Deloitte.

How to safeguard against the menace of deepfake technology

The battle against digital manipulation

March 2024

Contents

1. Introduction

In today's dynamic technological landscape, the emergence of Generative Artificial Intelligence (Generative AI) and its blend with deepfake technology has affected voice, image, and video. This amalgamation has opened a landscape of opportunities and risks. The recent spate of deepfake videos targeting women celebrities results in reputational and emotional damage, and privacy invasion. These videos can also lead to trauma.[1]

Consumers rely on digital multimedia content; there is a significant increase in risks associated with these new technologies. For example, people working remotely and relying solely on video/audio calls for their daily tasks, are more vulnerable to attacks, such as deepfake and phishing.[2]

The evolution of deepfake technology has raised significant concerns about its potential misuse and threats to various facets of society. To address these challenges, a proactive approach towards safeguarding against deepfakes is imperative. This point of view aims to discuss the convergence of Generative AI and deepfake technology, illuminating the potential weaknesses and risks they introduce to biometric authentication systems (voice and video-based authentication systems). Once seen as a robust security measure, biometric authentication faces extraordinary challenges due to the malicious use of Generative AI, which can create deceptively realistic voices/videos.

As the digital landscape continues to evolve, understanding and addressing these vulnerabilities is crucial to fortify the foundations of digital security in an age defined by Generative AI and deepfake innovation.

[1] https://www.indiatoday.in/movies/celebrities/story/rashmika-mandanna-on-deepfakes-we-have-normalised-them-but-it-is-not-okay-2468209-2023-11-27

[2] https://www.bloomberg.com/news/articles/2023-08-25/deepfake-video-phone-calls-could-be-a-dangerous-ai-powered-scam#xj4y7vzkg

2. What is deepfake technology?

Deepfake technology refers to using Artificial Intelligence (AI) and machine learning techniques, particularly deep learning, to create realistic-looking but fabricated or manipulated audio, videos, or image content. This trend goes back to 2014 research presented by Ian Goodfellow on deepfakes. Almost 10 years later making a deepfake is significantly easier due to the readily available resources/SDKs online.[3]

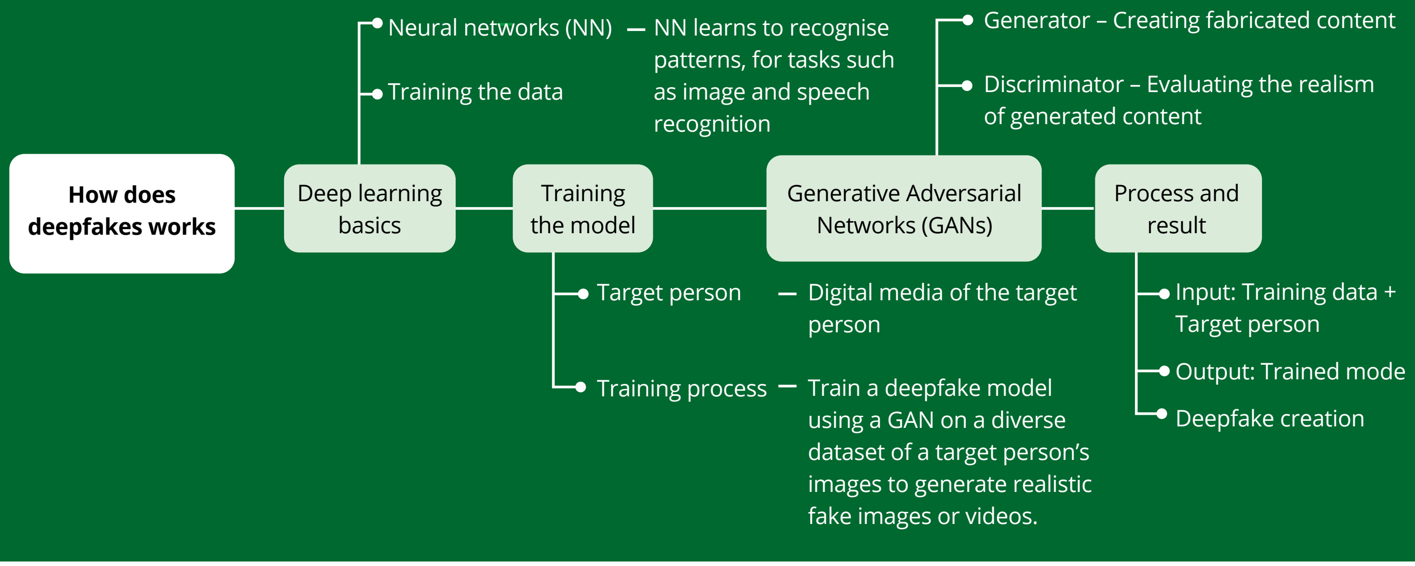
This technology uses Generative Adversarial Networks (GANs) to create content that appears highly realistic but is manipulated or fabricated.

