ahead of time will quickly clarify whether the kidnappers are legitimately holding a loved one for ransom. Simply asking the kidnappers for the proof-of-life statement, even if one has not been arranged beforehand, will throw them off their game if they do not have the loved one. Demanding to speak to the loved one will also allow for quick verification through shared secret knowledge. If using this technique, be sure to choose shared knowledge that is not commonly known or had been posted to social media.

The Never Do Policy - Another easy to implement policy-based countermeasure is for a high-ranking executive to state in no unclear terms what requests executives and supervisors will never make. After being hit by several gift card scams, a director of one organization sent a broadcast email to all employees stating unambiguously that he would never ask any employee to ever purchase gift cards for any reason or circumstance, full stop. This type of messaging gives employees clear guidance regarding the pretexts used in these scams. Knowing that gift cards would never be asked for, allows employees to immediately identify such a request as fraudulent and circumvents any reservation about reporting it.

The Multi-Person Authorization Policy – Requiring two people to authorized actions of a pre-specified nature or transactions that exceed a specified amount can be cumbersome and consume additional human resources. For this reason, this policy should be implemented cautiously. However, for actions that could be potentially devastating to the organization, having two sets of eyes reviewing the actions can be vital. This second review could substantially diminish the number of successful BIC frauds committed each year.

The Multi-Factor or *Multi-Channel Verification Policy* – Another highly effective policy is verifying the action through another channel. Using a second form of authentication would prevent a substantial number of BIC scams from being successful. The important point here is that the second factor or channel needs to be distinct from the primary channel being utilized. I once interviewed a victim of a BEC scam who confided to me that they had sent an email to the address they received the request from, asking if the first email was legitimate. The email account had been compromised with the

criminal having control of the inbox. The scammer assured the employee that the request was legitimate and that they should proceed with the transfer. Another method of verification is multi-factor authentication (MFA) which might employ a hardware token that generates a synchronized code which can be used to validate the other party. The greatest challenge when implementing these policies is overcoming the human proclivity toward circumventing them when circumstances dictate that their use is a hinderance because of time, convenience, or sympathy for a distressed employee "I can't verify the number right now because I dropped my fob in the toilet and don't want to tell my boss".

Behavior-Based Detection and Countermeasures – Higher degrees of interactivity imply that a human puppeteer will be needed to control the syn-puppet during an engagement with a potential victim. This opens opportunities for potential targets and defenders to detect these attacks based on puppeteer leakage of cues. This will be particularly true in cases of high familiarity where kinematic inconsistencies will likely reveal the deception. When the syn-puppet impersonates someone with whom we are highly familiar with there is a higher likelihood of detecting behavioral inconsistencies. Another potential point of weakness with highly interactive attacks is that the puppeteer will need to be a skilled actor to convey the subtle ancillary cues that are present during the pretextual emotional states, inconsistencies are likely to be detected by the target. More research is needed to determine where these weak points exist and train users on how to detect them. In the meantime, making employees aware that synthetic media technologies are being actively exploited may help in detecting these attacks.

Conclusion

Synthetic media represents one of the greatest revolutions in social engineering capability. This will lead to new challenges that the security will need to face. The Synthetic Media Attack framework that encompasses five dimensions: *Medium* (text, audio, video, or a combination), *Interactivity* (pre-recorded, high asynchrony, low asynchrony, or real-time), *Control* (human puppeteer, software, or a hybrid), *Familiarity* (unfamiliar, familiar, close), and the *Intended Target* (human or automation, an individual target, or a broader audience). I hope this will help guide security professionals and researchers during the challenges that lay ahead. While several technology-based methods to detect synthetic media such as currently exist, most technology-based solutions are not readily available to the average user and are difficult to apply in real-time. Easy to implement polices offer low tech and low-cost solutions for countering these threats. So called 'old-school' solutions will find new life in the emerging world of synthetic media attacks and security defenders will need to find ways to adapt to this new reality.

Key Takeaways

1. Online scammers are currently using deepfake technology and other synthetic media to launch social engineering synthetic media attacks against their intended victims. This technology will soon likely play a more significant role in social engineering attacks.

2. A common framework will help researchers better describe attack TTPs and defenders to anticipate future attacks. The Synthetic Media Attack Framework addresses this need and categorizes attacks along five dimensions: the medium used, the control of the syn-puppet, familiarity of the syn-puppet with the victim, interactivity of the syn-puppet, and the intended target of the attack.

