DEEPFAKES: DRIVING THE PROLIFERATION OF FRAUD, DECEPTION AND DISINFORMATION AT WARP SPEED

Hardly a day goes by without new warnings around ChatGPT and other generative AI technology.[1] While the focus on these large language model technologies is relatively recent (beginning not much more than six months ago), warnings about one form of AI-augmented threats – deepfakes -- have been with us for a few years now.[2] But like large language models, the number and sophistication of deepfakes online, due to the power of generative AI tools, have increased exponentially, and so too have the risks. Very few experts writing about deepfakes two years ago, let alone six years ago, fully appreciated the speed at which, and manner in which, the threats would proliferate. This week's deepfake Twitter report of an attack on the Pentagon that briefly sent financial markets down, prompting a response from the Pentagon Force Protection Agency and Arlington County Fire Department, is a timely reminder (among many) of the threat and concomitant need to respond quickly.

What are Deepfakes?

The term "deepfakes" covers a range of digitized textual, audio and visual content created through the use of AI. Deepfakes appear authentic, typically featuring people saying or doing things that they have never said or done, and would never say or do. While early deepfakes tended to be visual and pornographic, nearly all depicting women (prompting an increasing number of states to provide legal recourse for victims), since 2019, audio deepfakes have become far more ubiquitous, largely because convincing voice deepfakes can be created with relatively small audio samples of a target's speech and are harder to identify as fake.[3] Deepfakes are created with varying degrees of sophistication and quality, ranging from so-called "shallowfakes"/"cheapfakes" to more labor-intensive deepfakes developed with neural network technology.

---

1 *See* my previous briefing notes, available here: AI and Election Disinformation, Regulating GAI and A Primer on GAI.

2 The first reports of deepfakes date back to 2017. (*See* "Increasing Threat of Deepfake Identities," published by DHS). A report by University College London published in August 2020 ranked deepfakes as the most serious AI crime threat.

3 Speaking recently, Tristan Harris and Aza Raskin co-founders of the Center for Humane Technology in a presentation entitled The A.I. Dilemma, demonstrated how just three seconds of captured audio can allow malign actors to "voice-clone," and complete indistinguishable sentences through generated speech. They cite the example of a teenager answering a phone call (it is "a wrong number" and the caller hangs up); sometime later the caller calls back and the synthetic voice of the "teenager" asks the parent to be reminded of his/her social security number for a job application. That social security number has just been stolen. Or, the "teenager" calls his parents to report he/she is in jail and needs cash for bail. The Washington Post reports this scam actually happened, repeatedly.

The FTC report on 2022 scams affecting consumers that was cited in the Washington Post article is available here. CNN recently reported on (and provides useful tips for thwarting) scam kidnappings and ransom demands based on an all-too realistic cloned voice. Voices can be cloned from social media accounts as well.

The relationship between generative AI and deepfakes is based on the ability of the former to be used to build more realistic versions of cloned people, better able to mimic how that person would sound uttering malicious content. The technology also facilitates the creation of synthetic people that are able to deliver messages at scale in multiple languages, at minimal cost.

Consequences of Deepfakes

Worryingly, the extent to which deepfakes can be deployed is limited only by one's imagination. As a consequence, we soon could find ourselves questioning everything, not knowing what is real and what is fake, which in turn means that businesses, firms and institutions will need to implement fundamental changes to the way they operate. Eurasia Group, in its ranking of the top risks of 2023, placed generative AI as #3 on its list, noting that "advances in deepfakes, facial recognition, and voice synthesis software will render control over one's likeness a relic of the past." It notes that, with no barriers to entry to create malicious content, the volume of content increases exponentially, rendering it impossible for anyone to "reliably distinguish fact from fiction."

An essay published in the Wall Street Journal in February, "The Deepfake Dangers Ahead," describes not only criminal risks but threats to national security. It cites, for example, Jihadi extremists mobilizing recruits through deepfake denigrations of Islam and deployed troops despairing after reading divisive or provocative social media posts ostensibly by fellow soldiers, but in fact deepfakes. Imagine Japan's wartime Zero Hour radio programs, on steroids.

The Department of Homeland Security provides additional examples of threat scenarios:

- Incitement of unrest and violence: a malign actor produces a deepfake interview of a police chief endorsing mistreatment of citizens.
- Production of false evidence about climate change: malign actors produce deepfake images showing ice accumulation in Antarctica rather than melting.
- A fake kidnapping: families are targeted through deepfake videos of staged kidnappings of family members to obtain ransom without the "kidnappers" having to physically kidnap the purported victim; deepfakes make it easier to show horrific treatment of victims to pressure the families.
- Production of false evidence in criminal cases: deepfakes could be used to make up alibis.
- Corporate sabotage: addressed in detail in this note.
- Cyberbullying: malign actors can seek to blackmail families by threatening to disseminate deepfake videos of purported criminal behavior of a family member or non-consensual deepfake pornography.
- Election influence.
- Child predator threats.

The Wall Street Journal essay catalogues a range of likely responses, by way of example:

- The military will need to roll out highly secure systems for verifying orders and making sure that automated systems cannot be overridden by deepfakes.

- Political leaders responding to crises will have to build in delays to ensure that information before them is not fake or even partially manipulated by an adversary.
- Media will have to become far more sceptical of breaking news stories. Where there is doubt, an outlet might need to regularly post "information not verified" warnings.

Considerations for Businesses/Firms/Institutions

It is no overstatement that disinformation has found its way into every corner of the online world, and therefore of society. And, just as the business community needs to confront the scourge of disinformation, so too must business enterprises be thinking in terms of the following four categories of deepfake threats. Similarly, professional service firms, financial service firms and non-governmental institutions are not immune to at least two of these four threats, and thus have similar exposures. Government institutions have their own sets of issues.

- ***Reputational risk***: arising from malign uses of audio or visual deepfakes to damage brands or individuals. Deepfakes could be used, for example, to falsely humiliate a senior executive or other employee to damage the reputation of the executive and, by extension, the company. The threat could come from a disgruntled employee or a malign competitor. Deepfakes could be used to fabricate injury from products to obtain settlements – in effect a form of blackmail.

  Investigative or other journalists may be victims of deepfake harassment by or on behalf of targets of their investigations/reporting in order to silence them (malign actors will have less need to resort to SLAPPs).
- ***Stock price manipulation***: disinformation purportedly from senior executives or others at a company or from a respected third-party commentator (including a stock analyst) could trigger a sell-off, or sharp rise in the price, of the company's securities. A malign actor may have shorted the stock ahead of a deepfake triggering a sell-off.
- ***Social engineering fraud/business identity compromise***: in 2021, the FBI warned via a Private Industry Notification of the ease with which deepfakes (what it then termed "synthetic content") could create highly believable synthetic spearphising messages. It called the threat "business identity compromise" ("BIC"), a potentially more malicious threat than the more traditional "business email compromise" ("BEC"). BIC involves "the use of content generation and manipulation tools to develop synthetic corporate personas or to create a sophisticated emulation of an existing employee." Sophisticated attacks may involve a combination of traditional phishing (*i.e.,* BEC) and BIC, with impersonations of trusted senior corporate officers or key personnel in treasury functions intended to cause sensitive proprietary or confidential information to be released or unauthorized payments to be made, augmented by doctored emails.[4]

  Much attention, rightly, has been paid over the past decade to the risks of traditional cybersecurity threats, particularly the risks of unauthorized access to computer systems to fraudulently access information. BIC adds a new layer to this, as it

[4] The FBI's Internet Crime Complaint Center (IC3), in a report posted last May that BEC scams alone between June 2016 and December 2021 led to domestic and international losses of $43 billion. This period does not cover BIC-driven scams.

essentially co-opts internal staff unwittingly to release sought-after information, obviating the need to deploy malware or gain access to login credentials. The Securities and Exchange Commission has issued guidance, proposed disclosure rules in March 2022 for public companies and risk management in March for market participants, and has brought enforcement actions in respect of cybersecurity policies and procedures. It remains to be seen how this regime, which focuses on "occurrences on or conducted through … information systems," might be expanded to cover deepfakes, which pose the same threats to registrants and market participants.

Last June, the FBI's Internet Crime Compliance Center warned that deepfakes were being deployed by malign actors during job interviews to obtain remote work and work-from-home positions, particularly for IT positions. Impersonations were aided by stolen personally identifiable information.

- ***Authentication risk***: use of malicious technology to fool facial recognition (visual deepfakes) or voice recognition (audio deepfakes) security verification systems.

