Contextualizing Deepfake Threats to Organizations

Executive summary

Threats from synthetic media, such as deepfakes, present a growing challenge for all users of modern technology and communications, including National Security Systems (NSS), the Department of Defense (DoD), the Defense Industrial Base (DIB), and national critical infrastructure owners and operators. As with many technologies, synthetic media techniques can be used for both positive and malicious purposes. While there are limited indications of significant use of synthetic media techniques by malicious state-sponsored actors, the increasing availability and efficiency of synthetic media techniques available to less capable malicious cyber actors indicate these types of techniques will likely increase in frequency and sophistication.

Deepfakes are AI-generated, highly realistic synthetic media that can be abused to:

- Threaten an organization's brand
- Impersonate leaders and financial officers
- Enable access to networks, communications, and sensitive information

Synthetic media threats broadly exist across technologies associated with the use of text, video, audio, and images which are used for a variety of purposes online and in conjunction with communications of all types. Deepfakes are a particularly concerning type of synthetic media that utilizes artificial intelligence/machine learning (AI/ML) to create believable and highly realistic media. [1] The most substantial threats from the abuse of synthetic media include techniques that threaten an organization's brand, impersonate leaders and financial officers, and use fraudulent communications to enable access to an organization's networks, communications, and sensitive information.

Organizations can take a variety of steps to identify, defend against, and respond to deepfake threats. They should consider implementing a number of technologies to detect deepfakes and determine media provenance, including real-time verification capabilities, passive detection techniques, and protection of high priority officers and their communications. [2] [3] Organizations can also take steps to minimize the impact

of malicious deepfake techniques, including information sharing, planning for and rehearsing responses to exploitation attempts, and personnel training.

In particular, phishing using deepfakes will be an even harder challenge than it is today, and organizations should proactively prepare to identify and counter it. Several public and private consortiums also offer opportunities for organizations to get involved in building resilience to deepfake threats, including the Coalition for Content Provenance and Authenticity and Project Origin. [4] [5]

This cybersecurity information sheet, authored by the National Security Agency (NSA), the Federal Bureau of Investigation (FBI), and the Cybersecurity and Infrastructure Security Agency (CISA), provides an overview of synthetic media threats, techniques, and trends. It also offers recommendations for security professionals focused on protecting organizations from these evolving threats through advice on defensive and mitigation strategies.

Synthetic media threats

Tools and techniques that can be used to manipulate authentic multimedia have been around for decades [6]; however, the scale in use of media manipulation has dramatically increased as the complexity of leveraging manipulated media has fallen. Making a sophisticated fake with specialized software previously could take a professional days to weeks to construct, but now, these fakes can be produced in a fraction of the time with limited or no technical expertise. This is largely due to advances in computational power and deep learning, which make it not only easier to create fake multimedia, but also less expensive to mass produce. In addition, the market is now flooded with free, easily accessible tools (some powered by deep learning algorithms) that make the creation or manipulation of multimedia essentially plug-and-play. As a result, these publicly available techniques have increased in value and become widely available tools for adversaries of all types, enabling fraud and disinformation to exploit targeted individuals and organizations. The democratization of these tools has made the list of top risks for 2023. [7]

Apart from the obvious implications for misinformation and propaganda during times of conflict, national security challenges associated with deepfakes manifest in threats to the U.S. Government, NSS, the DIB, critical infrastructure organizations, and others. Organizations and their employees may be vulnerable to deepfake tradecraft and techniques which may include fake online accounts used in social engineering attempts, fraudulent text and voice messages used to avoid technical defenses, faked videos used to spread disinformation, and other techniques. Many organizations are attractive targets for advanced actors and criminals interested in executive impersonation, financial fraud, and illegitimate access to internal communications and operations.

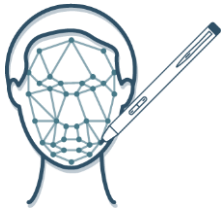
Deepfake social engineering includes:

- Fraudulent texts
- Fraudulent voice messages
- Faked videos

Defining the problem

Several terms are used to describe media that have been synthetically generated and/or manipulated. Some of the most common include:

- **Shallow/Cheap Fakes:**
  Multimedia that has been manipulated using techniques that do not involve machine/deep learning, which in many cases can still be as effective as the more technically sophisticated techniques, are often referred to as shallow or cheap fakes. These fakes are often generated through the manipulation of an original message conveyed in some real media. Some explicit examples of this include:



  - Selectively copying and pasting content from an original scene to remove an object in an image and thereby change the story. There are many examples of this in history, including when Josef Stalin removed an individual from an image after they became enemies. [8]
  - The slowing down of a video by adding repeat frames to make it sound like an individual is intoxicated. [9]
  - Combining audio clips from a different source and replacing the audio on a video to change the story. [10]

- Using false text to push a narrative and cause financial loss and other impacts. [11]

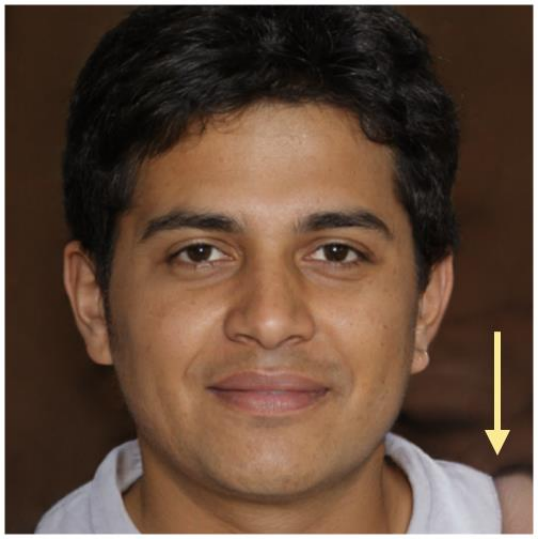
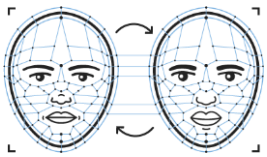


Figure 1: Deepfake image, courtesy of https://farid.berkeley.edu/misc/deepfakefaces.html [12]

- **Deepfakes:**
  Multimedia that have either been created (fully synthetic) or edited (partially synthetic) using some form of machine/deep learning (artificial intelligence) are referred to as deepfakes. Some explicit examples include:



  - LinkedIn experienced a huge increase in deepfake images for profile pictures in 2022. [13]
  - An AI-generated scene that is the product of AI hallucination—made up information that may seem plausible but is not true—that depicts an explosion near the Pentagon was shared around the internet in May 2023, causing general confusion and turmoil on the stock market. [14]
  - A deepfake video showed Ukrainian President Volodomyr Zelenskyy telling his country to surrender to Russia. [15] More recently, several Russian TV channels and radio stations were hacked and a purported deepfake video of Russian President Vladimir Putin was aired claiming he was enacting martial law due to Ukrainians invading Russia. [16]

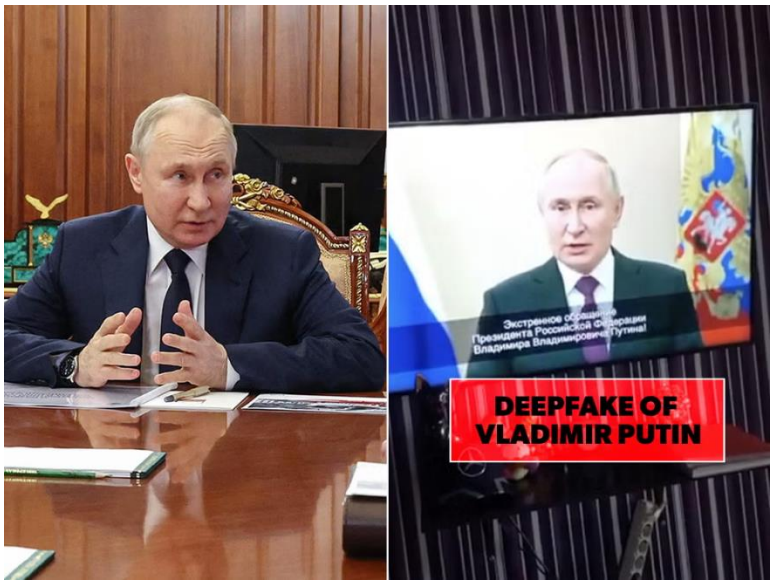
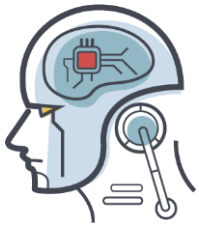


