

► **BRIEF**

SYNTHETIC MEDIA: A PRIMER

ZEROFOX INTELLIGENCE

B-2023-03-06a

Classification: TLP:CLEAR
March 6, 2023



Synthetic Media: A Primer

Executive Summary

Synthetic media is the production, manipulation, and modification of artificial data by automated means for the purpose of misleading those who consume it. While synthetic media is often referred to as “deepfakes,” that is a common misnomer and a misapplication of the term. Deepfakes is just one of six distinct segments of synthetic media, alongside image synthesis, audio synthesis, speech synthesis, natural language generation (NLG), and Artificial Intelligence (AI) art. This product will provide an overview of each of these segments, context and current threats, and recommendations for countering the threat presented by synthetic media.

This product is **not** intended to provide a complete understanding of any of these segments, nor does it address the entire threat landscape associated with synthetic media. Succeeding products focused on each of the six segments of synthetic media will represent a comprehensive ZeroFox Intelligence resource on this subject.

Key Findings

- The availability and sophistication of synthetic media capabilities have increased significantly in recent years due to advancements in AI and machine learning (ML).
- While technologies associated with synthetic media have many positive applications, each technology has the potential to be used maliciously—and threat actors have already seen the results.
- In recent years, tools and services dedicated to deepfakes proliferated to the point of commoditization across the open, deep, and dark web (DDW).
- Widespread use of synthetic media within the cyber threat landscape remains low, but ZeroFox Intelligence assesses the threat is likely growing based on ease of access and use of supporting technologies combined with an observed increase in utilization by cybercriminals and nation-state actors over the last 12 months.

Overview

Synthetic media—also known as generative media—refers to digital content created or modified by AI and ML. For the purposes of this brief and in accordance with our research, ZeroFox Intelligence has segmented synthetic media into six categories:

Deepfakes	Image Synthesis	Audio Synthesis
Speech Synthesis	Natural Language Generation (NLG)	Artificial Intelligence (AI) Art

Each of these categories represents great technological leaps forward with a multitude of valid and benign applications that can greatly improve people's lives and productivity. Concurrently, each of these synthetic media segments presents opportunities for use by malicious actors intent on improving their efficiency and effectiveness to meet their goals—financial or otherwise.

Deepfakes

ZeroFox Intelligence defines deepfakes as synthetic audio and visual content that is produced to depict a person saying or doing something that they have not and can be undertaken with either malicious or non-malicious intent. This practice is often attributed to the “r/deepfakes” Reddit forum that appeared in late 2017 and was dedicated to face-swapping female celebrities into adult content videos. Deepfakes typically fall into four main categories:

- **Face-swap:** Replaces the face of a subject by stitching it onto the body of someone else
- **Face re-enactment (aka puppeteering):** Manipulates the facial features of a subject
- **Face generation:** Uses Generative Adversarial Network (GAN)*-generated images to create entirely synthetic yet convincing individuals
- **Speech/voice synthesis:** Creates a model of a subject's voice that can read text out loud in the same manner, intonation, and cadence as the subject

