RESEARCH NOTE

Abstract

Deepfakes are digital forgeries that use artificial intelligence to create believable but misleading images, audio, and videos. When organizational insiders interact with this manipulated media, they may inadvertently take actions that compromise their organization's security. To address this problem, we've developed a short video to raise awareness of deepfakes and provide guidance on how to counteract them. The purpose of this effort is to decrease unintentional insider threat by building greater awareness of deepfakes and how they are used to influence workforce behavior.

the ThreatLab
Integrating Research Into Practice

About The Threat Lab

The Defense Personnel and Security Research Center (PERSEREC) founded The Threat Lab in 2018 to realize the DoD Counter-Insider Threat Program Director's vision to incorporate the social and behavioral sciences into the mission space. Our team is headquartered in Seaside, California, and includes psychologists, sociologists, policy analysts, computer scientists, and other subject matter experts committed to workforce protection.

Deepfakes and Unintentional Insider Threats

OPA Report No. 2021-087 • PERSEREC-RN-21-14 • FEBRUARY 2022

Christina R. Weywadt, Kristin G. Schneider, Mark Giuffre, Sandra Ellis, and Stephanie L. Jaros

Introduction

Unintentional insider threat occurs when an insider unwittingly compromises the security of their organization through neglect, carelessness, or human error. Unintentional insider threat is generally less costly than malicious insider threat, but twice as common (IBM Security, 2020) and can make an organization more vulnerable to targeted attacks. The vulnerabilities created by unintentional insiders can be exploited by bad actors, who attempt to lure unwitting insiders into providing access to controlled information or compromising their organization's systems in other ways. One method that bad actors increasingly employ to target unintentional insiders is deepfakes. Deepfakes are a pernicious form of digital forgery in which artificial intelligence is used to create misleading images, audio, and video hoaxes.

Deepfakes use artificial intelligence (AI) to alter digital content and can be very difficult to detect. Using algorithms (i.e., computer "rules") built from digital media (e.g., digital video or voice recordings), AI can realistically mimic familiar voices, manipulate expressions, swap faces, or sync them with speech to make it seem like people said or did something they never actually said or did. Deepfakes are increasingly easy to produce and less costly to disseminate than traditional print media forgeries (Rosenbaum, 2021).

While not all AI-generated deepfakes are created for malicious purposes, bad actors use deepfakes to influence people and manipulate their behavior (Leibowicz et al., 2021; Ovadya & Whittlestong, 2019). For example, in a high-profile fraud case a British energy company was cheated out of more than two hundred thousand dollars when an employee transferred funds to a fraudulent account following a call that used deepfake technology to recreate the voice of his boss. The employee reported that the request seemed odd, but that the tone, intonation, and accent was

so realistic that they complied with the request without verifying it (Institute for the Future, 2020). In cases like this, deepfakes have been used to steal data, personally identifiable information (PII), intellectual property, and money, bringing devastating losses of integrity, intellectual property, and State secrets, as well as posing risks to infrastructure. Deepfakes can also contribute to the spread of disinformation and misinformation (e.g., COVID-19, election fraud), destabilize communities, and undermine people's confidence in the information they consume, making these forgeries a threat at every level of society.

Current DoD cyber security training already addresses some forms of digital deception, such as phishing, a practice in which bad actors use manipulated email addresses to fool recipients into communicating with an impostor. But expanding this training to educate the general workforce on the emerging threat from deepfakes could help improve organizational security and protect against insider threats (University of Maryland, Applied Research Laboratory for Intelligence and Security, 2020). The rapid advancement of deepfake technology will quickly outpace prevention efforts that encourage the use of perceptual indicators to identify deepfakes (e.g., distortions or misspellings). Raising awareness of this form of digital forgery in a broader sense will encourage people to protect themselves using critical thinking and healthy skepticism. If employees understand how bad actors can use deepfakes to manipulate and lead them into unintentional actions that harm their organization, then they can better protect themselves and their organization from this vulnerability.

The National Insider Threat Task Force (NITTF) and the Office of the Under Secretary of Defense for Intelligence & Security (OUSD[I&S]) tasked The Threat Lab, a division of the Defense Personnel and Security Research Center (PERSEREC), to raise awareness of unintentional insider threats through an informational video. This Research Note describes the process we used to identify a topic for the video that would highlight how technology can be used to manipulate insiders, how insiders can counteract this threat, and the process we used to develop the video.

Content Development and Design

We began the video development process by identifying and reviewing the existing research literature on unintentional insider threat. The research team selected emerging threats related to social engineering, with deepfakes as the primary focus. The review included a search of databases including Google, Google Scholar, and ProQuest for published literature. We reviewed identified books, reports, scholarly articles, conference papers, theses and dissertations, government publications, magazine articles, blog postings, and online materials. The review focused on unintentional insider threats and risks related to social engineering including definitions of and motivations for creating deepfakes, descriptions of different types of deepfakes, the relationship of deepfakes to unintentional insider threat, and the future of deepfakes.

