Seeing Isn’t Believing: Navigating the Deepfake Dilemma in the Age of AI



ARTICLE 02.18.24 GUI COZZI

“I’ll believe it when I see it.” An idiom we are all familiar with is about to be rendered obsolete thanks to recent technological innovations enabled by artificial intelligence (AI).

Deepfakes: A Growing Cyber Threat

As has been the case throughout our history, the rush to be the first to market with innovations and technologies without consideration of their long-term consequences is rearing its head again. Powerful new tools, enabled by machine learning and AI, allow those who wield them to create “deepfake” videos that can imitate anything imaginable.

“Deepfakes” are synthetic media, such as videos or audio recordings, created using AI algorithms to create, manipulate, or superimpose content onto existing images or recordings. Anyone with access to the Internet – with just a few minutes of effort and minimal skill – can create powerful videos of anything they can imagine by simply providing a short text prompt. These tools are not only available to governments and private companies – but more recently the public. When such powerful, digital cyber capabilities are released, there is never a shortage of actors ready to pounce and leverage these capabilities with nefarious intent. Due to these technologies, cyber threats have become increasingly sophisticated, posing significant risks to individuals, organizations, and even entire nations.

Focusing on the risks posed to organizations and enterprise networks – AI deepfake technology enables attackers to impersonate any individual. By creating realistic videos or audio recordings of key personnel within an organization, attackers can deceive employees into believing they are communicating with a trusted authority figure, like a CEO or manager. This can lead to fraudulent activities, such as unauthorized fund transfers or disclosing sensitive information.

A recent CNN article highlights an incredible incident in which this attack came to fruition:

> *A finance worker at a multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company’s chief financial officer in a video conference call, according to Hong Kong police. The elaborate scam saw the worker duped into attending a video call with what he thought were several other members of staff, but all of whom were in fact deepfake recreations, Hong Kong police said at a briefing on Friday. “(In the) multi-person video conference, it turns out that everyone [he saw] was fake,” senior superintendent Baron Chan Shun-Ching told the city’s public broadcaster RTHK.*
>
> *CNN*

Defending Against Deepfakes: The Human Element

So how do organizations stay ahead of the threats and prevent the dangers of deepfake technologies?

Employee awareness and training are paramount in combating social engineering attacks involving deepfakes.

- Organizations should conduct regular training sessions to raise awareness among employees about the existence and potential risks of deepfake technology, as well as how an attack may take form.
- Employees should know how to verify the authenticity of communications, especially those involving sensitive information or financial transactions.
- Organizations should establish clear communication protocols for verifying the authenticity of requests, particularly those related to financial transactions or sensitive information.
- Employees should confirm instructions received via unusual channels, such as video calls or audio messages, through established communication channels or in-person verification.
- Organizations should create out-of-band verification channels for authenticating transactions conducted over non-in-person communications channels.

Fight AI with AI

Utilize AI Detection Tools to counteract AI with AI. Organizations can leverage AI-based detection tools to identify and flag potential deepfake content. These tools utilize machine learning algorithms to analyze videos or audio recordings for signs of manipulation or inconsistencies, helping organizations identify fraudulent content before it causes harm. Availability of these tools today is limited – but organizations should look for solutions that can help supplement the protection provided by cyber-aware employees.

No longer can we believe it if we see it. Organizations need to make efforts to ensure that – when it comes to high dollar transactions or releasing sensitive information – employees use a critical eye and believe it when it is authenticated in person or through a pre-approved out-of-band channel/process.