How does deepfake technology work (Figure 1)?

By creating convincing copies of individuals' voices or appearances, deepfakes can deceive the biometric authentication mechanisms, compromising the security and integrity of these systems. As the deepfake technology advances, this potential for manipulation raises concerns about the vulnerability of such biometric security measures.

Recently, a deepfake audio of a chief executive was used to trick the CEO of a UK-based energy firm. The CEO believed that he was talking to his boss (the chief executive) who requested him to transfer US$2,43,000 as funds to a Hungarian supplier. This is an example of an advanced spear phishing attack, which is performed through deepfake technology and can easily result in the manipulation of the target.[4]

Figure 1: Deepfake technology


How does deepfakes works
Deep learning basics
Neural networks (NN)
NN learns to recognise patterns, for tasks such as image and speech recognition
Training the data
Training the model
Target person
Digital media of the target person
Training process
Train a deepfake model using a GAN on a diverse dataset of a target person's images to generate realistic fake images or videos.
Generative Adversarial Networks (GANs)
Generator – Creating fabricated content
Discriminator – Evaluating the realism of generated content
Process and result
Input: Training data + Target person
Output: Trained mode
Deepfake creation

[3] https://arxiv.org/pdf/1406.2661.pdf

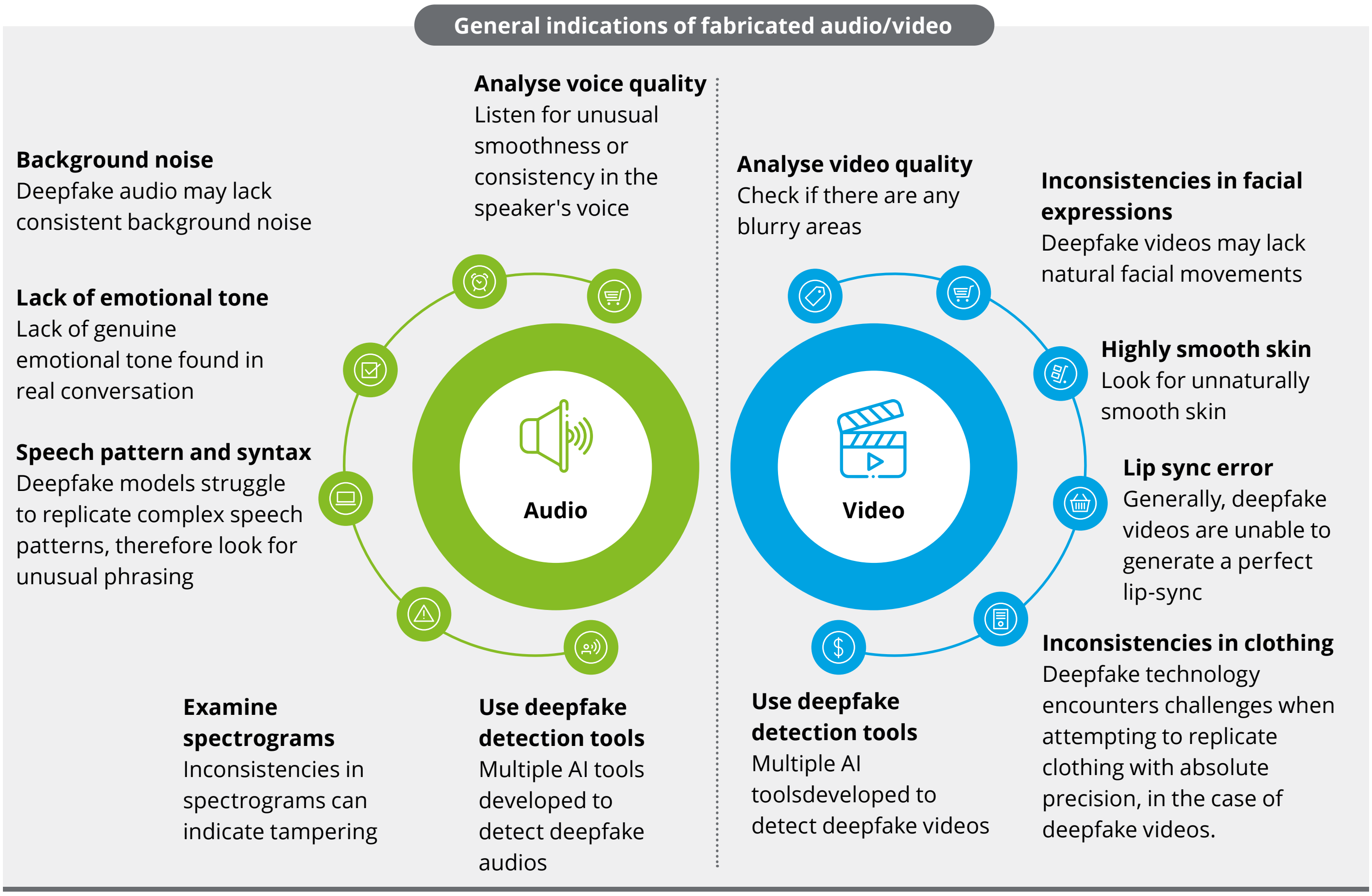
[4] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