Content Development

To identify the primary message of our video, we met with Subject Matter Experts (SMEs) from the Cybersecurity and Infrastructure Security Agency (CISA) to discuss emerging forms of socially engineered messaging that affect organizational security by increasing the risk of unintentional insider threats. The discussion confirmed our approach to focus the video on the future of manipulated media and countering this emerging threat. In particular, we discussed and opted to focus on the emerging threat of deepfakes, a form of digitally altered media that uses artificial

intelligence to create believable messaging engineered to encourage people to provide access to information or systems they would otherwise protect.

We developed the video content for a general and diverse population who may be familiar with altered media, but may not understand how advancements in altered media increase their risk of becoming an insider threat. We designed the video content to address five key concepts:

1) Media is not necessarily an accurate portrayal of reality; it can be enhanced, modified, and manipulated;
2) Enhancement, modification, and manipulation of media is evolving due to advancements in artificial intelligence (AI);
3) Deepfakes are an example of how AI can be used to modify media to generate visual and audio content with a high potential to deceive others;
4) Deepfake technology can be used for harmless fun or as tool for deception that can mislead people; and
5) There are ways people can protect themselves and others from being misled by deepfakes.

Look and Feel

We begin the video with an upbeat introduction to establish a personal connection with the viewer, followed by examples of digitally enhanced media that are harmless and humorous (see Figure 1). The video progresses to explain how the same methods of digital enhancement can be used for malicious and active deception. The video reminds viewers that, as the technology behind deepfakes continues to advance, identifying them via technological safeguards will be insufficient and it is our responsibility to practice vigilance and take a critical stance to prevent the spread of misinformation. The video concludes by focusing on the human solution, and how we can improve security by improving how people *think* about security.

Figure 1: Face Swap Technology Used for Harmless Fun

Overall Message and Distribution

The overall message of the video is that deepfakes threaten to become an increasingly destructive political and social force. Although technical methods to detect manipulation in photos and videos are advancing, they still fall short. Empowering the workforce to protect itself from misinformation and disinformation is essential to combat this emerging threat. Included in the video are a list of resources (See Figure 2) people can use to learn more about protecting themselves and others from deepfakes. Included in these resources are general recommendations that encourage people to be:

1) Aware of advancing technologies that make it difficult to identify when media has been enhanced or altered;
2) Skeptical and to increase their critical awareness without provoking cynicism or distrust in media generally;
3) Proactive by engaging selectively with trusted media outlets and avoiding media that does not provide details about the context or origin of the message; and
4) Mindful of how their actions and choices can contribute to the spread of misinformation and disinformation.

Figure 2: List of Resources Provided in the Video

As part of the development process, members of CISA reviewed and approved our key concepts and ensured our messaging aligned with the general principles promoted by their STOP.THINK.CONNECT© campaign.

A link to the final video (https://vimeo.com/602102971/f1f1012cf5) will be posted on the CDSE website. The link will be disseminated to insider threat program leaders and members of The Threat Lab distribution list, it could also be re-distributed as part of National Insider Threat Awareness Month. We recommend that this video be used to supplement current DoD cyber security training to help the DoD workforce keep a forward focus on the type of emerging threats that could be used to target them.

Future Research

We hope this effort will encourage future work to control the spread of misinformation and disinformation. This goal can be advanced through efforts to increase media literacy in the general public and Federal workforce, and to develop tools and techniques that can mitigate the risk posed by deepfakes and other deceptive media.

References

IBM Security. (2020). Cost of insider threats: Global report. https://www.ibm.com/security/digital-assets/services/cost-of-insider-threats/#/

Institute for the Future (2020). Moving upstream 2030. Protecting the DoD workforce against future insider threats. https://www.iftf.org/upstream2030

Lastdrager, E. E. (2014). Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, *3*(9), 1-10. https://doi.org/10.1186/s40163-014-0009-y

Leibowicz, C., McGregor, S., & Ovadya, A. (2021). The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media. *arXiv preprint arXiv:2102.06109.* https://arxiv.org/abs/2102.06109

Ovadya, A., & Whittlestone, J. (2019). Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning. arXiv preprint arXiv:2004.07213. https://arxiv.org/abs/1907.11274

Rosenbaum, S. (2019, September 23). What is synthetic media? MediaPost. https://www.mediapost.com/publications/article/341074/what-is-synthetic-media.html

University of Maryland, Applied Research Laboratory for Intelligence and Security (ARLIS). (2020). *Inaugural Report 2020*. http://www.arlis.umd.edu/index.php/inaugural-report