Figure 2: Deepfake of Vladimir Putin

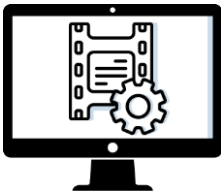
- Another example of technology being developed in video is Text-to-Video Diffusion Models, which are fully synthetic videos developed by AI. [17]
- In 2019, deepfake audio was used to steal $243,000 from a UK Company [18] and, more recently, there has been a massive increase in personalized AI scams given the release of sophisticated and highly trained AI voice cloning models. [19] [20]
- Openly accessible Large Language Models (LLMs), are now being used now to generate the text for phishing emails. [21]

- **Generative AI:**



As of 2023, Generative AI is gaining popularity for many capabilities that produce synthetic media. Generative AI (Machine Learning Techniques), such as Generative Adversarial Networks, [22] Diffusion Models, [23] and Large Language Models [24] (or a combination thereof) are the machinery that is enabling the production of highly realistic synthetic multimedia based on much larger datasets.

- **Computer Generated Imagery (CGI):**



CGI is the use of computer graphics to create or improve visual media (image and video). Traditionally, these methods have been the go-to visual effect for most major movies, but now that

Generative AI techniques are getting better and cheaper, the merging of these two technologies is being used to produce even more convincing fakes. [25]

Detection vs authentication

For several years, public and private organizations have expressed concern over manipulated multimedia and developed means to detect and identify countermeasures. Many public and private partnerships have since emerged, focusing on cooperative efforts to detect these manipulations and verify/authenticate multimedia. There are many differences between the detection and authentication efforts, as they have different goals. The greatest of these is that detection methods are often passive forensic techniques, whereas authentication methods are active forensic techniques that are purposely embedded at the time of capture or time of edit of the media in question. [26]

Detection efforts typically focus on developing methods that seek evidence of manipulation and present that evidence in the form of a numerical output or a visualization to alert an analyst that the media needs further analysis. These methods are developed under the assumption that modifications to original media or completely synthetic media contain statistically significant traces that can be found. This form of detection is a cat and mouse game; as detection methods are developed and made public, there is often a quick response from the generation community to counter them. However, until there is a universal adoption of authentication standards, these methods are necessary to support forensic analysis.

Authentication methods are designed to be embedded at the time of capture/creation or time of edit to bring transparency to the provenance of the media. Some examples include digital watermarking [27] which can be used in synthetically generated media, active signals in real-time capture to verify liveness, [28] and cryptographic asset hashing on a device. [29]

Some of the ongoing efforts in detection and authentication include several public/private collaborative initiatives, such as:

- The DARPA Semantic Forensics program is currently developing advanced semantic capabilities for media forensics and authentication. Program

participants include NVIDIA, PAR Government Systems, SRI International, and several research institutions. [30]

- The Air Force Research Lab (AFRL) recently awarded a contract to the small business, DeepMedia, for the development of deepfake detection capabilities. [31]
- Deepfake detection tools have been fielded by several companies, including Microsoft, Intel, and Google.
    - Prior to the 2020 elections, Microsoft introduced the Microsoft Video Authenticator and in 2023 they rolled out the “About this Image” tool to get more context for the authenticity of images they may receive. [32]
    - Intel introduced a real-time deepfake detector in late 2022 labeled FakeCatcher which detects fake videos. [33]
    - Google, in collaboration with academic researchers in Europe, contributed a large dataset of visual deepfakes to the FaceForensics Benchmark in 2019. [34] [35]
- Adobe launched the Content Authenticity Initiative (CAI) in 2019 to push for provenance of digital content. CAI has several hundred members seeking to develop open content attribution standards. [36] CAI developed the Coalition for Content Providence and Authenticity (C2PA). “C2PA unifies the efforts of the Adobe-led Content Authenticity Initiative (CAI) which focuses on systems to provide context and history for digital media, and Project Origin, a Microsoft- and BBC-led initiative that tackles disinformation in the digital news ecosystem.” [4]

How deepfakes can impact organizations

Public concern around synthetic media includes its use with disinformation operations, designed to influence the public and spread false information about political, social, military or economic issues to cause confusion, unrest, and uncertainty. However, synthetic media threats that organizations most often encounter include activities that may undermine the brand, financial position, security, or integrity of the organization itself. The most significant synthetic media threats to the DoD, NSS, the DIB, and critical infrastructure organizations, based on potential impact and risk, include, but are not limited to:

- **Executive impersonation for brand manipulation:** Malicious actors may use deepfakes, employing manipulated audio and video, to try to impersonate an organization's executive officers and other high-ranking personnel. Malicious actors may employ convincing audio and video impersonations of key leaders to damage the reputation and value of an organization's brand, such as quickly distributing a convincing deepfake publicly across social media platforms before it can be stopped or disputed. Types of manipulated media operations have been observed targeting high profile political figures such as Ukraine's Volodymr Zelenskyy to spread disinformation and confusion. This technique can have high impact, especially for international brands where stock prices and overall reputation may be vulnerable to disinformation campaigns. Considering the high impact, this type of deepfake is a significant concern for many CEOs and government leaders.
- **Impersonation for financial gain:** Malicious actors, many of them likely cyber criminals, often use multiple types of manipulated media in social engineering campaigns for financial gain. These may include impersonating key leaders or financial officers and operating over various mediums using manipulated audio, video, or text, to illegitimately authorize the disbursement of funds to accounts belonging to the malicious actor. Business Email Compromise (BEC) schemes are counted among these types of social engineering, costing companies hundreds of millions of dollars in losses. Similar types of techniques may also be used to manipulate the trade or sale of crypto assets. These types of schemes are far more common in practice and several partners reported being targeted by these types of operations.
- **Impersonation to gain access:** Malicious actors may use the same types of manipulated media techniques and technologies to gain access to an organization's personnel, operations, and information. These techniques may include the use of manipulated media during job interviews, especially for remote jobs. In 2022, malicious actors reportedly employed synthetic audio and video during online interviews, although the content was often unaligned or unsynchronized, indicating the fraudulent nature of the calls. These malicious attempts were enabled by stolen personal identifiable information (PII). [37] Successful compromises may enable actors to obtain sensitive financial, proprietary, or internal security information. Manipulated media techniques used

to impersonate specific customers may also be employed to gain access to individual customer accounts for account access or other information gathering purposes.

The following are examples of synthetic media attempts used to target organizations. In many cases these appear to be executed by cyber criminals intending to defraud the organization for financial gain. Beyond examples of faked personal profiles on social networking sites and deceptive smishing or vishing—phishing using SMS texts or phone calls—attempts, there is limited evidence that state-sponsored actors are specifically using sophisticated deepfakes so far.

- In May 2023, an unknown malicious actor targeted a company using synthetic visual and audio media techniques to impersonate the CEO of the company. A product line manager in the company was contacted over WhatsApp and invited to an interactive call with a sender claiming to be the CEO of the company. The voice sounded like the CEO and the image and background used likely matched an existing image from several years before and the home background belonging to the CEO.
- In May 2023, an unknown malicious actor targeted a company for financial gain using a combination of synthetic audio, video, and text messages. The actor, impersonating the voice of a company executive, reached a member of the company using a poor quality audio call over WhatsApp. The actor then suggested a Teams meeting and the screen appeared to show the executive in their office. The connection was very poor, so the actor recommended switching to text and proceeded to urge the target to wire them money. The target became very suspicious and terminated the communication at this point. The same executive has also been impersonated via text message on other occasions by likely financial criminals.

Emerging trends in deepfakes and generative AI

Dynamic trends in technology development associated with the creation of synthetic media will continue to drive down the cost and technical barriers in using this technology for malicious purposes. By 2030 the generative AI market is expected to exceed $100 billion, growing at a rate of more than an average of 35 percent per year. [38] However, while capabilities available to malicious actors will dramatically increase, technologies and techniques available to defenders seeking to identify and mitigate deepfakes will also improve substantially. One example announced in late 2022, GPTZero, is a program designed to identify computer generated text, including ChatGPT, Google's LaMDa, and other AI models, that has already garnered more than a million users. [39] However, detectors of AI-generated content can suffer from false positives where they identify human-written content as AI-generated as well. A technological escalation is expected in synthetic media technologies and capabilities to detect AI generated content and authenticate legitimate content. [40]

Major trends in the generation of media include the increased use and improvement of multimodal models, such as the merging of LLMs and diffusion models; the improved ability to lift a 2D image to 3D to enable the realistic generation of video based on a single image; faster and tunable methods for real time modified video generation; and models that require less input data to customize results, such as synthetic audio that captures the characteristics of an individual with just a few seconds of reference data. [41] All of these trends point to better, faster, and cheaper ways to generate fake content.

