WHITE PAPER

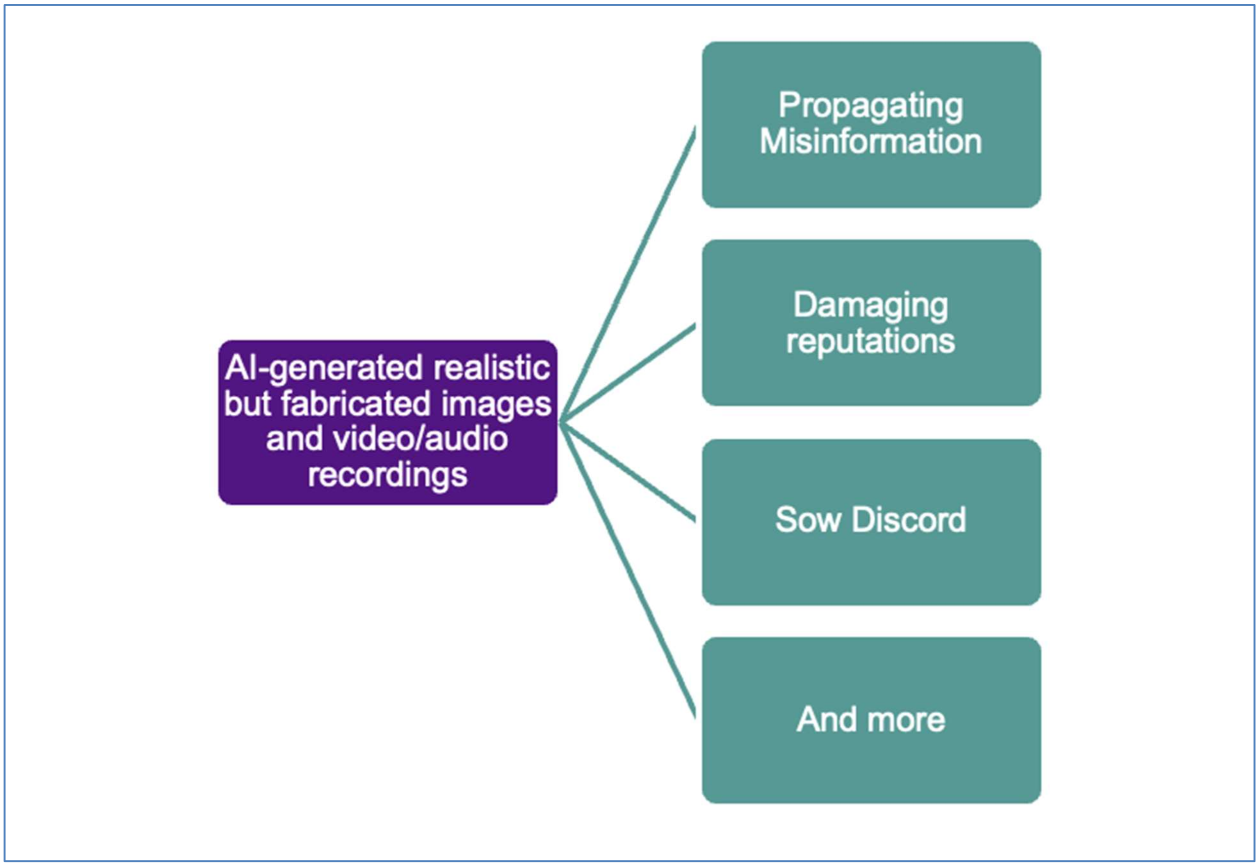
Managing protocols and best practices countering deep fakes

Table of Contents

Abstract:

Deepfakes, synthetic media manipulated by AI, pose a significant threat. This white paper explores deepfake detection, identification, and prevention strategies to build a deepfake-resilient future.