3. General indications of fabricated audio/video

Detecting fabricated audio and video content, especially deepfakes, can be quite challenging.

Some general indications and techniques that people can use to identify potential manipulation include the following (Figure 2):

Figure 2: General indications of fabricated audio/video



General indications of fabricated audio/video

Analyse voice quality
Listen for unusual smoothness or consistency in the speaker's voice

Background noise
Deepfake audio may lack consistent background noise

Lack of emotional tone
Lack of genuine emotional tone found in real conversation

Speech pattern and syntax
Deepfake models struggle to replicate complex speech patterns, therefore look for unusual phrasing

Examine spectrograms
Inconsistencies in spectrograms can indicate tampering

Use deepfake detection tools
Multiple AI tools developed to detect deepfake audios

Analyse video quality
Check if there are any blurry areas

Inconsistencies in facial expressions
Deepfake videos may lack natural facial movements

Highly smooth skin
Look for unnaturally smooth skin

Lip sync error
Generally, deepfake videos are unable to generate a perfect lip-sync

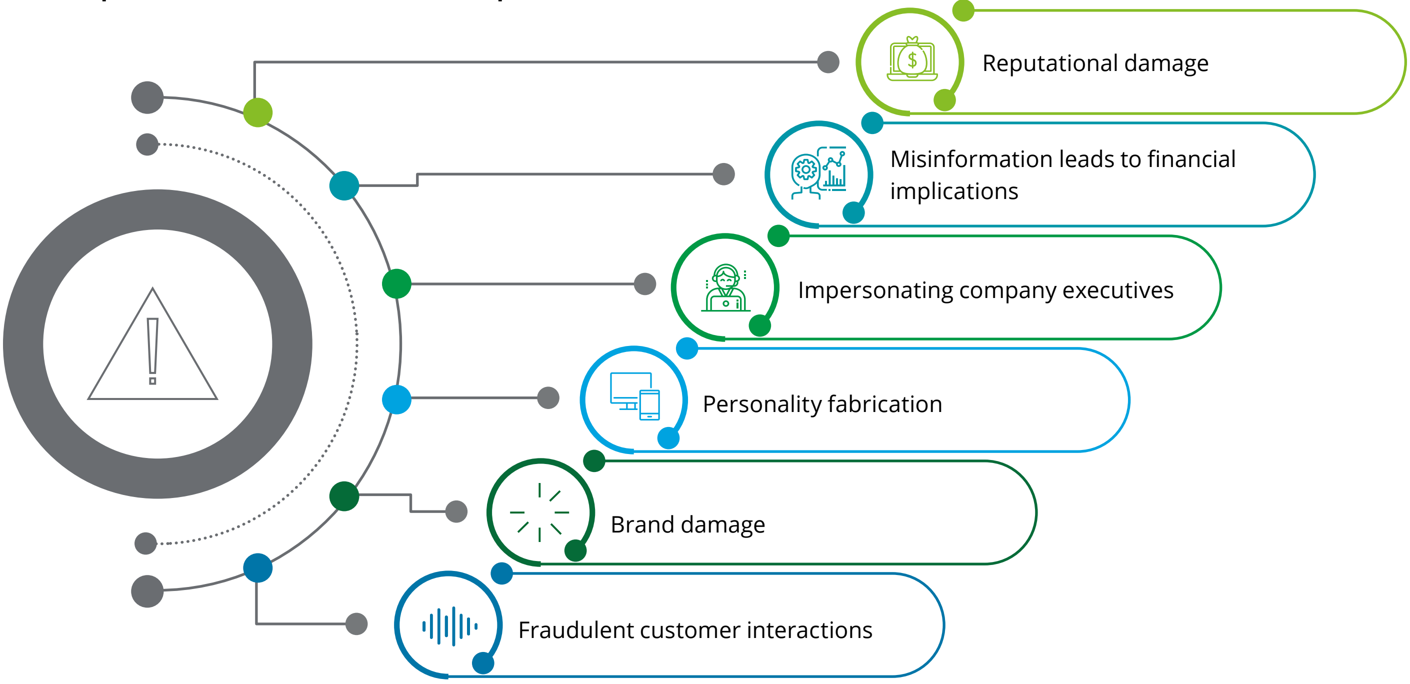
Inconsistencies in clothing
Deepfake technology encounters challenges when attempting to replicate clothing with absolute precision, in the case of deepfake videos.

Use deepfake detection tools
Multiple AI toolsdeveloped to detect deepfake videos

4. Risks, threats, and impact of deepfakes

Figure 3: Deepfakes impact on corporate

How deepfakes (video and AI) can affect corporates?


Reputational damage
Misinformation leads to financial implications
Impersonating company executives
