A CISO'S GUIDE TO DEEPFAKES

DR. JENNIFER BAYUK

My first encounter with deepfakes was circa 2000. Pornographic images were circulating among the male technology staff. The images on their screens were notice by other staff who reported the activity to Human Resources. Corporate Security monitored the physical activity of the culprits. They found a kiosk-like set-up in the desktop image-build laboratory. The CIO had recently purchased a 24-disc CD duplicator for the purpose of distributing standard builds. At a recent office Christmas party, desktop administrators had taken pictures of all the women who worked in IT at the time (not many). Back in the lab, they had pasted the faces on pornographic images downloaded from the Internet and burned them onto a CD-ROM. They had then used the new duplicator to copy the CD in bulk and sold them for $10 each. The lab was behind a locked door, so the suspects were limited to the desktop admins who had physical access. Nevertheless, Corporate Security brought in Information Security to assist in gathering digital evidence. An unfortunate recent computer science graduate on the cybersecurity staff (called "Information Security" at the time) was assigned to image the machines, search for each photo on the CD-ROM, and connect it to digital evidence incriminating each desktop administrator. Back in the day, this took a few weeks.

Fictional British TV presenter Max Headroom circa 1987

Employee misconduct and deepfake fake news are the two most widely known members of the class of deepfake attacks.

Of course, these were not deepfakes in the true sense of the word because it was mostly pretty easy to tell that the faces were pasted. The incident was significant because it was a precedent for cases that were not traditionally within the domain of cybersecurity. That is, computer security, information security and their descendent cybersecurity were originally solely concerned with business confidentiality, integrity and availability issues from the perspective of operational risk management. Code of conduct cases such as this one were typically handled solely by Legal and Human Resources. However, the skill sets required to investigate this case were found internally only in the cybersecurity group. Even the CIO realized that it made sense for Corporate Security to limit the dissemination of information related to the investigation to internal cybersecurity staff subject matter experts rather than conduct research or hire vendors to find out what needed to be done to collect appropriate digital evidence. This is why, when a deepfake that negatively impacts

the organization is identified today, any required forensic analysis is likely to land in the CISO's lap. Even if the CISO has outsourced forensic activity, the oversight of the vendor has to fall in an area where the technology itself is understood well enough to agree on the investigation's deliverables.

Today's caseload of deepfake incidents goes well beyond employee misconduct. Investigating employees is cake compared to investigating the origin of malicious deepfakes from anonymous sources that appear designed to discredit public figures. Although it is becoming more common for individuals to see their own pictures tampered with on the internet, it is typically not a threat to their employers unless they are in a position of leadership and/or personally represent the company to external parties. Common deepfakes that target public figures are videos of politicians edited to create the appearance of drunkenness or stammering. Anything public figures like politicians, celebrities and business leaders do may be considered news, so I call this type of attack "deepfake news." The threat level is dependent on the identity and affiliation of the target. Note that this type of attack is not new, but happened to **Nancy Pelosi** and **Mark Zuckerburg** as far back as 2019.

If you remember **Max Headroom**, then it does not take much to envision a deepfake of a news anchor. In fact, broadcasting companies are **experimenting with clones** of their own anchors for use in "breaking news" broadcasts. A deepfake attack scenario wherein a cloned anchor delivers fake news in combination with an advanced signal hijack attack is a timebomb waiting to happen. If you remember Orson Welles's "**War of the Worlds**," déjà vu. (If you don't, it's worth the click to see.)

Deepfake of South Korean news anchor Kim Joo-Ha, 2020

Another case where public figures, including business executives, may be the target of deepfakes does not involve public video, but video presumed to be private, or "**fabricated private remarks**." For example, a deepfake targeting a business executive announcing less than expected revenue on a privately created recording mimicking an earnings call would have no widely trusted public version with which to compare. Such "private fake news" propagation could be used to alter investor sentiment and/or fabricate a market-moving event.

A variation on deepfake attack tactics related to public figures is **deepfake doxing**. Video is doctored to show a public figure committing an ethically questionable act or even a crime, then is posted online in combination with the target's current location. This happened to an anti-porn crusader. Not only did they put her face on porn, they published her location and encouraged others to rape her.

Employee misconduct and deepfake fake news are the two most widely known members of the class of deepfake attacks. The full class includes a much wider variety of potential attack tactics. Many are variants on existing attack tactics like phishing and account takeover.

For example, a typical phishing attack path looks something like Figure 1. An inbound email is faked to look like it came from a source related to a bank website to which the target may have login credentials. The target is fooled into clicking on a link to a faked site that looks like the real one and enters valid credentials. The lookalike site sends the attacker the valid credentials, displays a "login

failed" message, then redirects the user to the actual website. The attacker logs into the bank using the real credentials and transfers funds.