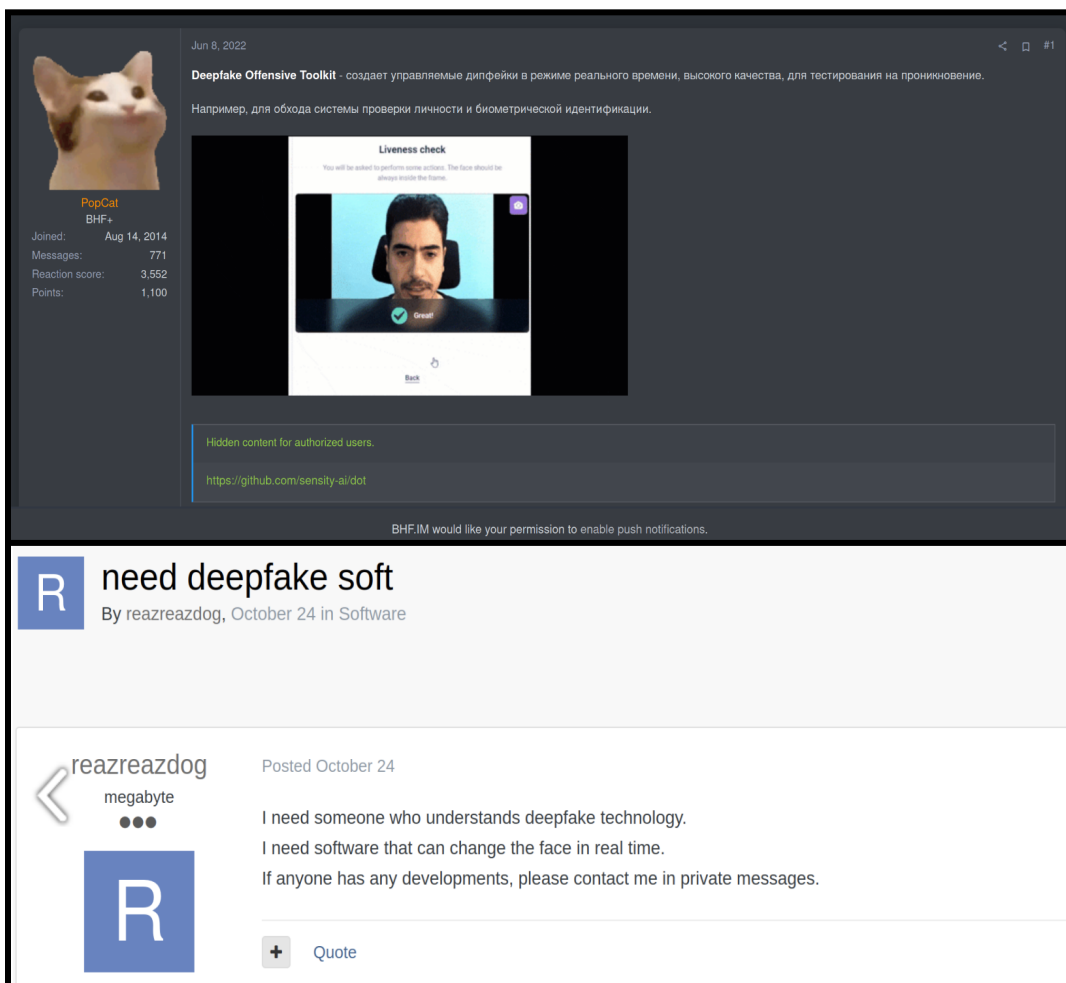
** A Generative Adversarial Net (GAN) consists of two neural networks, a generator, and a discriminator; the generator tries to produce realistic samples that fool the discriminator, while the discriminator tries to distinguish real samples from generated ones.¹*

Deepfakes can be created from a wide set of tools that vary in quality and complexity. Simple tools—including mobile apps like ReFace or Zao—provide a low barrier to entry when compared to sophisticated options that require greater expertise, time, and

¹ [hXXps://arxiv.org/pdf/1803.04469\[.\].pdf](https://arxiv.org/pdf/1803.04469[.].pdf)

technology to generate convincing deepfakes. The February 2018 takedown of the previously-referenced r/deepfakes Reddit forum likely increased the threat posed by deepfakes, as the tools have since proliferated and become commoditized and many “as-a-service” models have been discovered. Researchers and hobbyists also often share projects and ready-made software on open-source code repositories like GitHub, contributing to the availability of this powerful technique.

Deepfake communities, some of which include tutorials, are active on the Discord, Reddit, Telegram, and YouTube platforms. ZeroFox Intelligence also continues to observe deepfake chatter in closed forums and DDW marketplaces, with users often seeking deepfake services to help bypass security measures. To date, ZeroFox has not witnessed any references to specialized tools beyond those available in the open source. This speaks to the availability and utility of those free options, which contribute to the growing threat posed by deepfakes due to the sustained low barrier to entry.