Personality fabrication
Brand damage
Fraudulent customer interactions

In cybersecurity, malicious actors can use deepfake technology to create counterfeit content and enable social engineering attacks, such as phishing or spear-phishing campaigns. These risks underline the urgent need for robust countermeasures to safeguard against the detrimental impact of deepfake-generated cybersecurity threats and privacy breaches. A host of vulnerabilities and threats have emerged. These are primarily driven by advancements in Generative AI and deepfake technology. The imminent threats these technologies pose to the integrity of voice and video authentication systems include the following:

1) **Evasion of authentication –** Threat actors can create fabricated audio or video through deepfakes to bypass authentication systems.

   Incident: Multiple banks across the world use voice ID as an authentication mechanism for individuals to log into their accounts. However, deepfake technology has made it easy to bypass such authentication methods.[5]

2) **Impersonation through deepfakes –** Deepfake technology can be used to create realistic fake audio, videos, and images that help the attacker perform impersonation.

   Incident: A branch manager of a Japanese company in Hong Kong received a call from a man whose voice he recognised—the director of his parent business requesting him to transfer US$35 million. The branch manager even received an e-mail confirmation and initiated the process. A deepfake audio was used to commit this crime.[6]

3) **Theft of personal data –** Theft of personal data, such as voice print or facial data of a user is another risk related to deepfakes. This data can be used to train multimedia artefact.