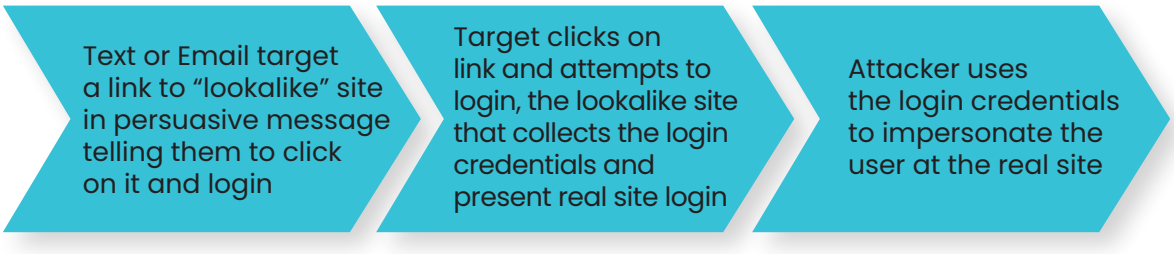

Figure 1: Typical phishing attack path

The past decade has seen countless permutations of a phishing attack path. There are now variations that employ deepfakes, such as those in Figure 2.

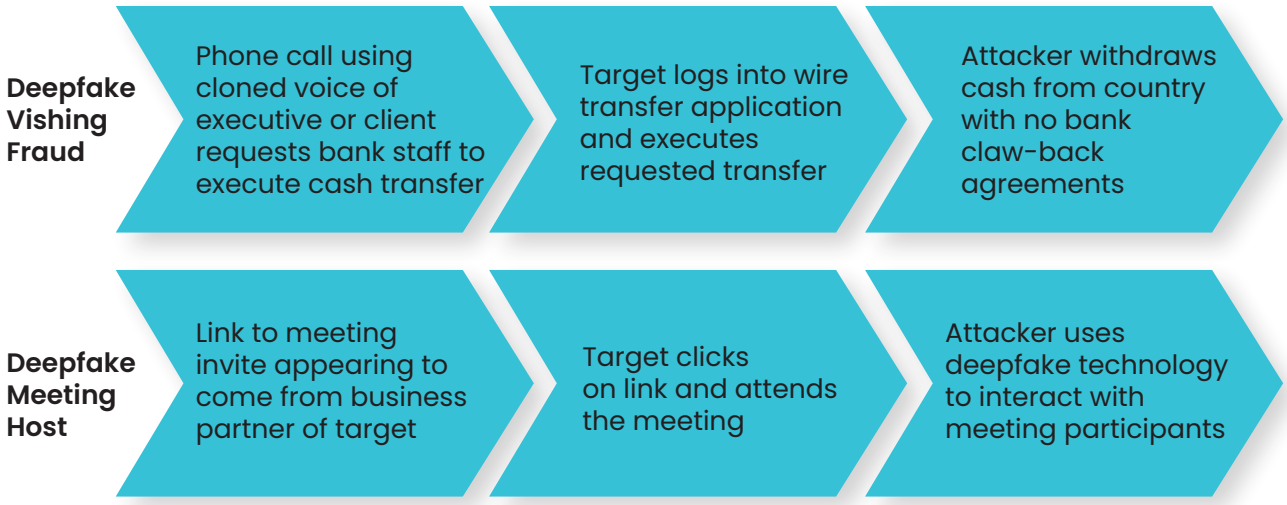

Figure 2: Deepfake phishing variations

In the first variation, deepfake vishing, a person's voice is faked rather than an email address and a website. The voice-phished target is bank staff who typically take cash transfer orders from executives or clients with verbal confirmation. The staff recognize the faked voice and so execute the corresponding instructions. The deepfake vishing fraud example in Figure 2 is based on an actual **event at Centennial Bank.**

The second variation, deepfake meeting host, uses traditional phishing tactics to target the victim and send them a fake email that looks like it came from a person of influence. The email dupes them into attending an online meeting. When they attend, they are met with a deepfake of the influential figure who puts false words in the mouth of the figure. The unfortunate "meeting host" is totally ignorant of the meeting's existence. This example happened at **Binance**.

An example of a deepfake tactic that is a variant of account takeover is to spoof biometric authentication. We have known for years that it was theoretically possible to use AI technology to **deepfake biometrics**, and in the past two years, we have seen evidence that these attacks have successfully occurred. Facial impersonation has often been tried but now meets the present definition of deepfake because attackers are using a person's cloned image to dupe account login modules. Fingerprints and voice credentials have been faked as well.

So far the deepfake tactics described have been related to a single target. In a whole other class of attacks the target is not the person cloned, but **deepfake identify theft**. Individuals have used deepfakes of qualified applicants **to apply for remote workforce positions** with access to sensitive data.

Given the wide variety of deepfake attack tactics, there is no one solution to reduce the risk of negative impact related to malicious use of deepfake technology. Nevertheless, there are some common sense remediations for the deep fake risk issues discussed above. A necessary tool in this toolkit is to train your security operations teams on techniques that are useful to quickly ascertain where an online news story originated. For example, the **SIFT method**: Stop, Investigate the source, Find better coverage, Trace the original context. Another tool is a fast-track procedure to publish an internal investigation's results using your standard press release process. If internal resources are inadequate for these purposes, consider a collaborator like **Logically** or **RealityDefender**. Figure 3 lists more specific deepfake risk issue remediation approaches for each of the attack tactics discussed above.

Of course, these efforts come with cost, so it makes sense to come up with a credible scenario for each type of deepfake attack class and run through it in a systematic manner to determine what level of preparation will be required to mitigate it. That said, you will not find deepfake separately listed as a tactic in MITRE ATT&CK. The cybersecurity profession is just starting to identify the creative ways hackers are intermingling deepfakes into their tactics. Still, prioritizing deepfake responses does not necessarily require hiring external experts. The Cybersecurity and Infrastructure Security Agency (CISA) has published comprehensive instructions on how to use tabletop exercises to analyze

Deepfake Tactic	Potential Remediation
Biometrics	Use the same antifraud techniques to biometric authentication as for eCommerce logins, e.g. block known bad actors, lock out user behavioral anomalies and multiple failed attempts, require session-level two-factor authentication. Update algorithms as required to maintain state-of-the-art fake detection.
Doxing	Provide public figures with an emergency call button for executive protection services that can prevent their arrest and/or get them out of jail promptly. Make sure executive protection staff know to call security when they suspect a deepfake.
Employee Misconduct	Adapt existing procedures for any other employee fraud. Train incident response team on techniques like SIFT.
Identity Theft	Outsource in-person verification of identity prior to hiring remote staff, or in any other situation where in-person identity verification is not practical when establishing a trusted relationship. If you do not currently verify that new hires have actually quit their old job once they have been hired, add that step to your onboarding process.
Meeting Host	Fast-track publish a disclaimer. Prioritize as you would a critical security incident. Train response team on techniques like SIFT. Prosecute where possible for future deterrence.
News Anchors	Fast-track publish a disclaimer. Contact the public relations department of the news source, and if possible, the anchor personally. Request that they issue a public statement disclaiming the fake.
Public News	Fast-track publish a disclaimer. Rapidly produce the original video and make it available for comparison. Prioritize as you would a critical security incident. Train response team on techniques like SIFT. Fast-track publish investigation results. Engage legal for cease-and-desist/defamation proceedings.
Private News	Follow public news remediation. Also enlist the assistance of an independent outsider to publicly opine on its lack of authenticity (e.g. Logically or RealityDefender).
Vishing Fraud	Require identity-based checks and balances on outbound cash transfers above a preset risk limit, regardless of the seniority of those ordering transfer.

Figure 3: Example deepfake risk issue remediations

cybersecurity risk. Although CISA does not yet have a template for a deepfake tabletop, you can roll your own with their **generic tabletop how-to resources.**

One way to prepare for deepfake attacks is to construct tabletop exercises using resources CISA offers.

With your portfolio of potentially significant deepfake risk issues and remediations in hand, identified deepfakes may be treated with standard cybersecurity incident triage. There will be some sifting of attributes to determine which falls though the sieve to be declared security incidents. Depending on the size and threat profile of an organization, there could be several incidents involving deepfakes in a year, month, week or day. Of those, some small percentage might be declared security incidents worthy of response—that is, bona fide enacted threats in contrast to some **script kiddie** spoof or comic video that would be beneath the dignity of the organization to take seriously. As with any security incident, it may be directly observed by cybersecurity staff, or referred for investigation by another internal or external source. If you do not yet have a "deepfake" indicator flag that you can count in your security metrics, best to create it now. If deepfake incidents start creeping up in your security operations trend metrics, you may want to consider getting ahead of the adversaries with threat intelligence solutions that target deepfake activity (e.g. **ActiveFence** or **Blackbird.ai**).

Although not every organization is at high risk for deepfake attacks, a recent annual **survey** of 125 cybersecurity and incident response professionals reported that deepfake attacks increased by 13 percent over 2021 and 66 percent of respondents claimed to have witnessed a deepfake attack in the past 12 months. It is just a matter of time before the cybercrime industry refines its deepfake products to increase attack efficiency and effectiveness. Forewarned is forearmed.

"Richard, I am not a deepfake. I'm a real person. And you are really fired."

(Reprinted from the TAG Cyber Security Annual, 1st Quarter 2023.)