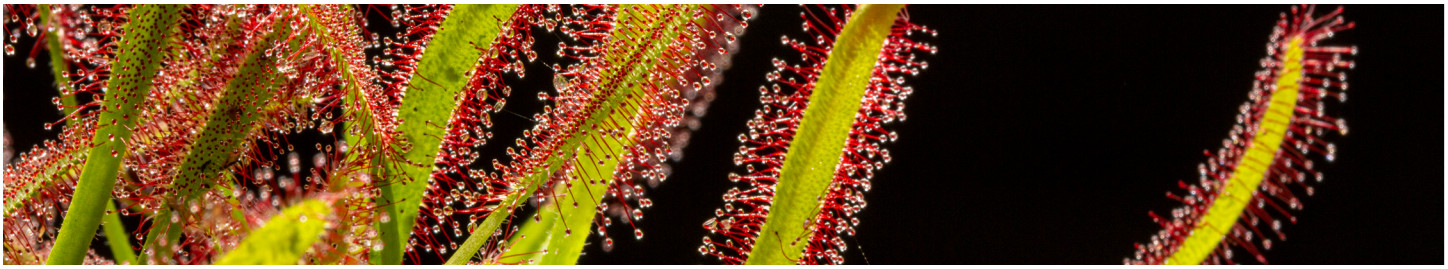

How to protect your business from deepfakes

Deepfakes are increasingly being used to imitate executives and target organizations. Learn how to shield your business from this emerging threat.

Key takeaways

- Deepfakes are videos, audio, photos and text — created using artificial intelligence — that are extremely hard to differentiate from authentic media.
- While there are legitimate uses for deepfakes, they can also put businesses at risk for exploitation and fraud, and pose a significant threat to authentication technologies.
- The best protection strategy is to tighten your business' identity verification protocols, and to incorporate deepfakes into incident response training and into your cyber awareness education.

Discerning between legitimate and inauthentic content online has never been more difficult than it is today. The spread of disinformation — verifiably false information disseminated with the intent to mislead — has affected every corner of the internet, from social media to search engine results (see *The threat misinformation and disinformation pose to business*). One of the most effective and dangerous tools of disinformation is the deepfake — a type of synthetically modified media used to impersonate real humans.

"While deepfake technology is still fairly new (first developed in 2017), deepfakes rank as one of the most dangerous AI crimes of the future."

Manipulation of digital media has blurred the lines of reality, making it easier for cyber criminals to target and deceive individuals and businesses. In the past, media such as video, audio and photos could only be modified manually by humans with editing skills and/or sophisticated software programs. Today, digital video, audio, images and even text can be created and reshaped using artificial intelligence (AI). And while there are legitimate uses for this type of synthetic media, it is more often deployed in disinformation campaigns seeking to subvert the truth, which can damage an organization's reputation or include fraudulent requests for payment.

Deepfakes are digital content created or modified using a subset of AI known as deep learning, or deep neural networks. Deep learning algorithms are generated by a mesh network of computers that sync like a human brain to produce video, audio, photos or text depicting a fake event or spoofing an identity. Many deepfakes are convincing enough to fool most viewers, and the technology has already been successfully abused by cyber criminals for financial gain. In fact, while deepfake technology is still fairly new (first developed in 2017), deepfakes have been called one of the most dangerous AI crimes of the future.[1]

To create a deepfake, you need three main elements: the input, deep learning and the output. The input is a large data set, such as dozens of podcast episodes or hours of video footage of the individual being impersonated. The more data you have to feed to the AI, the more accurate the deepfake will be. Next, deep learning is applied to analyze this input and determine patterns from the data set. Once finished, deep learning composes the output. This could be dynamically generated audio or video that's manipulated in real time, or a post-production clip that can be used any number of ways, such as a social media post.

The most common deepfakes today are artificially manipulated post-production videos that use an actor to impersonate a real person. For the input in this type of deepfake, developers need a target video of the actor to use as the foundation of the synthetic media, as well as a collection of video clips of the real person to be inserted in the target. After analyzing the clips to identify and map the person's face from different angles and in different lighting and environments, autoencoders (a form of deep learning) then map that person's likeness onto the actor in the target video by finding common features.

Some deepfakes employ generative adversarial network (GAN) technology to detect and improve flaws, which makes the deepfake more believable — and less likely to be identified by the human eye or even deepfake detectors. In addition, cloud-based and mobile applications are now available to manufacture deepfakes for those who don't have the computing power, capital and know-how to create them on their own. For now, these synthetic videos are focused on the face and are unable to map below the neck, but they're starting to evolve to catch mannerisms and behaviors.

Legitimate uses for deepfakes

Deepfakes are a dangerous tool in the disinformation arsenal. However, they're unlikely to be outlawed, as there are several legitimate applications for the technology. The biggest adopter of legitimate deepfakes so far is Hollywood, where the technology has been used to restore an actor's vocals, improve foreign-language dubbing, age down actors in flashbacks, or even complete works after an actor has died or retired. In addition, deepfakes have been used for educational purposes. In St. Petersburg, Florida, for example, the Dalí Museum has a deepfake video of the surrealist painter Salvador Dalí introducing his art and taking selfies with visitors.[2]

"Two out of three cyber security professionals saw malicious deepfakes used as part of a strike against businesses in 2022, a 13% increase from the previous year, with email as the top delivery method."

Deepfakes can also be used for legitimate corporate purposes, as a time- and cost-saving measure. For example, instead of sitting through multiple takes, a busy executive could record a video once and let deepfakes make rapid corrections. In addition, a recorded video can be translated and localized in multiple languages using deepfakes. The technology can even be used to create completely synthetic characters, replacing costly actors for professional development courses or commercials.

[1] University College London, "Deepfakes' ranked as most serious AI crime threat," August 2020.
[2] Dami Li, The Verge, "Deepfake Salvador Dalí takes selfies with museum visitors," May 2019.

Deepfake risks for companies

While deepfakes can be used for sanctioned business content, organizations must acknowledge their inherent risks. In 2021, the FBI issued a warning to businesses about deepfake fraud, saying that malicious actors "almost certainly will leverage synthetic content for cyber crime and foreign influence operations in the next 12-18 months."[3] In fact, two out of three cyber security professionals saw malicious deepfakes used as part of a strike against businesses in 2022, a 13% increase from the previous year, with email as the top delivery method.[4]

Because well-crafted deepfakes require high-end computing resources, time and technical skill, cyber criminals typically use them for operations against large enterprises and demand steep payments — but as technologies evolve to make deepfakes easier and cheaper to create, criminals will be able to target companies (or third-party vendors) of all sizes.

Business identity compromise risks

In 2020, threat actors used an audio deepfake to steal $35 million from a Hong Kong bank, the largest publicly disclosed amount lost to inauthentic content yet.[5] They pulled off the sophisticated heist using a newly defined threat vector called business identity compromise (BIC). BIC uses deepfake technology to create synthetic corporate personas or imitate existing employees, often posing as a well-known, high-ranking professional in the organization. Put simply, BIC builds trust where there shouldn't be. Once it's established, criminals can seize trade secrets and patents, impact company culture with political commentary, undermine relationships with customers and partners, tank stock values, create turmoil in the supply chain and otherwise sow chaos. Both audio and video deepfakes have already been used to impersonate executives at Fortune 500 companies, including CEOs, CFOs, treasurers and other senior leadership. In some cases, deepfakes have been deployed to humiliate or harass the executive, making it seem like they said or did something that they did not, in order to damage the reputation of the company and executive.

Deepfake phishing risks

Deepfake phishing is another emerging threat for businesses, combining disinformation (in the form of deepfakes/BIC) and phishing to fool employees into making unauthorized payments or volunteering sensitive proprietary or customer information. Often, deepfake phishing begins with an audio deepfake of a trusted figure in the organization. The criminal, disguised as the figurehead, will reach out via web conferencing or voicemail, then follow up with other forms of social engineering, such as business email compromise (BEC) or dynamic voice manipulation, using a sense of urgency to pressure employees into releasing funds or data.

Authentication risks

Finally, deepfakes pose a significant threat to authentication technologies, including facial recognition and voice recognition. Researchers at the University of Chicago found that AI-generated deepfake voices were able to fool three popular real-world voice recognition systems.[6] Similarly, studies have shown that some deepfake-generating techniques have been able to trick common web-based facial recognition APIs[7] and even certain types of technologies used to unlock smartphones,[8] though smartphones that use three-dimensional mapping as part of facial recognition are not yet vulnerable to two-dimensional deepfakes.

As deepfake technology becomes more readily available, organizations with less sophisticated security capabilities — and fewer awareness and mitigation policies around deepfakes — will be at greater risk. As mobile and cloud-based deepfake applications further penetrate the market, criminal focus may shift from executives to high-net-worth individuals or professionals with access to critical infrastructure. If threat actors have access to enough input data, anyone from a power plant manager to a social media influencer could be targeted with deepfakes.

[3] FBI, "Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations," March 2021.
[4] VMWare, "Global Incident Response Threat Report," August 2022.
[5] Thomas Brewster, *Forbes*, "Fraudsters Cloned Company Director's Voice in $35 Million Bank Heist, Police Find," October 2021.
[6] Emily Wenger *et al*., University of Chicago, "Hello, It's Me: Deep Learning-based Speech Synthesis Attacks in the Real World," September 2021.
[7] Kyle Wiggers, VentureBeat, "Study warns deepfakes can fool facial recognition," March 2021.
[8] Jessica Hallman, Penn State, "Deepfakes expose vulnerabilities in certain facial recognition technology," August 2022.

Shallowfakes

Although true deepfakes require the use of deep learning or deep neural networks, shallowfakes (also called cheapfakes) are similar in concept but use simpler methods. For instance, they can consist of media presented out of context or doctored with simple editing tools, such as filters or airbrushing.

Shallowfakes are often used in disinformation campaigns, as well as in business scams. For example, criminals have hacked into video conference calls using previously recorded video with the sound turned off, so they could impersonate an executive while feigning audio problems. They then call back by phone, pretending to be the executive and request a wire transfer. Other examples of shallowfakes include the following:

Audio/video speed — To cause reputational damage, scammers may slow down audio to make a speaker sound intoxicated or speed up video to make a person's actions look violent.

Proof of identity or address — By using simple photo editing tools to doctor photo IDs, utility bills or bank statements, for example, documents can be manipulated to falsely prove identity or location.

Documentation — Manipulated invoices, receipts or even photo evidence can be used to make fraudulent expense reports or insurance claims.

Tips to mitigate deepfake risks

Deepfake incidents have only recently started targeting businesses. Here are proactive steps organizations can take to protect their businesses from deepfakes:

- **Educate employees and partners**
  Most professionals outside of cyber security likely haven't heard of deepfakes and will benefit from learning about different formats and likely threat scenarios. Show a variety of deepfake examples — legitimate and inauthentic, video and audio — to raise awareness of their risks. Additionally, leadership should reinforce their risk expectations and clarify when senior leaders can request payments and how employees can validate those requests. Remember, employees are your first line of defense.

- **Maintain cyber security best practices**
  Cyber security best practices, especially those related to social engineering and fraud prevention, can help fortify companies against deepfake phishing and other disinformation campaigns.

  - Review foundational security policies with employees, especially how to spot relevant scams and sophisticated phishing techniques. This should include scrutinizing communications with skepticism and verifying their validity through secondary channels — especially those that ask for personal information or payments outside of usual billing structures, and that make such requests with insistence.
  - Empower your teams to "pause" and raise a concern, so you can validate or mitigate it.
  - Provide positive reinforcement when employees spot fraud and prevent losses.

- **Strengthen identity verification and validation protocols**
  This should include strengthening login credentials and authentication methods, as well as adhering to the principle of "least privilege" — and to a lesser degree "trust, but verify."

  - Least privilege asserts that users should only be given access to those accounts deemed necessary to perform their jobs. Whereas trust, but verify can be applied upon suspicion of a false identity. Even if you know the face on the screen or voice on the line, it's important to have a secondary confirmation channel.

  - If you receive a call asking for payments that exceed acceptable thresholds, you should have a clear exceptions process that requires verification — even if the request comes from the C-suite.

- **Include deepfakes in incident response planning**
  After developing deepfake awareness and reinforcing identity verification policies, businesses should incorporate deepfakes into their incident response planning. It's critical to publicly acknowledge and squash a deepfake (or shallowfake) as fast as possible — especially if it's circulating on social media. It's a race against the clock: The longer a deepfake spreads without being addressed, the more people can be convinced and the more believable it becomes. You want to tamp down disinformation quickly and succinctly in a press release and/or in messaging delivered through verified social media channels. Pre-approved communication templates can be prepared (and most importantly, approved by legal) in advance to address media, employee or vendor concerns.

Deepfake detection

As deepfakes evolve, it will become more and more difficult to distinguish them from authentic media. However, synthetic media created with rudimentary deepfake technology or by threat actors with lesser skills can still be detected by human senses. Training in deepfake detection can improve the likelihood that employees will catch BIC and deepfake phishing attempts before they cause irreparable harm. Even if personnel are unable to pinpoint a specific flaw, they may experience the "uncanny valley" phenomenon, where the slight difference between a humanlike deepfake and an actual human causes discomfort or revulsion.[9]

How to spot a deepfake

There are a few telltale signs of audio and video deepfakes.

Audio: Listen for longer-than-usual pauses between words and sentences. The person's voice may also sound flat and lifeless. If it sounds off, it likely is.

Video: Watch for long periods without blinking, patchy skin tones and poor lip syncing. Jawlines can sometimes reveal flaws, such as blurriness or flickering around the edge of the face. Ears may also appear a completely different skin color from the face.

Because deepfake threats against organizations are so new, there aren't many established technical solutions for detection or protection purposes. Those that do exist tend to be pre-profit startups that allow users to upload a suspected deepfake video, as well as photo and audio files to be analyzed for anomalies or traces of spoofing. As technology companies iterate on detection algorithms, they'll develop ever more robust models and systems to meet the need for more reliable solutions. Until then, we can only trust — but verify. ■

[9] Jeremy Hsu, *Scientific American*, "Why 'Uncanny Valley' Human Look-Alikes Put Us On Edge," April 3, 2012.

For use in external marketing and communications materials when the content of the material discusses both products and services offered through the bank and broker/dealer affiliates.

BANK OF AMERICA