[5] https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice

[6] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=550684d75591

Incident: A deepfake attack was orchestrated against the CFO of a multinational firm resulting in a loss of US$25 million, by utilising publicly available facial data from videos and images.[7]

4) **Bypassing eKYC –** In the digitalisation of processes, within the financial sector, a person can complete their eKYC process online through automatic video identification. This can become a challenge with the entrance of deepfake technology.[8]

5) **Forging fake documents through Artificial Intelligence –** An alarming concern revolves around the capability of malicious actors using Generative AI and deepfake technology with remarkable accuracy to create legitimate-looking documents, such as government IDs and other legal documents.

6) **Reputational damage –** A good reputation is important for an organisation as it influences its clients, credibility, and overall success.

   Incident: A fake but realistic-looking video of a CEO making foul comments or contradicting statements can hamper a brand's image and lead to reputation loss.[9]

7) **Deepfake employment interviews –** Criminals are using a combination of deepfake videos and stolen personal data to misrepresent themselves and gain employment in various work-from-home positions. This increases the risk of insider attacks.[10]

8) **Fraudulent customer interactions –** Deepfake voices can be used in social engineering attacks when threat actors use AI-generated voices to impersonate customers, bank executives, or any other individual. Social engineering may lead to the target divulging confidential information and falling victim to fraudulent transactions.

9) **Misinformation leading to financial implications –** An AI-generated fake image/video can be used to spread misinformation, which can lead to financial implications for the organisation.

   Incident: An AI-generated fake image of the Pentagon Explosion circulated over social media platforms resulted in the S&P 500 falling 30 points within minutes. This is a clear example of how AI-generated misinformation could affect the financial markets.[11]

[7] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[8] https://complyadvantage.com/insights/what-is-ekyc/

[9] https://economictimes.indiatimes.com/news/company/corporate-trends/moodys-says-deepfake-disinformation-campaigns-pose-reputational-risks-to-businesses/articleshow/70496252.cms?from=mdr

[10] https://www.darkreading.com/attacks-breaches/criminals-deepfake-video-interview-remote-work

[11] https://www.thestreet.com/technology/s-p-sheds-500-billion-from-fake-pentagon-explosion

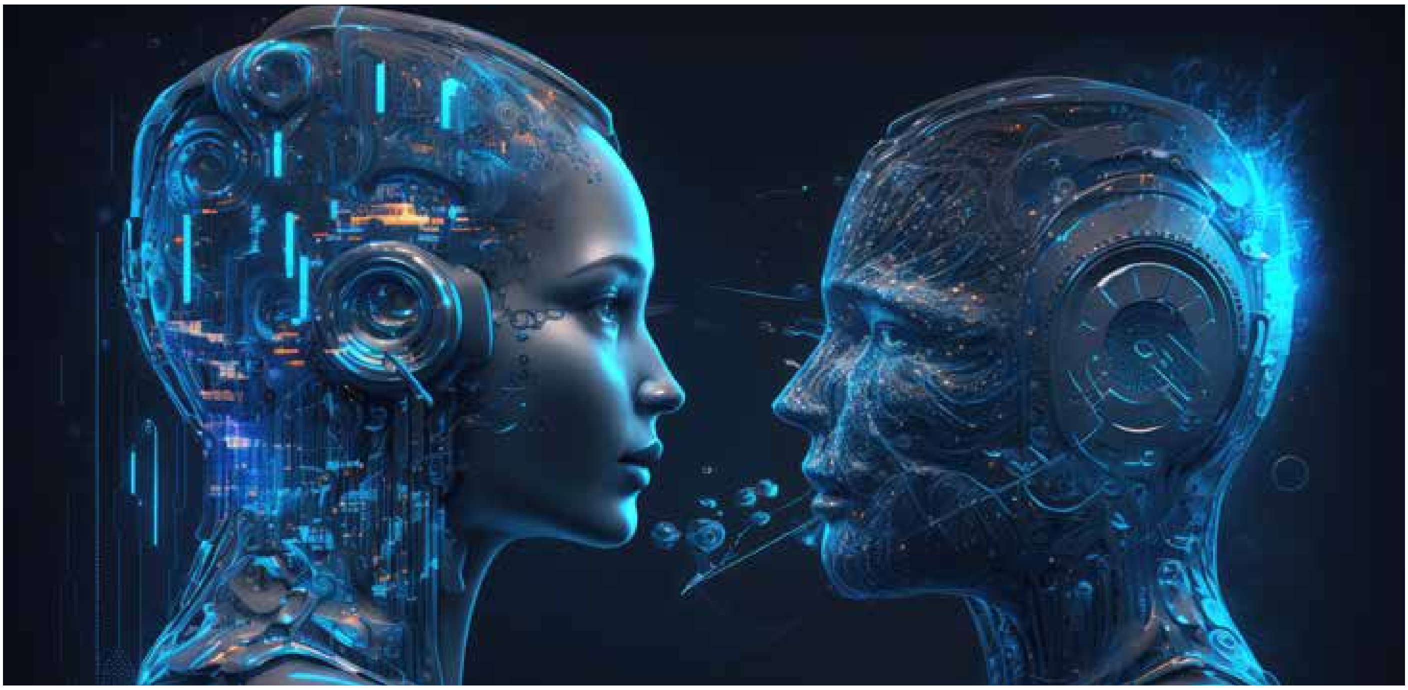
5. Global regulatory perspectives on deepfake and Generative Artificial Intelligence

