IM0802-232434M - Responsible Artificial Intelligence assignment 1

Deepfakes detection and attribution

Alexander Van Hecke (852631385)

April 6, 2024

Introduction

In this essay we want to describe a dark scenario "Surely this is something I can trust?". In this scenario we discuss Peter, a young man who is just starting to get settled, is in a stable financial situation and has a couple of years of working experience. Peter tries to stay informed about current events and uses diverse sources of information for this. He watches the news on national television regularly, but also checks several social media feeds to see what's happening in the world. The social media companies decide what Peter gets to see on his feed, and they try to maximize his screen time, as more screen time means more advertisements can be shown to Peter. Trying to maximize screen time and selling ads is the business model of the social media companies, and they distribute all kinds of media (text, short stories, videos, etc) to maximize their profits.

One day, Peter gets a video on his feed where a well known financial journalist Michael praises an investment. As Peter has been doing well financially lately, he has some financial reserves and he has already been thinking about investing his money in some way. However, since he is no financial expert, he didn't know how to get started. This seems like an excellent opportunity to Peter, since he knows Michael from national television where he often discusses economy and financial topics on the news. Peter fully trusts Michael, as he always has a nuanced opinion and generally seems very knowledgeable about investments. After looking at the website of the investment, he decides it looks good and invests all his money in this opportunity.

Unfortunately, the video was a deepfake video. Fraudster Kris created this video using freely available software, and was able to use a lot of video and speech from Michael's television appearances. Kris was abusing Michael's good name and financial reputation to lure people in a fake investment. Needless to say, all money invested goes straight to Kris. Michael was not aware of this video, and certainly does not endorse the investment praised in the deepfake video.

This scenario is happening today. Two very recent cases in Belgium illustrate this. A deepfake video of financial journalist Michael Van Droogenbroeck was distributed on social media (1), and news anchor Annelies Van Herck appeared in a deepfake video where she advertises a "get rich quick" game (2). These videos were widely distributed on social media, and lots of people fell for this fraud. Social media companies are aware of this issue (3,4) but don't do enough to prevent this from happening as it goes against their business model. Meanwhile, fraudsters are getting away with this and trick people in their fake investments.

Discussion of the issue

Sources

Deepfake videos are pervasive in many media. Recently, hyper-realistic videos have emerged that depict someone saying and doing things that never happened. Given the reach and speed of social media, convincing deepfakes can quickly reach millions of people and have negative impacts on our society (5). Academic research has focused not only on generative models and generative adverserial networks (GANs) capable of generating deepfake videos, but also on ways to detect deepfake videos by these architectures. A convolutional neural network (CNN) with a classifier network is proposed by the authors of (6) for the detection of deepfake videos. They extract faces from videos and compare these against a face dataset (7)

to detect deepfake videos. Their technique works especially well on videos generated by an autoencoder. The use of transfer learning is examined in (8). Transfer learning in autoencoders and a combination of CNN and Recurrent Neural Network (RNN) models are used to increase detection rates of deepfake videos when these are generated using unseen manipulations and datasets. They conclude fine-tuned models yield better accuracy as compared to models trained from scratch. A survey of different manipulation techniques, public datasets and performance evaluation for several categories of deepfakes is given in (9). Some of the shortcomings of the current detection techniques are described in (10).

Several news anchors of VRT, a public news outlet in Belgium (11), have been the subject of deepfake videos. Michael Van Droogenbroeck, a financial journalist working for VRT, featured in a deepfake video that was distributed on social media. In this video, he praised a fake investment opportunity. Colleague Annelies Van Herck, news anchor at VRT, appeared in another deepfake video where she was depicted as advertising an online "get rich quick" game. Fraudsters were behind these videos, having mainly financial motives. Given the yield of social media and the trust the public places in these news anchors, this has resulted in many people falling for the fraud.

Deepfake videos appear in popular media as well. The BBC TV show "The Capture" (12) centers around the idea of deepfake videos being created to explicitly discredit people. This leads to unjustified charges of kidnapping and even murder. In this TV show, the validity of the deepfake evidence is called in to question by a detective and everything ends well. However, it does illustrate a point : deepfake videos can not only be used for financial gain, but also to discredit and falsely accuse (political) opponents and organisations.

Use of AI in the issue

We argue that in this case, the use of AI is not only causing the problem, but is also **part** of the solution :

- deepfake videos can be generated using generative autoencoders or GANs. Academic research on the topic is mostly freely available, and more importantly the code implementing this research is often freely available as well on platforms like GitHub. This implies these implementations are open to use by anyone, free of charge. Given the increasing volumes of training data available (datasets of video, text, speech, ...), this gives fraudsters powerful tools to generate convincing deepfake videos. Europol has found that for fraudsters lacking the necessary skills, "deepfake as a service" platforms (13) exist.

- AI can be used as part of the solution as well, by devising solutions for detecting deepfake videos. Note this is only part of the solution. A technical solution without the necessary legal framework (regulation and penal laws for abuse) and sufficient awareness in social media companies, will not alleviate the problem. Social media are already aware of the problem (3,4), but are incentivized not to intervene in the distribution of deepfake videos, which are often highly popular and hence, generate a lot of ad revenue.

Stakeholders and consequences

Several different stakeholders can be identified in this scenario.

(i) victims who get tricked into fake investments lose money and give away confidential financial information. Often these victims are pressured to invest more and more money. This can lead to serious financial problems for the people involved, leading to a deteriorating quality of life. **(ii)** The people whose identity is being abused (news anchors in our example). They have to endure reputational damage, and may get criticized by people for something they weren't even aware of. Related to this, **(iii)** the companies affiliated with the identity victims endure reputational damage as well. Their businesses might be impacted. **(iv)** The fraudsters creating and distributing the deepfake videos are clearly stakeholders as well. They can make a lot of money with relatively little effort by abusing people's trust. **(v)** Social media companies have clear financial incentives to distribute popular videos, even if they are deepfake videos. More views yield larger ad revenues. However, they risk facing stricter regulation. **(vi)** Companies advertising on social media platforms are stakeholders as well. They do not want to be associated with these kinds of deepfake videos as they would risk reputational damage. **(vii)** Policy decision makers (e.g. European commission) need to invest time and effort in regulating social media companies, video generation software, and passing penal laws to take legal action against offenders. Finally, **(viii)** resellers specialised in video

manipulation software may not have bad intentions, but may face increased scrutiny nonetheless. This is summarized in Table 1.

Table 1: Stakeholders and consequences

stakeholder	consequence
victim	financial loss, data privacy
identity victim	abuse of identity
companies affiliated with identity victims	reputation damage
fraudster	financial gain
social media companies	financial gain, subject to more regulation
companies advertising on social media	reputation damage
policy decision makers	regulation, legal action
deepfake and special effects software resellers	regulation

Why is this an issue?

Several ethical issues can be identified in this scenario.

- The **(data) privacy** of victims is violated. Sensitive data such as financial and personal details fall into the wrong hands.
- **Technological abuse**. This is an example of a technology that can be used for good and bad, but in this scenario there is abuse for personal financial gain.
- **Do no harm**. This scenario is a fraudulent and harmful use of AI technology. It can lead to financial and general quality of life issues for those involved as victims. People whose identity is abused can suffer from serious reputational damage.
- It is becoming increasingly difficult to see **what is real and what is not**. People not trained in recognizing deepfake videos may have a hard time detecting them.
- There is little **accountability** of fraudsters and social media companies. Social media companies are aware of the issue and try to counter this with technology, but at the same time claim to have little responsibility when it comes to the content distributed on their platforms. Fraudsters should be held accountable for posting harmful deepfake videos.
- **Accountability** of other stakeholders. Should people not falling for the fraud but sharing the deepfake video nonetheless be held accountable as well?
- **Transparency** from social media companies. When trying to come up with solutions to deepfake video abuse, it is important to be transparent about how, when and why these solutions will be implemented. Note these solutions could be using technology (deepfake video detection), but they could also use people as part of the solution (factcheckers).
- **Collaboration** between different stakeholders. Firstly, policy decision makers and social media companies need to agree on how to deal with these kinds of issues, as trying to solve these issues is a complex problem. Secondly, social media companies should share best practices around tackling this issue. This is not happening right now, as there is a lot of competition between the different platforms. When sharing best practices, this must be done using the scientific method, i.e. through peer reviewed papers containing enough details and not through web blogs containing very little detail (3,4).

Note that trying to steal money and personal information from victims is not a new problem, nor is it restricted to the use of AI. People have always been trying to steal from other people. Likewise, people have tried to pretend to be someone they aren't for financial gain. The use of AI however, makes identity abuse easy and convincing. We would argue AI is an accelerator for abuse in this case.

Solution

In our literature search, we have focused on ways to automatically detect deepfake videos (6,8). In theory, this provides a solution for our scenario. If deepfake videos can be detected, they can be banned and these kinds of frauds could be avoided. However, it is not as simple as just being able to detect deepfake videos. We believe that a solution will have to be multi-layered and will need to include aspects of **detection**,

transparency, **collaboration**, **regulation** and **accountability**. We discuss these aspects next, and focus on policy decision makers and social media companies as principal stakeholders.

The **detection** of deepfake videos by social media companies is a prerequisite for being able to ban illegal or unwanted content. Detection could be automatic or manual. Automatic detection is the focus of the studied papers (5,6,8–10), but it has shortcomings. Firstly, automatic detection of deepfake videos will never be 100% accurate. This means some videos that are deepfake videos will not be identified as such. Secondly, prohibiting every deepfake video detected also prohibits all legal, interesting and recreational uses of deepfake videos. After all, technology might be able to detect a video is a deepfake, technology cannot (at this moment in time) determine what the video is about and whether this video could be potentially harmful if distributed. Therefore, manual judgement by factcheckers will remain an important aspect to ensure **fairness** for content creators. After all, not all deepfake video authors have criminal intentions. Meta recently announced (14) that instead of removing AI generated videos, they are looking into labeling videos as being "Made with AI".

Assuming deepfake videos can be properly detected by machines, humans, or a combination of both, social media companies will need to be **transparant** about the criteria used for banning videos. When is a video considered to be harmful? This will need to take into account aspects like comedy and satire as well. Transparency will lead to best practices, and best practices can be implemented by all relevant stakeholders.

Collaboration between different stakeholders is key. Firstly, there will need to be collaboration between the different social media companies. If not all social media companies uphold the same set of criteria, fraudsters can easily abuse this fact by avoiding platforms having strict criteria and choosing platforms having less strict criteria to distribute their content. As of today, we don't see this kind of collaboration, because the different platforms are in a fierce competition against each other. Each platform tries to get as big a share as possible of the global ad revenues market. They are therefore incentivized to keep internal details of their respective platforms secret. Secondly, collaboration between policy decision makers and social media companies is very important. Policy decision makers will need to regulate, but need to take into account the technical complexities of operating social media platforms on a global scale as well. Furthermore, AI technology is advancing at a rapid pace, meaning regulation and the implementation of this regulation will need to be kept up-to-date.

Many steps have already been taken for **regulation**. The recently passed EU AI act does not ban deepfakes, but requires **transparency** from creators. This includes people creating deepfake content, but also people or systems disseminating deepfake content like social media platforms. The EU AI act mentions four categories of AI systems : minimal-risk (can be deployed without additional restrictions), limited-risk (risk of manipulation or deceit, transparency required), high-risk (potential to cause significant harm) and unacceptable risk (prohibited, thinks like general social scoring or subliminal manipulation). Generative AI systems are categorized as limited-risk for now, but should perhaps be categorized as high-risk instead. Europol has studied deepfakes as well (13) and stresses that any regulation should take the perspective of law enforcement into account. Of note here is that the EU AI act regulates the use of AI in Europe, but social media platforms have a global reach. In particular, many of the most prominent social media platforms are located in the US, where generative AI regulation is very active but still in early stages. The same can be said about China (15), which also hosts a number of large social media platforms and is perhaps less transparant about policies than the EU or US. If regulation is not equally strict across the globe, the risk of fraud remains.

Finally, **accountability** is important as well. It encompasses two aspects : **(i)** people or organisations could be called on to answer for actions and decisions they have taken, and **(ii)** being transparant about actions and decisions. We have discussed the transparency aspect for social media platforms before. Unfortunately, we cannot expect fraudsters to be transparent about their actions or intentions. Law enforcement will have to deal with the first aspect.

As a society, we are just getting to know generative AI in all its forms (large language models (LLM), image, audio and video generation, etc). This will have a profound impact, as it will be increasingly difficult to know what content was generated by humans and what content was generated by AI. AI generated content will be pervasive, and will not be limited to social media. We will have to find ways to deal with this new reality in general. Several iniatives have already been taken (13,16,17), but many more will need to follow as AI technology progresses.

Conclusions

In this essay, we have looked at the abuse of deepfake videos by fraudsters for personal financial gain and investigated some technical solutions (5,6,8) for the detection of deepfake videos. Detection accuracy of deepfake videos is good, but we lack the means to automatically judge whether the content of a deepfake video should be considered harmful.

It is our opinion that we cannot rely on automatic methods alone to tackle this issue. A solution will need to be multi-layered, taking into account aspects of **detection**, **transparency**, **collaboration**, **regulation** and **accountability**. The problem is aggravated by the global reach of social media platforms, meaning global consensus on best practices, regulation and law enforcement will have to be reached. Given the proper regulation and transparency, comedy, satire and recreational deepfake videos will remain justified and legitimate uses of this powerful AI technology.

References

1. NWS V. CHECK - Alweer VRT-gezichten misbruikt in valse nieuwsberichten: Zo herken je online oplichting [Internet]. vrtnws.be. [cited 2024 Mar 24]. Available from: https://www.vrt.be/vrtnws/nl/2024/03/08/check-alweer-vrt-gezichten-in-online-scams-misbruikt-hoe-herk/
2. NWS V. CHECK - Opgepast voor valse AI-reportage met VRT NWS-anker Annelies Van Herck over spel waar je geld mee zou verdienen [Internet]. vrtnws.be. [cited 2024 Mar 24]. Available from: https://www.vrt.be/vrtnws/nl/2024/01/16/deepfake-reclame-game/
3. Enforcing Against Manipulated Media [Internet]. Meta. 2020 [cited 2024 Mar 24]. Available from: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/
4. How we're helping creators disclose altered or synthetic content [Internet]. Google. 2024 [cited 2024 Mar 24]. Available from: https://blog.google/intl/en-in/products/platforms/how-were-helping-creators-disclose-altered-or-synthetic-content/
5. Westerlund M. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review. 2019;9(11):40–53.
6. Mitra A, Mohanty SP, Corcoran P, Kougianos E. A Novel Machine Learning based Method for Deepfake Video Detection in Social Media. In: 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS) [Internet]. 2020 [cited 2024 Mar 23]. p. 91–6. Available from: https://ieeexplore.ieee.org/abstract/document/9426086
7. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: Learning to Detect Manipulated Facial Images. In 2019 [cited 2024 Mar 23]. p. 1–1. Available from: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html
8. Suratkar S, Kazi F. Deep Fake Video Detection Using Transfer Learning Approach. Arabian Journal for Science and Engineering [Internet]. 2023 Aug [cited 2024 Mar 23];48(8):9727–37. Available from: https://doi.org/10.1007/s13369-022-07321-3
9. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence [Internet]. 2023 Feb [cited 2024 Apr 3];53(4):3974–4026. Available from: https://doi.org/10.1007/s10489-022-03766-z
10. Mirsky Y, Lee W. The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys [Internet]. 2021 Jan [cited 2024 Apr 3];54(1):7:1–41. Available from: https://doi.org/10.1145/3425780
11. NWS V. VRT NWS: nieuws [Internet]. VRTNWS. 2024 [cited 2024 Mar 24]. Available from: https://www.vrt.be/vrtnws/nl/
12. *The Capture* (TV series) [Internet]. Wikipedia. 2024 [cited 2024 Mar 24]. Available from: https://en.wikipedia.org/w/index.php?title=The_Capture_(TV_series)&oldid=1213822347
13. European Union Agency for Law Enforcement Cooperation. Facing reality?: Law enforcement and the challenge of deepfakes : An observatory report from the Europol innovation lab. [Internet]. LU: Publications Office; 2024 [cited 2024 Mar 24]. Available from: https://data.europa.eu/doi/10.2813/158794

14. Belanger A. Meta relaxes “incoherent” policy requiring removal of AI videos [Internet]. Ars Technica. 2024 [cited 2024 Apr 6]. Available from: https://arstechnica.com/tech-policy/2024/04/meta-relaxes-incoherent-policy-requiring-removal-of-ai-videos/

15. Sheehan M. China’s AI Regulations and How They Get Made [Internet]. Carnegie Endowment for International Peace. [cited 2024 Mar 24]. Available from: https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117

16. Rachel M. Tackling deepfakes in European policy.

17. EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act [Internet]. [cited 2024 Mar 24]. Available from: https://artificialintelligenceact.eu/