The major trends on detection and authentication are toward education, detection refining, and increased pressure from the community to employ authentication techniques for media. Eventually, these trends may lead to policies that will require certain changes. For now, efforts like the public/private detection and authentication initiatives referenced in this report and ethical considerations prior to releasing models [42] will help organizations take proactive steps toward more transparent content provenance. [27]

Recommendations for resisting deepfakes

Organizations can take a variety of steps to prepare to identify, defend against, and respond to deepfake threats. Synthetic media experts from several U.S. Government

agencies collaborated between 2021-2022 to consider these threats and establish a recommended list of best practices to employ in preparation for and response to deepfakes. [2] [3] Another good reference for best practices and understanding what policies can be enacted from a company perspective can be found on the Columbia Law Blue Sky blog. [43] Several of these recommendations are described below and explored in further detail.

1. **Select and implement technologies to detect deepfakes and demonstrate media provenance:**

   - **Real-time verification capabilities and procedures:** Organizations should implement identity verification capable of operating during real-time communications. Identity verification for real-time communications will now require testing for liveness given the rapid improvements in generative-AI and real-time rendering. Mandatory multi-factor authentication (MFA), using a unique or one-time generated password or PIN, known personal details, or biometrics, can ensure those entering sensitive communication channels or activities are able to prove their identity. These verification steps are especially important when considering procedures for the execution of financial transactions.

     - Companies that offer liveness tests powered by virtual injection techniques include ID R&D [44], Facetec [45], IProov [46], and many more.

     - The Center for Identification, Technology Research (CITeR) is a research initiative, funded in part by the National Science Foundation and other partners from the academic, commercial, and government sectors, that conducts research on techniques to achieve these goals. [47]

     - Passive detection of deepfakes: Passive detection techniques should be used when trying to determine the authenticity of previously created media. In these cases, recommendations for forensic analysis are as follows:

     Basic recommendations:

     - **Make a copy** of the media prior to any analysis.

     - **Hash both the original and the copy** to verify an exact copy.

- **Check the source** (i.e., is the organization or person reputable) of the media before drawing conclusions.
- **Reverse image searches**, like TinEye, [48] Google Image Search, [49] and Bing Visual Search, [50] can be extremely useful if the media is a composition of images.
- **Visual/audio examination** – look and listen to the media first as there may be obvious signs of manipulation
  - Look for physical properties that would not be possible, such as feet not touching the ground.
  - Look for presence of audio filters, such as noise added for obfuscation.
  - Look for inconsistencies.
- **Metadata examination** tools can sometimes provide additional insights depending on the situation.
  - All metadata intact is an indication of authenticity.
  - Some metadata stripped indicates the media was potentially manipulated, but further investigation is required.
  - All metadata stripped may indicate the media was obtained through a social media platform or other process that automatically strips the information.

Advanced recommendations:

- **Physics based examinations** – complete checks to verify vanishing points, reflections, shadows, and more using ideas from Hany Farid [see Chapter 1 of Fake Photos for more information] [51] and other methods that use Fluid Dynamics. [52]
- **Compression based examination** – Use tools designed to look for compression artifacts, knowing that lossy compression in media will inherently destroy lots of forensic artifacts.

- **Content based examinations (when appropriate)** – Use tools designed to look for specific manipulations when suspected. For example:
  - Use tools like those available on GitHub [53] if a GAN was suspected for deepfake production.
  - Consider plug-ins to detect suspected fake profile pictures. [54]
  - Explore the Antispoofing Wiki with various deepfake detection tools and software. [55]
  - Use open source algorithms and papers for various manipulation tasks, such as grip-unina [56] and the deepfake detection challenge. [57]
  - In addition to the techniques and categories mentioned above, other techniques can be deployed to detect deepfakes of high priority individuals. Such techniques are based off the unique characteristics of the individual and are sometimes referred to as Person of Interest (POI) models. Training these models for a particular person can be time consuming and, in some cases, requires hours of data. However, if the concern is to protect a particular individual, these methods are designed just for that. Some examples include:
    - ID-Reveal [58] and Audio-Visual Person-of-Interest DeepFake detection; [59]
    - Protecting World Leaders Against Deepfakes [60] and Protecting President Zelenskky; [61] and
    - Person Specific Audio Deepfake Detection. [62]
    - Note on POI models: if organizations wish to protect their executives with POI models, they should consider actively collecting and curating legitimate video and audio recordings of these individuals. Such collections of data will be necessary to develop detection models.

2. Protect public data of high-priority individuals.

To protect media that contains the individual from being used or repurposed for disinformation, one should consider beginning to use active authentication techniques such as watermarks and/or CAI standards. This is a good preventative measure to

protect media and make it more difficult for an adversary to claim that a fake media asset portraying the individual in these controlled situations is real. Prepare for and take advantage of opportunities to minimize the impact of deepfakes.

- **Plan and rehearse:** Ensure plans are in place among organizational security teams to respond to a variety of deepfake techniques. These should be prioritized by the likelihood and unique vulnerabilities of each organization and their industry. Some organizations will be more susceptible to executive impersonation or misinformation which may impact brand status or public stock shares. Other organizations relying on high volumes of virtual financial transactions may be more vulnerable to financial fraud.

  Once a plan is established, do several tabletop exercises to practice and analyze the execution of the plan. These should involve the most likely targets of deepfakes and include executives who may be prime targets. [63]

- **Reporting and sharing experiences:** Report the details of malicious deepfakes with appropriate U.S. Government partners, including the NSA Cybersecurity Collaboration Center for Department of Defense and Defense Industrial Base Organizations and the FBI (including local offices or CyWatch@fbi.gov), to spread awareness of trending malicious techniques and campaigns.

- **Training personnel:** Every organization should incorporate an overview of deepfake techniques into their training program. This should include an overview of potential uses of deepfakes designed to cause reputational damage, executive targeting and BEC attempts for financial gain, and manipulated media used to undermine hiring or operational meetings for malicious purposes. Employees should be familiar with standard procedures for responding to suspected manipulated media and understand the mechanisms for reporting this activity within their organization.

  Training resources specific to deepfakes are already available from the following sources:

  - SANS Institute – “Learn a New Survival Skill: Spotting Deepfakes;” [64]
  - MIT Media Lab – “Detect DeepFakes: How to counteract information created by AI” [65] and MIT Media Literacy; [66] and

  - Microsoft – “Spot the Deepfake.” [67]

- **Leveraging cross-industry partnerships:** C2PA is a significant effort launched in 2021 to address the prevalence of misleading information online through the development of technical standards for certifying the provenance of media content. Specifications and principles for ensuring media provenance can be found on the C2PA website. [4] Additional information on issues relating to misinformation and content provenance is available from C2PA associated efforts at CAI [36] and Project Origin. [5]

  As of 2023, CAI encompassed more than 1,000 private companies across tech, media, news publishers, researchers, and NGOs. CAI offers several free open source tools to implement media provenance, a regular newsletter, and a community channel on Discord.

  Project Origin, a cross industry effort involving Microsoft and several major media producers, aims to similarly establish a chain of content provenance through secure signatures and web browser extensions. Technical background can be found on their website at originproject.info.

- **Understand what private companies are doing to preserve the provenance of online content:** Organizations should actively pursue partnerships with media, social media, career networking, and similar companies in order to learn more about how these companies are preserving the provenance of online content. This is especially important considering how they may be working to identify and mitigate the harms of synthetic content, which may be used as a means to exploit organizations and their employees.