In the past, we have seen a lot of regulatory traction in the space related to Generative-AI. Some advise or lay the foundation on how the Generative-AI space is supposed to function or be governed and others take a more restrictive approach. For instance, the US white house recently established a new standard "Executive Order on The Safe, Secure and Trustworthy Development and Use of Artificial Intelligence".[12]

The European Union has published the AI Act, which categorised various implementations per risk levels viz. unacceptable risk, high risk, and limited risk. For instance, any kind of ranking based on socioeconomic factors is classified as an unacceptable risk. In addition, cognitive behavioural manipulation of people or specific groups, such as children being encouraged to do something harmful driven via an AI-based voice system is also classified as an unacceptable risk.[13]

The Indian government also expressed its concern about the deepfake technology during the G20 summit this year. The government highlighted the growing concerns about the misuse of generative AI and called for a global regulation where the world can work together to find a better solution.

In addition, the IT Minster highlighted imposing a financial penalty on the deepfakes (even if they are not produced in India) that are misused. The responsibility of these deepfakes will be attributed to both the author of the deepfake and the platform used to spread the deepfake.[14]



[12] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/]

[13] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

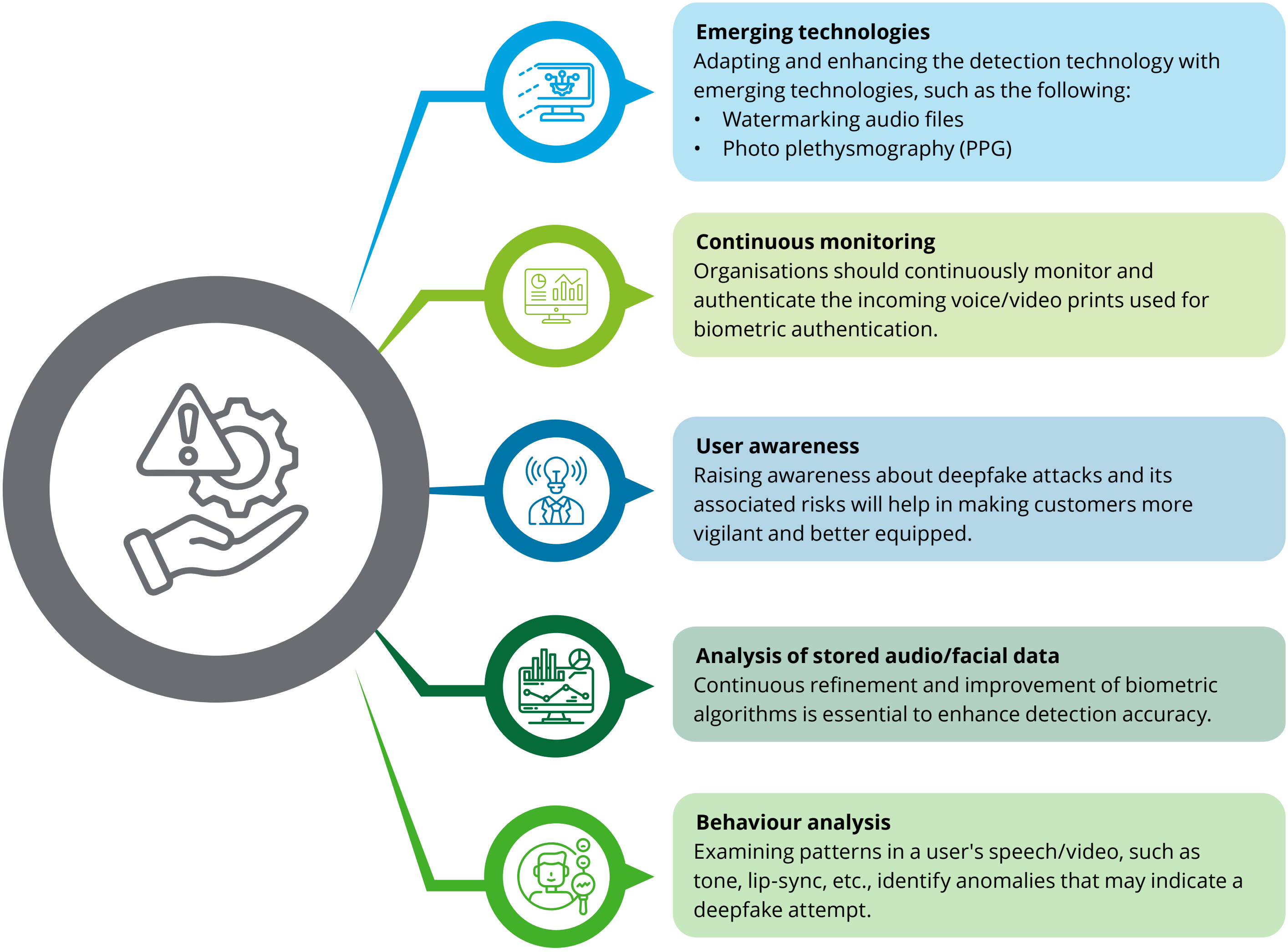
[14] https://www.businesstoday.in/technology/news/story/ashwini-vaishnaw-on-deepfake-menace-govt-considering-penalties-on-both-creator-and-platform-406830-2023-11-23

6. Risk mitigation strategies for deepfake attacks

Mitigating deepfake attacks demands a multi-layered strategy that combines technological advancements, proactive detection mechanisms, and robust authentication tools. Implementing AI-driven detection systems and constantly staying updated with evolving algorithms (to recognise patterns unique to deepfakes) forms a critical layer in combating these threats.

Some of the risk mitigation approaches for organisations against deepfakes include the following (Figure 4):