The rise of Deepfakes and the threats they pose


AI-generated realistic but fabricated images and video/audio recordings
Propagating Misinformation
Damaging reputations
Sow Discord
And more

- Deepfakes are synthetic media videos that use artificial intelligence to realistically superimpose a person's likeness onto another person's body. They can be used to create highly believable forgeries, posing a significant threat to individuals, businesses, and democratic processes. In this whitepaper, we'll explore the challenges of deepfakes and showcase our capabilities for managing protocols and practices to counter them.

Deepfake Detection: Identifying the Fabrications:

- The first line of defense against deepfakes is their detection. We leverage cutting-edge deepfake detection tools trained on comprehensive datasets to identify manipulated videos. These algorithms go beyond basic visual cues and analyze intricate details to ensure the authenticity of content. (For example, Zelenski asking troops to lay arms in conflict, a deepfake video detected before it caused a major threat to Ukraine.)

- Deepfake detection utilizes machine learning algorithms trained on vast datasets of real and fake videos.
- These algorithms analyze facial movements, lip-syncing, and other visual inconsistencies to identify deepfakes.
- Advanced techniques involve analyzing voice patterns, blinking rates, and subtle physiological details for increased accuracy.
- We believe in building strong partnerships to address the challenges of deepfakes. We work collaboratively with customers to understand their needs and customize our deepfake management strategy accordingly. Our ongoing monitoring and support ensure your continued protection in the ever-evolving digital landscape.

- The sample tools are available and have an accuracy percentage of >90%. In some cases, up to 95%

intel.
FakeCatcher
WeVerify
DUCKDUCKGOOSE
SENTINEL
FACIA
Microsoft
Video Authenticator

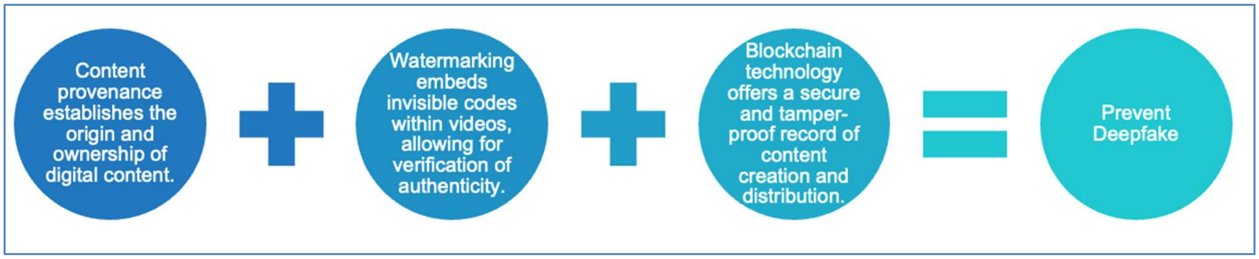
Technique Viseme Mismatches[1]:

Recent advancements in machine learning and computer graphics have made video and audio manipulation easier. Deepfakes range from complete face replacements to lip-syncing and partial audio manipulation. Detecting deepfakes with minor spatial and temporal manipulation is challenging. A technique to detect such manipulations exploits inconsistencies between mouth shapes ("visemes") and spoken phonemes. Focusing on visemes associated with words like "mama," "baba," or "papa" reveals inconsistencies in some deepfakes. These phoneme-viseme mismatches can be used to detect even subtle manipulations.

Alternate methods:

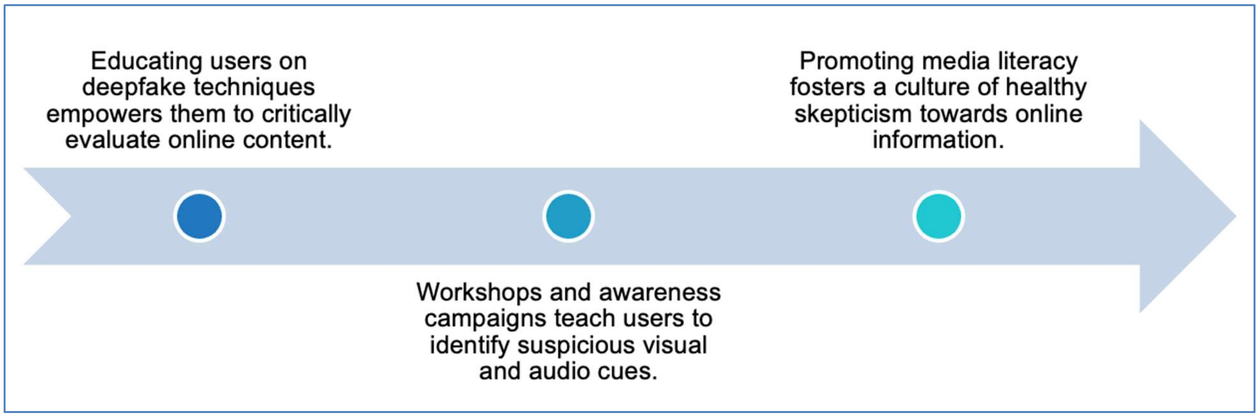
When deepfakes generate new facial expressions, the new images may not perfectly match the person's head position, lighting, or camera distance. To blend the fake faces with the surroundings, they are geometrically transformed, leaving digital artifacts. These artifacts can be subtle, but algorithms can be trained to detect them, even when human eyes fail to do so.

Content Provenance and Verification: Building Trustworthy Sources


Content provenance establishes the origin and ownership of digital content.
Watermarking embeds invisible codes within videos, allowing for verification of authenticity.
Blockchain technology offers a secure and tamper-proof record of content creation and distribution.
Prevent Deepfake

Building trust in the digital age requires establishing content provenance. To ensure the authenticity of content, we explore various techniques, including watermarking and blockchain. By identifying the source and ownership of videos, we can combat the spread of misinformation and deepfakes.

User Education and Media Literacy: Empowering Your Audience


Educating users on deepfake techniques empowers them to critically evaluate online content.
Promoting media literacy fosters a culture of healthy skepticism towards online information.
Workshops and awareness campaigns teach users to identify suspicious visual and audio cues.

Empowering the audience is crucial in the fight against deepfakes. We advocate for user education and media literacy initiatives. By equipping users with the skills to analyze online content critically, we can collectively mitigate the impact of deepfakes.

Recommended Deepfake Management Strategy: A Multi-Step Approach

- Deepfake management strategy combines the strengths of detection, provenance, and user education.

- Leverage modern tools such as Intel Fakecatcher, Facia, WeVerify, DuckDuckGoose, Sentinel, Microsoft Video Authenticator, and fakebusters2.0 for continuous monitoring and identification for different use cases.
- Implement content provenance techniques to ensure the authenticity of online sources.
- Prioritize user education through workshops and awareness campaigns to empower audiences.

Building a Deepfake Resilient Future

- Understand your specific needs and customize the Deepfake management strategy accordingly. Ongoing monitoring and support ensure your continued protection in the ever-evolving digital landscape.

Conclusion

Though it is an area of concern and complexities, if we approach it methodically, it is possible to detect most of the deep fakes by applying various techniques and tools and have customizations for specific use cases leveraging existing models to detect different types of deep fakes, be it audio, video, image or text/

References:

[1]: https://ieeexplore.ieee.org/document/9151013
published in https://ieeexplore.ieee.org/xpl/conhome/9142289/proceeding
and
https://openaccess.thecvf.com/content_CVPRW_2020/papers/w39/Agarwal_Detecting_Deep-Fake_Videos_From_Phoneme-Viseme_Mismatches_CVPRW_2020_paper.pdf

[2]: https://theconversation.com/detecting-deepfakes-by-looking-closely-reveals-a-way-to-protect-against-them-119218

Other references
https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1520
https://staysafeonline.org/resources/how-to-protect-yourself-against-deepfakes/
https://www.mcafee.com/blogs/internet-security/deepfake-defense-your-8-step-shield-against-digital-deceit/
https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection

sources: www.mdpi.com/1999-5903/13/4/93

Citations?

Write to connect@cuttingej.com