Deepfake Chatter on DDW Forums

Source: ZeroFox Intelligence

| Image Synthesis

Image synthesis is the process of artificially generating visual media (pictures and video) through technology.² A subset of image synthesis—and the area of particular concern in terms of mis/dis/malinformation (MDM) and cyberattacks—is human image synthesis that can produce realistic renditions of moving or still images using an amalgamation of images of real people. This technology, widely used in Hollywood since the first photorealistic human likenesses were included in the 2003 films *The Matrix: Reloaded* and *The Matrix: Revolutions*, has become increasingly dangerous as it becomes more widely available—and as people increasingly rely on social media as a trusted source for information.³



Samples of AI-Generated Faces

Source: *thispersondoesnotexist[.]com*

Threat actors can create unique and realistic profiles on social media platforms, such as Facebook and LinkedIn, that evade technologies designed to match a profile picture with known existing images to alert against impersonations. This ability to cheaply and rapidly create an unlimited number of fictional people—with photorealistic quality that avoids this detection—increases the volume and effectiveness of malicious campaigns, including, but not limited to, romance scams and fraudulent employment opportunities.

| Audio Synthesis

Audio synthesis refers to AI waveform manipulation that results in the creation (or recreation) of any conceivable sound. There are valid and useful applications for this technology, including the creation of music, videos, and movies that benefit from being able to manufacture the perfect sound—or series of sounds—to fit a creative need. Unfortunately, this same technology can be used to create sounds to increase trust in false messaging, such as a false 911 call reporting an active shooter with the benefit of that perfect sound in the background. This technology can also be used to create sounds that can be attributed falsely to fictional events or things. Some examples include

² [hXXps://www.sciencedirect\[.\]com/topics/engineering/image-synthesis](https://www.sciencedirect[.]com/topics/engineering/image-synthesis)

³

[hXXps://www.pewresearch\[.\]org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/](https://www.pewresearch[.]org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/)

creating audio files of events that never happened (e.g., gunfire, domestic abuse, or car accidents) or sounds that “prove” the existence of things not yet proven to exist (e.g., aliens or Bigfoot) for the purpose of advancing conspiracy theories.

NOTE: While related to Speech Synthesis, Audio Synthesis is specific to the creation of sounds that are not human voices or speech.

| Speech Synthesis

Speech synthesis, often included as part of the deepfake efforts covered above, refers to the use of software or hardware system programs to artificially produce human speech.⁴ This technology usually involves concatenating pieces of recorded speech that are stored in a database. According to work produced by McGill University, “for specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely synthetic voice output.”⁵

The creation of synthetic voices has many valid and useful applications, including text-to-speech capabilities that make it safer to text while driving and empower people with vision impairment or reading disabilities to participate in text-based communications. Unfortunately, nefarious applications of this technology—commonly referred to as deepfake voices—are a significant and growing concern. These deepfake voices, also known as audio deepfakes, are most commonly created through either synthetic-based or imitation-based deepfake processes, as shown in the diagrams below.

When used against selected targets, these realistic—but entirely fake—interactions contribute to vishing campaigns designed to further credit card fraud, corporate or personal account takeovers, and extortion, among other scams. Well-designed and targeted use of audio deepfakes has successfully impersonated executives on at least two occasions, resulting in over USD 35M in losses.^{6,7}

⁴ Tan, Xu; Qin, Tao; Soong, Frank; Liu, Tie-Yan (2021-07-23). "A Survey on Neural Speech Synthesis".

⁵ Rubin, P.; Baer, T.; Mermelstein, P. (1981). "An articulatory synthesizer for perceptual research". *Journal of the Acoustical Society of America*. 70 (2): 321–328.

⁶ [https://cisomag\[.\]com/deepfake-voice-attack/](https://cisomag[.]com/deepfake-voice-attack/)

⁷ [https://www.darkreading\[.\]com/attacks-breaches/deepfake-audio-scores-35-million-in-corporate-heist](https://www.darkreading[.]com/attacks-breaches/deepfake-audio-scores-35-million-in-corporate-heist)

| Natural Language Generation (NLG)

NLG, also commonly referred to as synthetic text, is the term for computer processing of data resulting in the creation of new text that appears to be human-generated. NLG is another technology with many valid and useful applications, including language translation, text summarization, and data augmentation in natural language processing (NLP) tasks. Some companies are currently using a combination of NLG and NLP to summarize current events and create daily or weekly content (e.g., newsletters) without human labor.

Unfortunately, as with nearly any new capability, there are multiple ways to use these technological advances maliciously. Bad actors can use NLG to automatically author phishing emails spanning a wide variety of topics or automate the spread of MDM, including state-sponsored propaganda.^{8,9} Each of these examples is a force multiplier that empowers small groups or individuals to have impact at a scale previously reserved for larger, well-funded enterprises.

The latest advancement in NLG is the OpenAI project known as ChatGPT. Launched in 2019 and recently updated for public use with the release of a conversational language model in November 2022, [ChatGPT](#) can provide text ranging from source code to short stories while communicating in a lifelike dialogue.¹⁰ Recognizing the power of this technology, in January 2023, Microsoft invested USD 10 billion to take a significant stake in ChatGPT and almost immediately integrated the technology into its proprietary search engine, Bing.^{11,12} Following that, both Google and Meta made waves in the NLG space by announcing their own technologies, [Bard](#) and [LLaMA](#), respectively.^{13,14} While all of these technologies are still in their infancy, investment by global technology leaders strongly suggests this technology, and its associated threats, will advance rapidly.

⁸ [hXXps://www.wired\[.\]com/story/ai-phishing-emails/](#)

⁹ [hXXps://cset.georgetown\[.\]edu/publication/truth-lies-and-automation/](#)

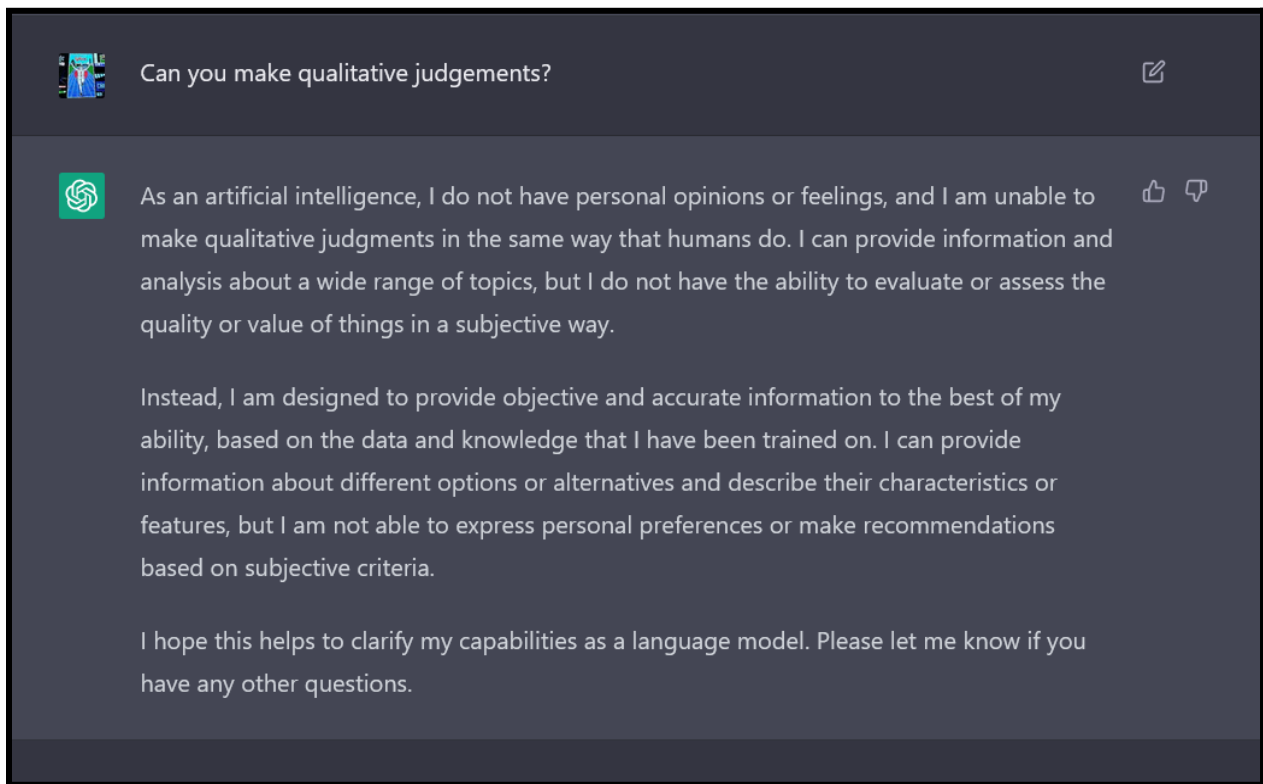
¹⁰ [hXXps://www.cyberscoop\[.\]com/chatgpt-ai-malware/](#)