Figure 4: Risk mitigation approaches



Emerging technologies
Adapting and enhancing the detection technology with emerging technologies, such as the following:

- Watermarking audio files
- Photo plethysmography (PPG)

Continuous monitoring
Organisations should continuously monitor and authenticate the incoming voice/video prints used for biometric authentication.

User awareness
Raising awareness about deepfake attacks and its associated risks will help in making customers more vigilant and better equipped.

Analysis of stored audio/facial data
Continuous refinement and improvement of biometric algorithms is essential to enhance detection accuracy.

Behaviour analysis
Examining patterns in a user's speech/video, such as tone, lip-sync, etc., identify anomalies that may indicate a deepfake attempt.

- Emerging technologies – To combat emerging technologies, organisations can:
  - **Watermark audio files**
    It is a process of embedding information into the audio signal that acts as a signature, which can be used to verify the authenticity of the audio content. This signature is generally inaudible to human ears. Since it is embedded, it is difficult to remove without affecting the audio quality.
  - **Use Photoplethysmography (PPG)**
    PPG is used to detect blood volume changes in the microvascular bed of tissues. Remote PPG, which is a technique to obtain PPG signals through non-contact methods, such as cameras, helps detect deepfake videos.[15]
- **User awareness –** Raise awareness about deepfake attacks and their associated risks as this will help customers be more vigilant and stay better equipped.
- **Continuous monitoring –** Continuous monitoring and authentication of incoming voice/video prints can be used for biometric authentication.
- **Multi-factor authentication –** Combine biometric authentication with other factors, such as passwords and PINs, to reduce the risk associated with a single compromised factor.[16]
- **Behaviour analysis –** Examine patterns in a user's speech/video, such as tone and lip-sync to identify anomalies that may indicate a deepfake attempt.
- **Analysis of stored audio/facial data –** Continuous refinement and improvement of biometric algorithms are essential to enhance detection accuracy.