3. Understanding these aspects of synthetic media attacks allows for the development of easily implemented policies that validate the legitimacy of the person on the other side of the communication and create more effective defenses against these and more traditional social engineering attacks.

Terms

Agent: An online entity under algorithmic control (aka a bot).

Avatar: An online representation of a human.

Syn-Puppet (aka **Digital Puppet**, or **Sock Puppet)**: The agent or avatar that is presented to the target.

Puppeteer: The human/algorithm controlling the puppet.

ROSE (Remote Online Social Engineering): The employment of social engineering techniques through technologically mediated communications.

Synthetic Media: The artificial production, manipulation, and modification of data and media by automated or semi-automated means.

TTPs (Tactics, Techniques, and Procedures): The methods used by cybercriminals and other threat actors to achieve their objectives.

Sources

[1] FBI. (2021). Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations. Federal Bureau of Investigations Private Industry Notification.

[2] FBI. (2021). 2020 Internet Crime Report. Federal Bureau of Investigations, Internet Crime Complaint Center (IC3).

[3] Ring, T. (2021). Security Firm: Deepfakes are Fraud's Next Frontier. Biometric Technology Today, June 2021, 2-3.

[4] https://en.wikipedia.org/wiki/Synthetic_media

[5] Newman, L. H. (2020). A Cheap 3D Printer Can Trick Smartphone Fingerprint Locks. Wired. Retrieved on July 15, 2021, from, https://www.wired.com/story/cheap-3d-printer-trick-smartphone-fingerprint-locks/

[6] Greig, J. (2020). 820% jump in e-gift card bot attacks since COVID-19 lockdowns began. Retrieved on March 21, 2021, from https://www.techrepublic.com/article/820-jump-in-e-gift-card-bot-attacks-since-covid-19-lockdowns-began/

[7] Wixey, M. (2018). Every ROSE has its thorn. Black Hat USA, Las Vegas.

[8] Derived from personal interviews with gift card scam victims.

[9] Greene, K. K., Steves, M., Theofanos, M., & Kostick, J. (2018, February). User context: an explanatory variable in phishing susceptibility. In Proceedings of 2018 Workshop Usable Security.

[10] Seymour, J., & Tully, P. (2018). Generative models for spear phishing posts on social media. arXiv preprint arXiv:1802.05196.

[11] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[12] Morales, C. (2021). Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders. Retrieved on March 17 2021 from https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html

[13] https://jollyrogertelephone.com/

[14] https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog

[15] Ryan, T., & Mauch, G. (2010, July). Getting in bed with Robin Sage. In Black Hat Conference (pp. 1-8).

[16] McCoy, T. (2014). DEA created a fake Facebook profile in this woman's name using seized pics — then impersonated her. Retrieved on April 30, 2021 from https://www.washingtonpost.com/news/morning-mix/wp/2014/10/07/dea-created-a-fake-facebook-profile-in-this-womans-name-using-seized-pics-then-impersonated-her/

[17] U.S. News (2015). U.S. Settles With Woman for $134K Over Fake Facebook Account. Retrieved on April 30, 2021 from https://www.nbcnews.com/news/us-news/u-s-settles-woman-134k-over-fake-facebook-account-n289901

[18] https://thispersondoesnotexist.com/

[19] Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017, October). Dolphinattack: Inaudible voice commands. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 103-117).

[20] Tabassi, E., Burns, K., Hadjimichael, M., Molina-Markham, A., & Sexton, J. (2019). A taxonomy and terminology of adversarial machine learning. NIST IR. Retrieved on July 19, 2021, from, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

[21] Arnoldi, J. (2016). Computer algorithms, market manipulation and the institutionalization of high frequency trading. Theory, Culture & Society, 33(1), 29-52.

[22] Domm, P. (2013). False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; AP Confirms Its Twitter Feed Was Hacked. Retrieved on July 19, 2021, from https://www.cnbc.com/id/100646197.

[23] https://www.youtube.com/watch?v=qcnnI6HD6DU

[24] Gallagher, R. (2016). Where Do the Phishers Live? Collecting Phishers' Geographic Locations from Automated Honeypots, 2016 ShmooCon, https://bitbucket.org/rgallagh/honey-phish

[25] Zhou, A. (2018). Bringing the Fight to Them.

[26] Zhang, L., & Thing, V. L. (2021). Three Decades of Deception Techniques in Active Cyber Defense-Retrospect and Outlook. arXiv preprint arXiv:2104.03594.

[27] Lyu, S. (2020, July). Deepfake detection: Current challenges and next steps. In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 1-6). IEEE.