¹¹ [hXXps://www.bloomberg\[.\]com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai](#)

¹² [hXXps://www.cnbc\[.\]com/2023/02/07/microsoft-open-ai-chatgpt-event-2023-live-updates.html](#)

¹³ [hXXps://blo\[g\].google/technology/ai/bard-google-ai-search-updates/](#)

¹⁴ [hXXps://ai.facebook\[.\]com/blog/large-language-model-llama-meta-ai/](#)

**ChatGPT Screenshot**

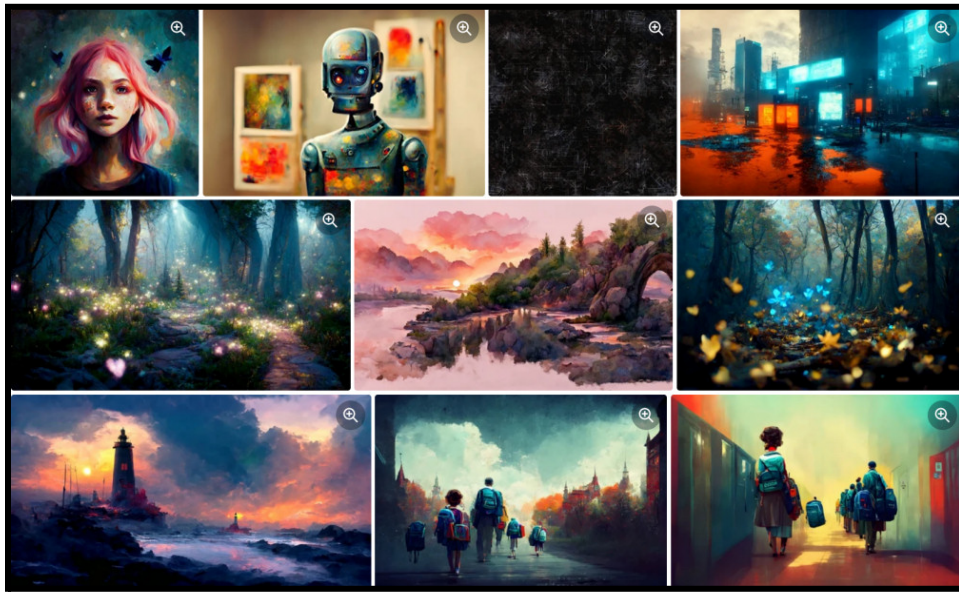
Source: [Medium\[.\]com](#)

| Artificial Intelligence (AI) Art

AI art refers to the computer generation of images, text-based art (e.g., poetry), and music. One of the biggest criticisms of AI art is that, unlike people, AI cannot create something new. Instead, AI art is created by scraping millions of existing samples and generating a derivative product without attribution or royalties paid to unwitting contributors. As several free applications for creating AI art are widely available, including [OpenArt](#), [Night Cafe](#), and [Hotpot](#), artists are losing work (and contest prizes) to AI.^{15,16}

¹⁵ [hXXps://www.vice\[.\]com/en/article/ake9me/artists-are-revolt-against-ai-art-on-artstation](https://www.vice[.]com/en/article/ake9me/artists-are-revolt-against-ai-art-on-artstation)

¹⁶ [hXXps://www.nytimes\[.\]com/2022/09/02/technology/ai-artificial-intelligence-artists.html](https://www.nytimes[.]com/2022/09/02/technology/ai-artificial-intelligence-artists.html)



Examples of AI-generated Art

Source: ArsTechnica[.]com

Beyond visual art, tools that author poetry—like [Verse by Verse](#) (from Google) and [Poem Generator](#)—and applications like [Soundraw](#) and [AIVA](#) that generate AI-created music are also abundant. While all of these tools may be fun for users and vexing to artists, they represent opportunities for threat actors. Threat actors can use these free tools to overcome language deficiencies and create professionalized, convincing websites, social media accounts, and emails that are more likely to earn the trust of their targets—leading to increased success in social engineering, phishing, business email compromise, brand impersonation, fake job postings, and romance scams.

| Threats Posed by Synthetic Media

The widespread use of synthetic media within the cyber threat landscape remains low, though evidence suggests its deployment in cybercrime is rising.¹⁷ Reported malicious incidents involving synthetic media remain sporadic and largely confined to the targeting of public figures. Nonetheless, many incidents have demonstrated the potential damage synthetic media can inflict on individuals, organizations, and society.

Cybercrime and Fraud

- July 2019: Threat actors used a voice-cloning tool to impersonate a chief executive, resulting in the fraudulent transfer of USD 243,000.¹⁸

¹⁷

[hXXps://blogs.vmware\[.\]com/security/2022/10/cybersecurity-awareness-month-amid-an-evolving-threat-landscape-defenders-continue-to-weather-the-storm.html](https://blogs.vmware.com/security/2022/10/cybersecurity-awareness-month-amid-an-evolving-threat-landscape-defenders-continue-to-weather-the-storm.html)

¹⁸ [hXXps://www.itpro\[.\]co\[.\]uk/social-engineering/34308/fraudsters-use-ai-voice-manipulation-to-steal-200000](https://www.itpro[.]co[.]uk/social-engineering/34308/fraudsters-use-ai-voice-manipulation-to-steal-200000)

- May 2022: A deepfake video of Elon Musk was used as part of a scam to encourage people to invest in an illegitimate trading scheme.¹⁹
- June 2022: The FBI warned of an increase in the use of deepfakes, combined with stolen credentials, to apply for remote work and work-at-home positions.²⁰
- August 2022: A Binance senior executive claimed threat actors used his likeness for deepfakes during Zoom calls in an attempt to scam other executives.²¹

Abuse, Harassment, and Extortion

- April 2018: Investigative journalist Rana Ayyub was subjected to prolonged harassment and extortion after opponents created deepfake sex videos of her.
- In 2020, 96 percent of deepfake content was non-consensual, fake pornography, and a single Telegram AI-bot reportedly victimized at least 104,000 women in under four months.^{22,23}

Mis/Dis/Malinformation

- Since at least 2019, Russia and China-affiliated actors reportedly leveraged GAN-generated images in false social media personas, negating reverse image searches as a means to detect online influence campaigns.²⁴
- February 2020: A video of Speaker of the House of Representative Nancy Pelosi slurring her words during an interview went viral. Slowing down the original video—a technique commonly known as “cheapfake” or “shallowfake” because of its lack of sophistication—was the cause of the alleged slurring.²⁵
- March 2022: A deepfake video of Ukrainian President Volodymyr Zelensky posted on a hacked Ukrainian news website shortly after Russia’s invasion of Ukraine advised Ukrainian soldiers and citizens to surrender to Russia.²⁶ The combination of the hacking of the website and the deepfake video lent credibility to the message to surrender.

ZeroFox Intelligence anticipates synthetic media will be increasingly leveraged by threat actors to enhance and evolve existing threats, especially as the tools for creating content become more sophisticated with low barriers for availability, cost, and expertise. The growing appetite for financially-motivated criminals to leverage

¹⁹

[hXXps://www.outlookindia\[.\]com/business/criminals-use-elon-musk-s-deepfake-video-to-dupe-crypto-investors-crypto-market-ri-ses-news-198403](https://www.outlookindia.com/business/criminals-use-elon-musk-s-deepfake-video-to-dupe-crypto-investors-crypto-market-ri-ses-news-198403)

²⁰ [hXXps://www.ic3\[.\]gov/Media/Y2022/PSA220628](https://www.ic3[.]gov/Media/Y2022/PSA220628)

²¹ [hXXps://www.malwarebytes\[.\]com/blog/news/2022/08/deepfake-hologram-targets-binance-and-crypto-community](https://www.malwarebytes[.]com/blog/news/2022/08/deepfake-hologram-targets-binance-and-crypto-community)

