Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 277–286
https://doi.org/10.54941/ahfe1005592

Emerging Threat of Deepfakes: Viability, Risks, Impacts and Mitigations Through a Practical Use Case

Levente Nyusti and John Eidar Simensen

Institute for Energy Technology, Halden, Norway

ABSTRACT

Early 2024 an employee was tricked to transfer 25 million dollars in a cyber-attack where live deepfakes were used to convince the employee of the legitimacy of the request (Chen and Magramo, 2024). Deepfakes are media (image, audio, and video) edited by an algorithm. In the case of the malicious deepfake e.g. video input, it enabled real time interaction between a target and the deepfake. The availability of software and hardware enabling anyone to create their own high quality deepfakes has become such a threat that Europol in 2022 stated: "*Many organisations have now begun to see deepfakes as an even bigger potential risk than identity theft (for which deepfakes can also be used), especially now that most interactions have moved online since the COVID-19 pandemic*" (Europol, 2022). This paper presents experiences from creating live deepfakes based on typical online, openly accessible media, using free software and a gaming computer, following a free online guide. Without previous experience with deepfakes, we were able to select settings that yielded high quality deepfakes with less than one hour of exploring. The activity of labelling faces requires very little skill. Based on the use case, we provide examples of the achieved quality and discuss cyber and information security implications for organisations. Interviews were performed with a set of small and medium sized organisations regarding their awareness and preparedness for dealing with deepfakes. Industry start to become aware of potential threats of deepfakes, but lack procedures, processes and awareness to be able to sufficiently mitigate deepfake risks. Finally, we suggest a set of best practices and procedures for identifying, and mitigating such threats, focusing on technology, organisation, and the human element.

Keywords: Cybersecurity, Artificial intelligence, Deepfake, Manipulation, Human influence, Misinformation, Social engineering

INTRODUCTION

There are several cases where deepfakes have been used by malicious actors in order to damage another person's reputation, disrupt democracy or steal money. Deepfakes are defined as media like image, audio or video edited by an algorithm. Early 2024, explicit images spread through social media, believed to be taken of Taylor Swift (Conger and Yoon, 2024). Eventually these images were removed and tagged as generated by Artificial Intelligence (AI), but by that time, these images have been seen over 47 million times. This year, deepfakes have been used to disrupt democracy: an attacker created an

audio deepfake of US president Joe Biden, that was used to call residents "*urging them not to cast ballots*" (Gordon, 2024). It allows malicious actors to perform very sophisticated attacks, which was demonstrated early 2024, where a financial worker was invited to a Teams meeting where several colleagues the employee recognized were present, including the Chief Financial Officer (CFO). During the meeting, this employee was tasked to transfer $25 million, which the employee did. The employee reported that the other attendees in the meeting looked and sounded just like his real colleagues (Chen and Magramo, 2024).

In 2022, Europol released a report on deepfakes and threats, providing malicious examples of deepfakes (Europol, 2022). This list includes activities like "*falsifying online identities and fooling 'know your customer' mechanisms*", "*facilitating document fraud*", "*disrupting financial markets*" and "*perpetrating extortion and fraud*". They also report that many organisations have become more aware of the damages deepfakes can cause, especially as most interactions have moved online since the COVID-19 pandemic.

In this paper, we present the results we achieved while developing a deepfake using an open-source framework, following an open-source guide, using off the shelf products for gamers. The developed use case is shown to three different organisations in a set of interviews addressing deepfake awareness and preparedness. We leverage the presented use case experience creating deepfakes with the state of practice in industry and suggest a set of good practices and activities to support industry getting started properly, as well as reference what we deem are good information repositories for developing an organisations capability to protect against deepfakes.

BACKGROUND

One early paper on algorithm based video manipulation (Bregler, Covell and Slaney, 1997), presented a methodology to edit a video in a way to alter lip movement for a person to align it with an audio file. The authors explained the usefulness of the technology for the purpose of movie dubbing, aligning the lip movement of the actors with the words spoken in different languages.

In (Pantserev, 2020), the author describes how the deepfake technology originally appeared to entertain, but today can be used in psychological warfare. The availability of such deepfake tools make this threat exponentially more dangerous, as now anyone with a laptop and access to the internet could create a deepfake where, e.g., an American general is burning the Koran. As there is high tension between these and other cultures, fake news like this could lead to serious consequences. The author also mentions research into developing detection mechanisms for such deepfakes, but they speculate that these will be outdated and inefficient in the near future.

A systematic literature review was performed in (Rana et al., 2022), reporting on current tools and methodologies used for deepfake detection. The different methodologies were compared, revealing that deep learning-based detectors outperform other solutions. However, the author also states that deepfake detection still faces many challenges.

The author of (Ahmed, 2023) has compared the general public's perception of their ability of detecting deepfakes and their actual ability of detecting deepfakes. The study showed that people generally believe that deepfakes influence others more than themselves, and that their ability of detecting deepfakes is higher than it actually is. However, the study also showed that people who are more frequently exposed to deepfakes are less naïve regarding their ability of detecting deepfakes, and the deepfake's influence effectivity.

USE CASE

We developed a deepfake model, using an open-source framework, called DeepFaceLab (Perov et al., 2021). The framework has several releases tailored for different operative systems. The Windows release is point and click and very beginner-friendly. The Linux version requires more environment setup, but it also provides more transparency as the code is human-readable, whereas in the case of the Windows version, the scripts are compiled to a machine-readable format.

Several guides are available for how to use the framework. The general workflow is the following:

1. Collect content about attacker and target (pictures and/or videos).
2. Extract image frames from video.
3. Extract face from frames. Manual action might be needed if the pictures contain several faces.
4. Data cleanup that includes - removing duplicates, blurry pictures and other unwanted pictures.
5. XSeg mask labelling and model training: This is the part that requires the most effort from the attacker. In this step, the attacker creates masks for both the attacker and the target. This requires some manual work, as these masks need to be precise in order to create good deepfakes.
6. Deepfake model training: Train the actual deepfake model, which then learns how to recreate both faces, then merge these two. This is the most time-expensive part, as this step requires several days of training.
7. Merge faces: In this step, the model trained merges the faces on the frames extracted from the video.
8. Convert from frames back to video.
9. Extract model (Optional): This script allows the attacker to extract the model in a format that is understandable by the live deepfake framework DeepFaceLive (Iperov, 2024).

Our deepfake was created by the first author, without previous knowledge of creating deepfakes, using one of the guides found online. From start to reaching a trained deepfake, it took 17 hours to examine the workflow and execute all the steps 1-9. As mentioned in step 6 however, the time-expensive part was training the deepfake model. It took approximately four weeks to train the model on a gaming PC that cost less than $3000. When upgrading the Graphical Processing Unit (GPU) to an RTX 4090 ($2000), training time was reduced to 2 weeks. The framework allows the use of several GPUs simultaneously, which will reduce the training time further. Our deepfake was

created using only four minutes of recording of the target, and ten seconds of recording of the attacker. Figure 1 shows the deepfake created in three pictures. The picture to the left shows the attacker, the center picture shows the target, and finally the result created by the deepfake model.

Figure 1: Deepfake use case.

The use case demonstrates how any normally technically savvy person, with access to of the shelf gaming computer gear, and openly available media of a target, quite easily can create their own high quality deepfakes. In the following we will explore existing advice and best practice to mitigate the threat of deepfakes, before exploring actual state of awareness and practice in industry.

DEFENDING AGAINST DEEPFAKES – BEST PRACTICES

One effective method to mitigate live deepfakes in an organisational setting is to implement a human equivalent for two-factor authentication (Hadnagy, 2024). One solution for this is to enforce a policy where the employee asked to perform an action that might leak confidential information or cause other type of harm performs a validation with the action responsible. One requirement then is that the validation happens through a trusted form for communication e.g. Microsoft Teams, where physical meeting is not an option. The employee in question then is required to find the correct contact information of the action responsible through the trusted platform. Like in the instance of the "25 million scam" mentioned earlier, the accident could have been avoided if the employee performed this validation with the CFO before completing the transfer.

As also mentioned in (Parliament et al., 2021), deepfake detection poses several challenges, and such technologies used for deepfake detection are limited. Here, the authors also mention technical prevention mechanisms, like adding a noise pattern to images shared, which is indistinguishable to

the human eye. This additional layer upon the original image while unseen for the human eye, confuses AI algorithms, making the media unusable for training. This however is irrelevant if the attacker successfully captures their own media about a target. Finally, the authors mention another preventative method, which is to raise awareness about deepfakes, and their sophistication.

The NSA, FBI & CISA recently created and released an information sheet, discussing deepfakes (NSA, FBI and CISA, 2023). In addition to discussing deepfakes in general and their impact on organisations, the authors also discuss detection and prevention mechanisms in addition to providing recommendations for resisting deepfakes. While the authors give examples for software that can be used to detect such manipulation, it is also mentioned that most technological detection methods are a cat and mouse game, as both the detectors and the deepfakes get better by the day. Finally, the list of recommendations includes implementing liveness tests, protecting public data of high-priority individuals and training personnel, also providing some training resources.

HOW WELL PREPARED ARE ORGANISATIONS AND INDUSTRY FOR DEEPFAKES

In order to better gauge awareness and preparedness in organisations we performed interviews with higher level information security and cyber security roles at several organisations. The interviews were performed as semi-structured interviews where the interviewees were informed about the topical area only prior to the interviews. In order to boost interviewees' willingness to the actual state of affairs, both individual persons and companies are kept anonymous. The goal of the interviews is to gain qualitative insights to state of practice and awareness on deepfakes, in preparation for a larger quantitative study on the topic. The semi-structured interview had a total of 18 guiding questions, developed through an iterative process. The questions cover 4 topical areas, including: understanding of what deepfakes are and what is required to produce good deepfakes, current level of maturity in the organisation with regards to deepfakes, future threat picture and need for knowledge, tools and procedures, and assessment of deepfake created in the use case presented in section Use case. A total of three interviews were performed with two private and one public organisation. The private organisations were one research institute and stock market registered retail company. The public organisation was a Norwegian academic institution. The interview personnel comprised two CISOs and one ICT consultant responsible for the organisation's cyber security. The participants average experience within IT and cybersecurity ranged 13–18 years. Identity of companies and interviewees are kept anonymous to protect the organisations. A standard interview consent form was used. In the following a summary of the results is provided with regards to the four main categories. A full overview of all questions will be provided upon request.

Understanding of Deepfakes

All three participants shared similar knowledge of what constitutes a deepfake and there were some variations on their views on how much material and effort is needed to create good deepfakes. They all reported that tools and how-to instructions are available for free and that not much knowledge is required to create quality deepfakes. A recurring discussion was on the quality of the deepfake with regards to potential effectiveness. There were some variations on the view of material needed and the quality of the material needed, ranging from 1–2 minutes of video of unspecified quality, to 30 minutes of high-quality video.

Organisational Maturity and Preparedness

All organisations report that their current level of maturity in the organisation with regards to deepfakes could be improved. At the same time two interviewees report that there are other threats and vulnerabilities they perceive as more pressing currently. All three agree that this can change quickly and that they need to prepare. Here, awareness-training of the organisation and of central stakeholders and leaders in the organisations were mentioned as the main mitigative activities by all three participants. None mention any procedures or initiative to control or decrease key-role stakeholders' presence on the internet. For preparedness, two organisations report that they have procedures for their economy departments to mitigate deepfakes and frauds. When it comes to the use of technical tools and mechanisms to detect, mitigate, and safeguard against deepfakes, the interviewees all stated they had little to no knowledge of what exists and is available.

FUTURE NEEDS

All interviewees saw deepfakes as a fast-developing threat that would need particular attention, including improved training, automatic detection, and technical protection mechanisms. Automatic detection capabilities were mentioned as a future need. None mentioned protection of digital material through anti-AI filters for example. During the interview, the participants were shown deepfake examples and the deepfake use case developed by the first author. All three asked for links to the material to evaluate for use for training in their own organisations, and one asked if it was possible to have the deepfake developed in the use case for the same purpose. The same interviewee also considered creating deepfakes of their own central employees to have more impactful examples and to gain more knowledge on deepfakes inhouse. All interviewees foresaw the need for more specific procedures for reporting deepfakes, and for describing work processes to better protect against deepfakes.

Use Case Feedback

At the end of the interview the participants were shown a set of deepfake examples from a training program from University of Washington and

Microsoft (Center for an Informed Public at UW, 2020), as well as the deepfake created by the first author. For the Microsoft examples they were asked if they think the pictures and videos are real or deepfakes, and the results varied. When they choose the right answer, they could not provide good reasons for their choices other than their "gut feeling". It is worth noting that in these cases the interviewees were primed that one would be real and one would be fake, something that will not happen in a real case. The interviewees were asked for their feedback on the deepfake created and presented in this paper. Feedback on its quality varied, one of the participants was clear that it would need improvement to fool someone, one stated that it was fairly good, and one was enthusiastically impressed. One shared sentiment was that the quality of the deepfake would depend on how well the target knows the faked person(s) and that the typical treats of the person' faked must be sufficiently covered. Further, the request given by the deepfake must be within what is typical and believable, which would require some social engineering efforts in preparation of the deepfake. When informed about the amount of knowledge, time and technology needed to create the deepfake, the participants were surprised and stated that they (their organisations) are not as ready as they need to be for the deepfake threat.

Other Observations From the Interviews

Across the topics covered and all three interviewees, we found that the provided answers and thinking regarding deepfakes were very much aligned, despite the fact that three different types of organisations were consulted. The interviews confirm what is found in literature regarding the lack of technical detection methods, as well as the need for emphasizing employee awareness training on the topic.

DISCUSSION

As mentioned earlier, deepfakes are not only openly available, but are also under constant development. As this technology gets increasingly better, it is safe to assume that malicious actors will also utilize it even more in the future. As the "25 million scam" shows, deepfakes are already at a point where both video and audio might be indistinguishable compared to the real person, even in a live setting. As such attacks are still in their infancy, sophisticated attackers will be able to engineer such deepfakes and attack scenarios more efficiently in the future. For this reason, it is of high importance to conduct further research on how attacks utilizing deepfakes can be reliably mitigated. Research might focus on how to appropriately educate personnel and what policies are recommended for organisations to implement.

In the interviews all participants made it clear that although they see deepfakes as a severe potential threat, they do not prioritize as much as they would have liked due to having other more pressing threats on their tables. Two stated that for deepfakes to get the same priority as e.g., ransomware, a large successful attack on a company or industry is needed domestically, or at least in Scandinavia. That way the threat comes closer and becomes 'more

real'. When it gets media focus, it becomes a common societal problem, and legislation and requirements from government will follow shortly after.

As interviewee candidates' replies aligned, and they represented three different types of public and private organisations and businesses, we would argue that the answers provided to a large extent represent a crosscut of organisations in Norway when it comes to awareness and preparedness for deepfakes. It is also worth mentioning that this might be due to the fact that no organisation in Norway have experienced a cyber incident involving deepfakes yet. The interviewees also state that this might need to happen before deepfakes are prioritized as a real threat, as prioritizing this would mean down-prioritizing other threats.

One of the interviewees also mentioned that it might be useful for them to create their own deepfake, featuring one of their central personnel. Based on our experience from the deepfake use case created, we believe that the threshold for successfully creating a deepfake without previous knowledge is sufficiently low, and that it could be useful for organisations to do this as it would result in more targeted awareness training material and improve the organisations knowledge and its' capability to safeguard against deepfakes. This would also show the employees how easy it is to create such believable deepfakes, making them more aware of the threats posed by this technology. Experiencing deepfakes of colleagues and closer related persons has a much more profound impact than seeing deepfakes of public figures.

Based on the literature, our experiences from the use case, and the interviews performed, we believe a good start for increasing awareness and developing good practices and procedures if the organisations develop their own deepfake which the employees can more easily relate to, featuring an employee they know. Using this deepfake during awareness training might trigger heightened awareness for employees, compared to deepfake introductory videos that can be found in popular cyber awareness training programs. Once the employees understand the reality behind the potential threats posed by deepfakes, we believe this might urge them to develop their own solutions to this problem. A general mitigation for companies can be to implement the human two-factor mentioned earlier while other reliable mitigation techniques are under development. We also suggest looking into preventative measures, like reducing the media presence where possible, and edit media before uploading it online, to include the invisible noise pattern used to confuse AI.

SUMMARY & CONCLUSION

This paper presented a use case creating a deepfake from typical openly available media, using open access recipes and openly available free tools, on store-bought gaming computer. The experiences gained were brought into three semi-structured interviews with different types of organisations on their awareness and preparedness for deepfakes. Findings from the interviews indicate that although organisations are very aware of the fast-evolving threat of deepfakes, they are not as prepared as they would like for different reasons such as legislative focus, threat prioritization, lack of

relevant training material and examples, and lack of available off-the-shelf tools and mechanisms to safeguard against deepfakes. It could be seen as a paradox that although one is aware of the threat, ability to prioritize against it properly seem difficult. With limited resources, a general industrial lack of cyber-experience personnel, and other more pressing threats craving attention, it is a quite natural expected reality for industry.

Based on the experiences from the deepfake use case, we believe the threshold for most organisations to create their own deep fakes of their own personnel could be useful to increase topical competence, awareness, and provide realistic 'wake-up' examples. Further, we suggested a set of information sources that would be good starting points for building the defensive knowledge and capabilities needed to protect against deepfakes.

FUTURE WORK

As a continuation of this work, we plan to survey what tools and training material are available to bring awareness to the threat posed by deepfakes. Our experience from the use case and the interviews show that one is able to create more awareness when using an example that is relatable to the training recipient. As this creates more awareness, future work might focus on methods and techniques that could be implemented as preventative measures. Once the preventative measures are in place, we suggest also focusing on tools and methods that can be used to detect deepfakes.

ACKNOWLEDGMENT

The authors would like to thank the interview participants for providing valuable information and for sharing their views, state of practice and future needs. We would also like to thank P-A. Jørgensen for volunteering as a deepfake target.

REFERENCES

Ahmed, S. (2023) 'Examining public perception and cognitive biases in the presumed influence of deepfakes threat: empirical evidence of third person perception from three studies', *Asian Journal of Communication*, 33, pp. 1–24. Available at: https://doi.org/10.1080/01292986.2023.2194886.

Bregler, C., Covell, M. and Slaney, M. (1997) 'Video Rewrite: Driving Visual Speech with Audio', in. *Computer Graphics, SIGGRAPH 97 Annual Conference Series*, pp. 353–360. Available at: https://doi.org/10.1145/258734.258880.

Center for an Informed Public at UW (2020) *Spot the Deepfake*. Available at: https://www.spotdeepfakes.org/en-US (Accessed: 28 June 2024).

Chen, H. and Magramo, K. (2024) *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'* | *CNN*. Available at: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html (Accessed: 11 June 2024).

Conger, K. and Yoon, J. (2024) 'Explicit Deepfake Images of Taylor Swift Elude Safeguards and Swamp Social Media', *The New York Times*, 26 January. Available at: https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html (Accessed: 11 June 2024).

Europol (2022) 'Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab'. Publications Office of the European Union, Luxembourg. Available at: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes#downloads (Accessed: 11 June 2024).

Gordon, R. (2024) *3 Questions: What you need to know about audio deepfakes*, *MIT News | Massachusetts Institute of Technology*. Available at: https://news.mit.edu/2024/what-you-need-to-know-audio-deepfakes-0315 (Accessed: 11 June 2024).

Hadnagy, C. (2024) 'Ep. 248 - The SE ETC Series - SE in the News - Tips & Tricks', *Security through education*, 26 February. Available at: https://www.social-engineer.org/podcasts/ep-248-the-se-etc-series-se-in-the-news-tips-tricks/ (Accessed: 20 June 2024).

Iperov (2024) 'iperov/DeepFaceLive'. Available at: https://github.com/iperov/DeepFaceLive (Accessed: 13 June 2024).

NSA, FBI and CISA (2023) *NSA, FBI, and CISA Release Cybersecurity Information Sheet on Deepfake Threats | CISA*. Available at: https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats (Accessed: 11 June 2024).

Pantserev, K. A. (2020) 'The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability', in H. Jahankhani et al. (eds) *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*. Cham: Springer International Publishing, pp. 37–55. Available at: https://doi.org/10.1007/978-3-030-35746-7_3.

Parliament, E. *et al.* (2021) *Tackling deepfakes in European policy*. European Parliament. Available at: https://doi.org/doi/10.2861/325063.

Perov, I. *et al.* (2021) 'DeepFaceLab: Integrated, flexible and extensible face-swapping framework'. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.05535.

Rana, M. S. *et al.* (2022) 'Deepfake Detection: A Systematic Literature Review', *IEEE Access*, 10, pp. 25494–25513. Available at: https://doi.org/10.1109/ACCESS.2022.3154404.