Discussion Paper: The Threat of Real Time Deepfakes

Guy Frankovits, Yisroel Mirsky

guyfrank@post.bgu.ac.il,yisroel@bgu.ac.il

Ben-Gurion University

Department of Software and Information Systems Engineering

Beersheba, Israel

Figure 1: An example of a high quality real-time deepfake used in America's Got Talent [1] created by Metaphysic [2]. The ability to impersonate people in real-time poses a significant threat to society. This because voice and video calls can not longer be trusted based on content alone.

ABSTRACT

Generative deep learning models are able to create realistic audio and video. This technology has been used to impersonate the faces and voices of individuals. These "deepfakes" are being used to spread misinformation, enable scams, perform fraud, and blackmail the innocent. The technology continues to advance and today attackers have the ability to generate deepfakes in real-time. This new capability poses a significant threat to society as attackers begin to exploit the technology in advances social engineering attacks. In this paper, we discuss the implications of this emerging threat, identify the challenges with preventing these attacks and suggest a better direction for researching stronger defences.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

Deepfake, deep fake, social engineering, cyber security, impersonation, phishing

1 INTRODUCTION

A deepfake is synthetic media (e.g. video, audio) generated by a deep neural network, where the content looks authentic to a human [16]. The most common use of deepfake technology is to replace the identity of individuals in images and audio. In the year since its introduction in 2017, the technology has witnessed significant improvements in both the quality of the fake media it generates and the time it takes to generate them. With modern deepfake

Guy Frankovits, Yisroel Mirsky

technology, it is possible to create content that looks authentic to humans. As a result, deepfakes have found numerous applications across various domains and industries. For example, the technology can be used to recreate a digital reanimation of the deceased [12], to enhance productivity in the film industry [7], can be used in the education field [8] and in many other areas where there is a need to create content in an easier and more flexible way.

Deepfake technology is highly accessible to even novice users [3, 11, 14, 17]. This means that anyone can make convincing content of any individual. Malicious actors have noticed this capability and have used the technology for defamation, blackmail, misinformation, and social engineering attacks [19]. For example, in 2017 deepfake pornography surfaced on the Internet in 2017 raising serious concerns over the impact on women's privacy and dignity [9]. More recently, in March 2022, amidst the ongoing conflict between Russia and Ukraine, a deepfake video was circulated that depicted the Prime Minister of Ukraine instructing his troops to give up and stop fighting. This video had the potential to spread false information and undermine the credibility of the Ukrainian government and military [21].

However, there is a new threat on the horizon: real-time deepfakes.[1] Imagine a technician at a power plant who receives a phone call from his manager. The manager tells him to urgently change some configuration or provide him access to some resources. However, the technician is not speaking to his manager but rather an imposter. This scenario is possible with the latest deepfake technology which can generate a face or voice in real-time, enabling attackers to interact with their victims (an example of this technology is in Fig. 1). This is a dangerous precedent because it gives attackers the ability to craft convincing false pretexts to manipulate their victims. This is because familiarity can be mistaken for authenticity. For example, in 2019 an attacker used the voice of a CEO to convince an associate to transfer $243,000 in support of a made-up business emergency [18]. Another case was in 2021 where criminals pulled off a $35 million dollar bank heist by tricking a banker by using a customer's voice [5].

Real-time deepfakes (RT-DFs) also pose a threat to everyday people. For example, scammers can call up the elderly using their children's voices. This is feasible since some technologies only need a few seconds of the target's voice to clone their identity [11]. In June 2022, the FBI released a warning that cyber criminals are using RT-DFs in job interviews in order to secure remote work positions and gain insider information. Then in August that year, cyber criminals attended Zoom meetings masquerading as the CEO of Binance [20].

2 DISCUSSION

Deepfakes are not a new threat. Since their discovery in 2017,researchers have proposed a wide variety of ways to detect these attacks and mitigate their misuse [4, 15, 16, 23]. However, most of these defences are designed to work on offline media, not streaming media. Detecting and preventing real-time deepfake attacks is

---

[1]Examples of RT-DF tools: https://github.com/iperov/DeepFaceLive
https://github.com/alievk/avatarify-python
https://github.com/yl4579/StarGANv2-VC
https://github.com/CorentinJ/Real-Time-Voice-Cloning

not as simple. Here are some aspects that make it hard to develop defences against these attacks:

Practicality. To protect users from fake calls, we can deploy detection models on the user's phone or laptop. However, this means that the models must be efficient. Moreover, they must also run in real-time to be able to monitor calls. These limitations are great challenges considering that most defences are based on deep learning which requires a considerable amount of resources to run. Moreover, attackers are cunning. If a detection model is only execute just before the call (e.g., as authentication), then the attacker could just switch between the real and deepfake content to evade detection.

Media Quality. Many defences assume that deepfake generation models will leave semantic, stylistic, or forensic evidence in the generated content. However, voice and video calls often undergo significant data compression which degrades signals left in the media. This also means that attackers can intentionally increase the level of compression to cover up their defects without raising suspicion.

Delivery. Recently, the world has moved to a wider use of remote communication. With so many different ways once can communicate over a call or virtual meeting, it becomes difficult to develop a single solution that fits all scenarios. Moreover, some victims may still proceed with a flagged call, simply because they are worried that the system may be mistaken (imagine the caller being your boss, angry and insisting that you follow through with the call).

However, even if you could make a generic and practical detection model that can overcome compression, the model would still be looking for artifacts in the content itself. This is a losing battle. Looking at the advancements of generative AI over the last 10 years, it seems that 'perfect' deepfakes are an inevitability (see Fig. 2). Therefore, we expect that *passive* solutions that rely solely on artifacts, will be rendered ineffective in a matter of years. Moreover, there are other ways that attackers can hide evidence in the content. For example, adversarial machine learning can be used to make detection models classify deepfakes as real [6]. Moreover, attackers can also remove specific artifacts. For example, models that search for blending boundaries can be mitigated by passing the final deepfake through a refiner such as a CycleGAN. Methods that look for biological signals [10] can be evaded by adding this ground truth to the model's [13] training. As a result, state-of-the-art defences do not provide long-term solutions.

To prepare for next-generation deepfakes, we cannot rely on passive methods of content analysis. In other words, we must seek solutions that expect that the media will reveal no forensic evidence of generative AI. There are at least two different defence strategies that accomplish this:

Active defences directly challenge the attacker. This gives the defender an advantage since he or she can now be proactive in his or her defence strategy. For example, in [22] the authors propose using a new kind of CAPTHCA to detect deepfake calls. Callers are authenticated by performing a task that is hard for a deepfake model but easy for a human to perform (e.g., press on the cheek or turn around). This is

Figure 2: A decade of deepfakes. The progression of facial deepfakes over the last ten years. 2014-2017 are images from GANs and such as StyleGAN. 2018 is from Buzzfeed. 2019 is a real-time deepfake that animates a single image. 2020-2023 are examples showing how deepfake quality is improving.

essentially a new kind of Turing test where the challenge is on creating content as opposed to interpreting content.

Out of band defences expose attacks by analyzing information surrounding the media. For example, by tracking the source of the media or by analysing the context surrounding the media. For example, a real-time deepfake caller could be revealed by simply determining the caller's origin or by verifying the status of the supposed caller (e.g., can we verify that the true individual is in a call right now?)

These approaches arguably provide better defences against the emerging threat real-time deepfakes. However, there is a lot of research which needs to be done to develop these defences. Therefore, we urge the research community to change paths and seek out more effective defences.

3 CONCLUSION

Real-time Deepfakes are an imminent threat and an urgent issue that must be dealt with. The technology provides malicious actors with the ability to perform powerful social engineering attacks. Although defences have been developed, many cannot be used to protect voice and video calls and many are likely to become obsolete as the quality of deepfakes improve. We encourage researchers to try to think of solutions that do not rely only on media content, thus giving the defenders a chance to keep ahead of the threat.

4 ACKNOWLEDGEMENTS

REFERENCES
[1] [n. d.]. America's Got Talen. https://www.nbc.com/americas-got-talent.
[2] [n. d.]. metaphysic AI. https://metaphysic.ai/.
[3] [n. d.]. Online Deepfake Maker. https://deepfakesweb.com/.
[4] Zaynab Almutairi and Hebah Elgibreen. 2022. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* 15, 5 (2022), 155.
[5] Thomas Brewster. 2021. Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=92c744755917. (Accessed on 02/20/2023).
[6] Nicholas Carlini and Hany Farid. 2020. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 658–659.
[7] Millie Chow. [n. d.]. What Are The Positive Applications of Deepfakes? https://www.jumpstartmag.com/what-are-the-positive-applications-of-deepfakes/.
[8] Matthew Griffin. [n. d.]. Edtech company Udacity uses deepfake tech to create educational videos automatically. https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/.
[9] Karen Hao. 2021. Deepfake porn is ruining women's lives. Now the law may finally ban it. | MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/. (Accessed on 12/15/2022).
[10] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. 2020. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400* (2020).
[11] Michael Kan. [n. d.]. Microsoft's AI Program Can Clone Your Voice From a 3-Second Audio Clip. https://www.pcmag.com/news/microsofts-ai-program-can-clone-your-voice-from-a-3-second-audio-clip.
[12] DAMI LEE. [n. d.]. Deepfake Salvador Dalí takes selfies with museum visitors. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.
[13] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. [n. d.]. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
[14] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. 2021. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. *arXiv preprint arXiv:2107.10394* (2021).
[15] Siwei Lyu. 2020. Deepfake detection: Current challenges and next steps. In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*. IEEE.
[16] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–41.

[17] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020).

[18] Catherine Stupp. [n. d.]. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[19] Jai Vijayan. 2022. Reshaping the Threat Landscape: Deepfake Cyberattacks Are Here. https://www.darkreading.com/threat-intelligence/threat-landscape-deepfake-cyberattacks-are-here. (Accessed on 12/15/2022).

[20] James Vincent. 2022. Binance executive claims scammers made a deepfake of him - The Verge. https://www.theverge.com/2022/8/23/23318053/binance-comms-crypto-chief-deepfake-scam-claim-patrick-hillmann.

[21] Jane Wakefield. 2022. Deepfake presidents used in Russia-Ukraine war - BBC News. https://www.bbc.com/news/technology-60780142. (Accessed on 11/27/2022).

[22] Lior Yasur, Guy Frankovits, Fred M Grabovski, and Yisroel Mirsky. 2023. Deepfake CAPTCHA: A Method for Preventing Fake Calls. *arXiv preprint arXiv:2301.03064* (2023).

[23] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A survey on deepfake video detection. *Iet Biometrics* 10, 6 (2021), 607–624.