²² [hXXps://www.vice\[.\]com/en/article/7x57v9/most-deepfakes-are-porn-porn-harassment-not-fake-news](https://www.vice[.]com/en/article/7x57v9/most-deepfakes-are-porn-porn-harassment-not-fake-news)

²³ [hXXps://www.wired\[.\]co\[.\]uk/article/telegram-deepfakes-deepnude-ai](https://www.wired[.]co[.]uk/article/telegram-deepfakes-deepnude-ai)

²⁴

[hXXps://www.aha\[.\]org/system/files/media/file/2021/03/fbi-tlp-white-pin-malicious-actors-almost-certainly-will-leverage-synthetic-content-for-cyber-and-foreign-influence-operations-3-10-21.pdf](https://www.aha[.]org/system/files/media/file/2021/03/fbi-tlp-white-pin-malicious-actors-almost-certainly-will-leverage-synthetic-content-for-cyber-and-foreign-influence-operations-3-10-21.pdf)

²⁵

[hXXps://www.reuters\[.\]com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI](https://www.reuters[.]com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI)

²⁶ [hXXps://www.npr\[.\]org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia](https://www.npr[.]org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia)

deepfakes will likely drive ongoing innovation within closed forums. Concurrently, politically-motivated actors, including nation-states, will almost certainly increase their use of synthetic media to create convincing false personas that more effectively amplify MDM narratives on social media.

While practical use cases will continue to emerge due to innovation and increased demand for AI solutions, the improving realism and prevalence of synthetic media also pose many ethical and security concerns over its potential to be weaponized against individuals, organizations, and society. To date, the threats observed from synthetic media used in a malicious or offensive context include non-consensual adult content deepfakes, social engineering, fraud, and influencing public opinion through mis-, dis-, and malinformation.²⁷ In most cases, this has involved deepfake audio and video with varying degrees of sophistication. However, it is important to note that the underlying frameworks across the full range of synthetic media have the potential to be repurposed for financial or political gain by threat actors ranging from cybercriminals to nation-states.

²⁷ ZeroFox Intelligence, September 9, 2022. Retrieved from [hXXps://cloud.zerofox\[.\]com/intelligence/advisories/18274](https://cloud.zerofox[.]com/intelligence/advisories/18274).

| Recommendations

Despite the nascency of synthetic media within the threat landscape and the low frequency of attacks, mitigating these threats should be part of any cybersecurity plan.

Prevent

- Incorporate synthetic media education into existing cybersecurity training, including examples designed to increase workforce awareness.
- Review compliance procedures for financial transactions, providing greater latitude to challenge senior leadership requests.
- Document and track executive exposure in open and closed sources and reduce digital footprints to minimize the availability of media that fuels impersonation.
- Consume ZeroFox Synthetic Media Intelligence for ongoing awareness and recommendations for defending against synthetic media-enabled threats.

Detect

- Leverage deepfake detection technologies, such as [Sensity](#), [DuckDuckGoose](#), [Reality Defender](#), [deepware](#), or tools from Microsoft and Intel.
- Monitor corporate social media and websites for signs of manipulation.
- Inspect images used by third-party profiles for distortions, indistinct and blurry backgrounds, and other visual artifacts often found in synthetic images.

Respond

- Create a crisis response plan to neutralize and contain any incidents of deepfake impersonation or MDM targeting the organization.

Protect

- Apply watermarks and/or use blockchain technology to verify media content.²⁸
- Explore digital provenance solutions through the [Coalition for Content Provenance and Authenticity \(CP2A\)](#) or the [Content Authenticity Initiative \(CAI\)](#).

²⁸ <https://c2pa.org/>

Appendix: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	Sources may use TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	Sources may use TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	Recipients may NOT share TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	Recipients may ONLY share TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. Note that TLP:AMBER+STRICT restricts sharing to the organization only.
	Green	Clear
WHEN SHOULD IT BE USED?	Sources may use TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	Sources may use TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	Recipients may share TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	Recipients may share TLP:CLEAR information without restriction, subject to copyright controls.