[15] https://link.springer.com/chapter/10.1007/978-3-031-27199-1_1

[16] https://economictimes.indiatimes.com/wealth/save/hello-upi-use-voice-commands-to-send-money-pay-bills-know-new-upi-features-and-how-they-work/articleshow/103464077.cms

7. Key recommendations

Fostering collaboration between consumers, tech experts, law enforcement, and policymakers to establish comprehensive legal frameworks and regulations is pivotal in addressing the rapidly evolving landscape of deepfake threats. This multi-pronged approach, including technological innovation, vigilant monitoring, and collaborative efforts, stands as a robust defence against the growing risks posed by deepfake technology.

Prevention strategies for organisations: Enterprises need to adopt appropriate measures to protect their data and reduce threats presented by malicious deepfake attacks. Organisations can employ the following prevention strategies to safeguard their data and proprietary information:

- Implement strong cybersecurity measures and privacy policies to safeguard employee data.
- Restrict internal access to sensitive information and use robust authentication methods.
- Add a layer of verification that can significantly bolster security and reduce the risk of unauthorised access, especially when a video or audio-based authentication is involved in business processes.
- Clarify verification protocols for payment requests, especially from senior executives to prevent fraud.
- Develop detailed incident response and business continuity plans to counter data breaches.
- Conduct routine cybersecurity trainings to educate employees about deepfake risks and threat identification/ reporting.

Prevention strategies for end-users: Our inclusive strategies equip end-users with security measures to safeguard against the inherent risks presented by these evolving technologies. Some of these include the following:

- As a consumer, it is crucial to be cautious about what information you share online and reflect on how it could be misused. Some social media platforms proactively implement security measures, such as making profile pictures non-downloadable by default. However, these settings often need to be enabled manually, highlighting the importance of awareness.
- End-users should always cross-check the information with reliable sources before believing it and check for the general indications of the fabricated audio/view before sharing it.
- Beware of the content shared from non-reputable sources.
- Enable two-factor authentication to add an extra layer of security.
- Update the software, such as antivirus software (used regularly) and operating system to ensure that none of the latest security updates are missed.
- Create strong and unique passwords for your accounts to reduce the risks of unauthorised access.
- Do cross-check the identity of the person who contacts you through digital media, especially if they request sensitive information.

8. Conclusion

In a landscape where deepfake threats are becoming increasingly sophisticated, businesses must prioritise proactive measures. Fortifying cybersecurity practices, enhancing employee awareness, and implementing stringent verification protocols will help businesses significantly reduce the risks associated with deepfake technology. These measures can safeguard its integrity and operations by investing in research and innovation. Since this landscape is still evolving there are no fool-proof solutions. The only way to stay secure is to be aware of trends and research ways by which consumers and companies can safeguard themselves. The only feasible way to tackle AI-based risk is to use AI-based solutions.

Connect with us

Deepa Seshadri
Partner & Leader – Cyber
Deloitte South Asia
deseshadri@deloitte.com

Gaurav Shukla
Partner
Deloitte India
shuklagaurav@deloitte.com

Anand Tiwari
Partner
Deloitte India
anandtiwari@deloitte.com

Tarun Kaura
Partner
Deloitte India
tkaura@deloitte.com

Praveen Sasidharan
Partner
Deloitte India
psasidharan@deloitte.com

Santosh Jinugu
Partner
Deloitte India
sjinugu@deloitte.com

Contributors

Anas Jamal

Arjuna KS

Rins John

Saubhagya Srivastava

Sanchi Gabrani

Swarup Sonar

Deloitte.

This communication contains general information only, and none of DTTL, its global network of member firms or their related entities is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication.