GAO Science, Technology Assessment, and Analytics

SCIENCE & TECH SPOTLIGHT:

COMBATING DEEPFAKES

GAO-24-107292, March, 2024

Accessible Version

WHY THIS MATTERS

Malicious use of deepfakes could erode trust in elections, spread disinformation, undermine national security, and empower harassers.

KEY TAKEAWAYS

- Current deepfake detection technologies have limited effectiveness in real-world scenarios.
- Watermarking and other authentication technologies may slow the spread of disinformation but present challenges.
- Identifying deepfakes is not by itself sufficient to prevent abuses. It may not stop the spread of disinformation, even after the media is identified as a deepfake.

THE TECHNOLOGY

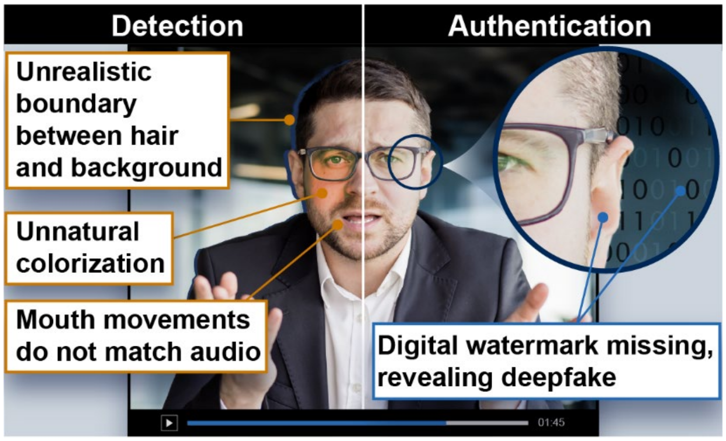
What is it?

Deepfakes are videos, audio, or images that have been manipulated using artificial intelligence (AI), often to create, replace, or alter faces or synthesize speech. They can seem authentic to the human eye and ear. They have been maliciously used, for example, to try to influence elections and to create non-consensual pornography. To combat such abuses, technologies can be used to detect deepfakes or enable authentication of genuine media.

Detection technologies aim to identify fake media without needing to compare it to the original, unaltered media. These technologies typically use a form of AI known as machine learning. The models are trained on data from known real and fake media. Methods include looking for (1) facial or vocal inconsistencies, (2) evidence of the deepfake generation process, or (3) color abnormalities.

Authentication technologies are designed to be embedded during the creation of a piece of media. These technologies aim to either prove authenticity or prove that a specific original piece of media has been altered. They include:

- **Digital watermarks** can be embedded in a piece of media, which can help detect subsequent deepfakes. One form of watermarking adds pixel or audio patterns that are detectable by a computer but are imperceptible to humans. The patterns disappear in any areas that are modified, enabling the owner to prove that the media is an altered version of the original. Another form of watermarking adds features that cause any deepfake made using the media to look or sound unrealistic.
- **Metadata**—which describe the characteristics of data in a piece of media—can be embedded in a way that is cryptographically secure. Missing or incomplete metadata may indicate that a piece of media has been altered.
- **Blockchain.** Uploading media and metadata to a public blockchain creates a relatively secure version that cannot be altered without the change being obvious to other users. Anyone could then compare a file and its metadata to the blockchain version to prove or disprove authenticity.


Detection
Authentication
Unrealistic boundary between hair and background
Unnatural colorization
Mouth movements do not match audio
Digital watermark missing, revealing deepfake
01:45

Sources: GAO analysis (data); Tetiana/sanchesnet1/stock.adobe.com (images). | GAO-24-107292

Figure 1. Examples of Deepfake Detection and Authentication

How mature is it?

Detection technologies. According to recent studies, existing detection methods and models may not accurately identify deepfakes in real-world scenarios. For example, accuracy may be reduced if lighting conditions, facial expressions, or video or audio quality are different from the data used to train the detection model, or if the deepfake was created using a different method than that used in the training data. Further, future advances in deepfake generation are expected to eliminate hallmarks of current deepfakes, such as abnormal eye blinking.

Authentication technologies. These technologies are not new, but their use in combating deepfakes is an emerging area. Several companies offer authentication services, including using digital watermarks, metadata, and blockchain technologies. Some claim to let website visitors authenticate media found on the internet, provided the original is in the company's database. Prominent social media companies are also beginning to label AI-generated content.

OPPORTUNITIES

- **Combined defenses.** Using multiple detection and authentication methods may help to identify deepfakes.
- **Updated training datasets.** Including diverse and recent media in training data could help detection models keep up with the latest deepfake generation techniques.
- **Competitions.** Deepfake detection competitions could encourage the development of more accurate detection tools and models. One 2019 competition included over 2,000 participants and generated over 35,000 models.

CHALLENGES

- **Disinformation and public trust.** Disinformation can spread from the moment a deepfake is viewed, even if it is identified as fraudulent. Further, trust in real media may be undermined by false claims that real media is a deepfake or if people do not trust a detection model's results.
- **Adaptation to detection.** Techniques and models used to identify deepfakes tend to lead developers to create more sophisticated deepfake generation techniques.

POLICY CONTEXT AND QUESTIONS

- Are current laws and regulations adequate to address evolving concerns about the malicious use of deepfakes? How do they address data security, privacy concerns, and First Amendment considerations, such as a deepfake creator's freedom of speech and expression?
- What entities (e.g., government, nonprofit, private company) should make decisions about identifying and blocking deepfakes, or about when and how to sanction those who produce or disseminate them?
- How can organizations across society coordinate on the development and improvement of deepfake detection and authentication technologies? What standards could be used or developed to evaluate these technologies?

SELECTED GAO WORK

Science & Tech Spotlight: Deepfakes, GAO-20-379SP

Technology Assessment: Blockchain, GAO-22-104625

SELECTED REFERENCES

Gourav Gupta, Kiran Raja, Manish Gupta, Tony Jan, Scott Thompson Whiteside, and Mukesh Prasad, "A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods," *Electronics*, vol. 13, no. 95 (2024) https://doi.org/10.3390/electronics13010095.

National Security Agency, Federal Bureau of Investigation, and Cybersecurity and Infrastructure Security Agency, *Contextualizing Deepfake Threats to Organizations*, Sept. 2023.

GAO SUPPORT:

The Government Accountability Office (GAO) meets congressional information needs in several ways, including by providing oversight, insight, and foresight on science and technology issues. GAO staff are available to brief on completed bodies of work or specific reports and answer follow-up questions. GAO also provides targeted assistance on specific science and technology topics to support congressional oversight activities and provide advice on legislative proposals.

For more information, contact: Brian Bothwell at (202) 512-6888 or bothwellb@gao.gov

Congressional Relations: A. Nicole Clowers, Managing Dir., (202) 512-4400, clowersa@gao.gov

Public Affairs: Charles Young, Managing Dir., (202) 512-4800, youngc1@gao.gov

This document is not an audit product and is subject to revision based on continued advances in science and technology. It contains information prepared by GAO to provide technical insight to legislative bodies or other external organizations. This document has been reviewed by Sterling Thomas, PhD, the Chief Scientist of the U.S. Government Accountability Office.

Staff Acknowledgments: Laura Holliday (Assistant Director), Douglas G. Hunker (Analyst-in-Charge), Shannon Brooks, Nirmal Chaudhary, Caroline Gross, Anika McMillon, and Benjamin Shouse.

Source: Debalina/stock.adobe.com (header). | GAO-24-107292