How to Mitigate the Threats

Experts recommend a range of solutions to mitigate the threat:

- Conduct **cybersecurity risk assessments**: easily available technology allows deepfakes to be created using minimal source material, whether online conference presentations, interviews with the media or others, or online earnings calls, as well as any other video or audio footage that can be captured digitally. Moreover, as anyone with any digital footprint could be at risk, internal risk assessments should consider whose voice or images could put the business/institution more easily at risk.
- **Educate** board members, senior executives and employees to raise awareness of the threats and update cybersecurity best practice protocols to cover deepfakes. Education has an additional salutary effect, namely increasing media literacy as a defense against ongoing attempts (by malign foreign or domestic actors) outside the business context to erode trust in institutions and democracy or for fraudulent purposes.
- Update **cyber-security awareness training modules** to test all members of the organization/institution on deepfake threats and thereby enhance the likelihood of deepfake detection.
- Borrow from the two-factor identification protocols now widely in use and establish parallel **internal validation/verification protocols** for at-risk activities such as online payment authorizations and release of sensitive proprietary or confidential information. Mandate "pause" protocols, as malign actors often rely on pressure tactics and create a sense of urgency to facilitate fraud. Note that people are more likely to be caught off guard if the deepfake interaction combines voice and visual (*e.g.*, on video calls on desk phones).
- Incorporate deepfake attacks in **incident-response exercises**. Deepfake attacks need to be responded to as a matter of urgency, which may involve public disclosure. Pre-approved communication templates should be prepared and vetted from legal, regulatory and investor relations/communications perspectives. Post-attack protocols

should cover who gets notified and when, whether law enforcement, regulators, insurers, auditors, customers, shareholders, workforce or other stakeholders.

- **Deploy detection software** to identify deepfakes.
- **Check insurance coverage.**

Spotting Deepfakes

Telltale signs that video is fake (from DHS, MIT, Norton and other sources) – starting with the face, as deepfake imagines almost always involve facial transformations:

- Little or no blinking, or too frequent blinking – essentially, unnatural eye movements (people's eyes tend to follow the people they are speaking to).
- Blurring evident in the face but not elsewhere in the image or video (or vice-versa).
- A perfectly symmetrical face.
- A change of skin tone near the edge of the face; the skin of the forehead or cheeks may seem incongruent with the hair or eyes, or the neck.
- Double chins, double eyebrows or double edges to the face.
- Unnatural lip movements (or lip movements that are out of synch with the voice).
- The face gets blurry when it is partially obscured by a hand or another object.
- Box-like shapes and cropped effects around the mouth, eyes and neck.
- Lack of detail in the mouth, especially teeth that do not look real.
- Lower-quality sections throughout the same video.
- Because so much attention is paid to constructing facial features, unnatural body shapes or awkward positioning of the body.
- Movements (*e.g.,* choppy or disconnected from frame to frame) that are not natural; deepfakes tend to avoid side views, as fake side-to-front transitions are easier to spot.
- Changes in the background and/or lighting.
- Too much/too little glare reflected on glasses.
- Poor audio quality, as malign actors tend to focus more on the visual than the audio.
- The background is in consistent with the foreground and subject.

Telltale signs that audio is fake (from DHS, among other sources):

- Choppy sentences/unnatural speech cadence.
- Varying tone inflection in speech.
- Phrasing – would the speaker say it that way?
- The message is out-of-context.
- Background sounds are inconsistent with the speaker's intent.

Concluding Thoughts

Just as there is growing realization of the need, and industry support, for regulatory responses to generative AI, so too there is ample justification for regulation addressing deepfakes. But regulation takes time. In the meantime, businesses, firms and institutions need to be

proactive in preparing for what is likely an inevitable general trend and, potentially, a directed threat. Sadly, while some of the more doomsday-like ramifications of generative AI may be speculative, the threats of deepfakes are here now, and they are real. And, incidentally, just as there are untold benefits of generative AI, so too are there significant legitimate use cases for deepfakes. And just as malign actors are seeking to game the system, the same advances in technology should up the game of detection systems.

Mitigation efforts will have benefits in terms of protecting the entity as well as making directors, senior executives and staff far more discerning when it comes to the digital media they, and the rest of us, all consume. Presumably, all of the cybersecurity best practices and warnings provided at the workplace over the years have translated into greater digital safety at home; so too should defense against deepfakes.

* * *

Mark S. Bergman
7Pillars Global Insights, LLC
Washington, D.C.
May 31, 2023