Works cited

[1] NSA, The Next Wave 2021 Vol. 23 No. 1: Deepfakes: Is a Picture Worth a Thousand Lies?, https://media.defense.gov/2021/Jul/06/2002756456/-1/-1/0/TNW_23-1.PDF
[2] DHS Public-Private Analytic Exchange Program, Increasing Threat of Deepfake Identities, https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf
[3] DHS Public-Private Analytic Exchange Program, Increasing Threats of Deepfake Identities – Phase 2: Mitigation Measures, https://www.dhs.gov/sites/default/files/2022-10/AEP%20DeepFake%20PHASE2%20FINAL%20corrected20221006.pdf
[4] The Coalition for Content Provenance and Authenticity (C2PA), https://c2pa.org/
[5] Project Origin: Protecting Trusted Media, https://www.originproject.info/
[6] Farid H., Digital Image Forensics, https://farid.berkeley.edu/downloads/tutorials/digitalimageforensics.pdf
[7] Eurasia Group, Top Risks 2023, https://www.eurasiagroup.net/issues/top-risks-2023
[8] 11 Points, 11 Famous Doctored Photos of Dictators, https://11points.com/11-famous-doctored-photos-dictators/
[9] Slate, Beware the Cheapfakes, https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html
[10] VERIFY, Video of First Lady at Eagles game manipulated to include audible boos, anti-Biden chants, https://www.verifythis.com/article/news/verify/national-verify/video-of-first-lady-at-eagles-game-manipulated-to-include-audible-boos-anti-biden-chants/536-c11989d4-e842-4a6b-bc18-9ed7c494dd84
[11] The Hustle, One fake Tweet may have cost Twitter a lot, https://thehustle.co/11152022-eli-lilly/
[12] Farid H., Deep-fake faces, https://farid.berkeley.edu/misc/deepfakefaces.html
[13] NPR, That smiling LinkedIn profile face might be a computer-generated fake, https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles
[14] CNN, 'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion, https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html
[15] Salon, Deepfake videos are so convincing — and so easy to make — that they pose a political threat, https://www.salon.com/2023/04/15/deepfake-videos-are-so-convincing--and-so-easy-to-make--that-they-pose-a-political/
[16] The Independent, Deepfake Putin declares martial law and cries: 'Russia is under attack', https://www.independent.co.uk/news/world/europe/deepfake-putin-martial-law-state-media-b2353005.html
[17] NVIDIA, Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, https://research.nvidia.com/labs/toronto-ai/VideoLDM/
[18] Trend Micro, Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company, https://www.trendmicro.com/vinfo/mx/security/news/cyber-attacks/unusual-ceo-fraud-via-deepfake-audio-steals-us-243-000-from-u-k-company
[19] TECHCiRCLE, Nearly half of Indian internet users faced AI-driven voice scams this year, https://www.techcircle.in/2023/05/02/nearly-half-of-indian-internet-users-faced-ai-driven-voice-scams-this-year
[20] The New York Times, Voice Deepfakes Are Coming for Your Bank Balance, https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html
[21] Schneier on Security, LLMs and Phishing, https://www.schneier.com/blog/archives/2023/04/llms-and-phishing.html
[22] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, Generative Adversarial Nets, https://arxiv.org/abs/1406.2661
[23] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah, Diffusion Models in Vision: A Survey, https://arxiv.org/abs/2209.04747

[24] Anjana Samindra Perera, Large Language Models, https://levelup.gitconnected.com/best-papers-on-large-language-models-ac01b13b94b3
[25] Couch Investor, The Impact of Generative AI on CGI Production in the Movie and Graphic Design Industries, https://couchinvestor.substack.com/p/the-impact-of-generative-ai-on-cgi
[26] Valentina Conotter, Active and Passive Multimedia Forensics, https://farid.berkeley.edu/downloads/publications/vcthesis11.pdf
[27] CNBC, Google will label fake images created with its A.I., https://www.cnbc.com/2023/05/10/google-will-label-fake-images-created-with-its-ai-.html
[28] Candice R. Gerstner and Hany Farid, Detecting Real-Time Deep-Fake Videos Using Active Illumination, https://farid.berkeley.edu/downloads/publications/cvpr22a.pdf
[29] Content Authenticity Initiative, How It Works, https://contentauthenticity.org/how-it-works
[30] Defense Advanced Research Projects Agency, Semantic Forensics (SemaFor), https://www.darpa.mil/program/semantic-forensics
[31] DeepMedia, DeepMedia to Help AFRL Spot Deep Fakes, https://www.deepmedia.ai/press/deepmedia-to-help-afrl-spot-deep-fakes
[32] Google, Get helpful context with About this image, https://blog.google/products/search/about-this-image-google-search/
[33] Intel, Intel Introduces Real-Time Deepfake Detector, https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.zllvh5
[34] Technical University of Munich, FaceForensics++: Learning to Detect Manipulated Facial Images, https://github.com/ondyari/FaceForensics/
[35] Google Research, Contributing Data to Deepfake Detection Research, https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html
[36] Content Authenticity Initiative, https://contentauthenticity.org/
[37] FBI, Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions, https://www.ic3.gov/Media/Y2022/PSA220628
[38] The New York Times, Another Side of the A.I. Boom: Detecting What A.I. Makes, https://www.nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html
[39] Forbes, With Seed Funding Secured, AI Detection Tool GPTZero Launches New Browser Plugin, https://www.forbes.com/sites/rashishrivastava/2023/05/09/with-seed-funding-secured-ai-detection-tool-gptzero-launches-new-browser-plugin/
[40] IEEE Spectrum, Detection Stays One Step Ahead of Deepfakes—for Now, https://spectrum.ieee.org/deepfake
[41] Microsoft Research, VALL-E (X): A neural codec language model for speech synthesis, https://www.microsoft.com/en-us/research/project/vall-e-x/
[42] MetaStellar, Adobe, Nvidia announce ethical AI image generation, https://www.metastellar.com/nonfiction/news/adobe-nvidia-announce-ethical-ai-image-generation/
[43] Columbia Law School, The CLS Blue Sky Blog, https://clsbluesky.law.columbia.edu/
[44] ID R&D, IDLive Face Plus Injection Attack Detection Helps Prevent Deepfake Fraud, https://www.idrnd.ai/idlive-face-plus-injection-attack-detection-deepfake-protection/
[45] FaceTec, https://www.facetec.com/
[46] iProov, https://www.iproov.com/
[47] Clarkson University, Center for Identification Technology Research (CITeR), https://citer.clarkson.edu/
[48] TinEye, https://tineye.com/
[49] Google Image Search (search by image), https://images.google.com/
[50] Microsoft Bing (search using an image), https://www.bing.com/
[51] Hany Farid, Fake Photos (2019), ISBN: 9780262537490
[52] The Conversation, Deepfake audio has a tell – researchers use fluid dynamics to spot artificial imposter voices, https://theconversation.com/deepfake-audio-has-a-tell-researchers-use-fluid-dynamics-to-spot-artificial-imposter-voices-189104
[53] NVIDIA Labs, StyleGAN3 Synthetic Image Detection, https://github.com/NVlabs/stylegan3-detector

[54] V7, Fake Profile Detector (Deepfake, GAN), https://chrome.google.com/webstore/detail/fake-profile-detector-dee/jbpcgcnnhmjmajjkgdaogpgefbnokpcc
[55] Antispoofing Wiki, Deepfake Detection Software: Types and Practical Application, https://antispoofing.org/deepfake-detection-software-types-and-practical-application/
[56] Grip-unina, https://github.com/grip-unina
[57] The unofficial deepfake-detection-challenge repo, https://github.com/drbh/deepfake-detection-challenge/
[58] Grip-unina, ID-Reveal: Identity-aware Deepfake Video Detection, https://github.com/grip-unina/id-reveal
[59] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva, Audio-Visual Person-of-Interest DeepFake Detection, https://arxiv.org/pdf/2204.03083.pdf
[60] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, Protecting World Leaders Against Deep Fakes, https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf
[61] Matyáš Bohácek and Hany Farid, Protecting President Zelenskyy Against Deep Fakes, https://arxiv.org/pdf/2206.12043.pdf
[62] U.S. Department of Energy Office of Scientific and Technical Information, Speaker-targeted Synthetic Speech Detection, https://www.osti.gov/biblio/1844063
[63] Debevoise & Plimpton, The Value of AI Incident Response Plans and Tabletop Exercises, https://www.debevoisedatablog.com/2022/04/27/the-value-of-airps-and-ai-tabletops/
[64] SANS, Learn a New Survival Skill: Spotting Deepfakes, https://www.sans.org/newsletters/ouch/learn-a-new-survival-skill-spotting-deepfakes/
[65] MIT Media Lab, Detect DeepFakes: How to counteract misinformation created by AI, https://www.media.mit.edu/projects/detect-fakes/overview/
[66] The MIT Center for Advanced Virtuality, Media Literacy in the Age of Deepfakes, https://deepfakes.virtuality.mit.edu/
[67] Center for an Informed Public at UW, Spot the Deepfake: Spotting deepfakes isn't as easy as you might think, https://www.spotdeepfakes.org/

Disclaimer of endorsement

The information and opinions contained in this document are provided "as is" and without any warranties or guarantees. Reference herein to any specific commercial entity, product, process, or service by trade name, trademark, manufacturer, or otherwise does not constitute or imply its endorsement, recommendation, or favoring by the United States Government, and this guidance shall not be used for advertising or product endorsement purposes.

Purpose

This document was developed in furtherance of the authoring organizations' cybersecurity missions, including their responsibilities to identify and disseminate cyber threats to National Security Systems, Department of Defense, Defense Industrial Base, and critical infrastructure information systems, and to develop and issue cybersecurity specifications and mitigations. This information may be shared broadly to reach all appropriate stakeholders.

Contact

Cybersecurity Report Feedback: CybersecurityReports@nsa.gov

CISA's 24/7 Operations Center to report incidents and anomalous activity: Report@cisa.gov or (888) 282-0870

General Cybersecurity Inquiries or Customer Requests: Cybersecurity_Requests@nsa.gov

Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov

Media Inquiries / Press Desk:

- NSA Media Relations: 443-634-0721, MediaRelations@nsa.gov
- CISA Media Relations: 703-235-2010, CISAMedia@cisa.dhs.gov