IM0802-232434M - Responsible Artificial Intelligence assignment 1

Deepfakes detection and attribution

Alexander Van Hecke (852631385)

March 23, 2024

The aim of this assignment is to write a critical essay on how AI is applied in a specific domain/problem which creates certain consequences for specific stakeholders. In this assignment you will either select one of the topics suggested to you or you will define a topic by yourself. A list with suggested topics is provided to you on yOUlearn in the Assignment 1 section. Further, you will search the following resources: - two recent scientific articles outside the resources used in the course for the chosen context, - one news/blog article that describes the topic selected, and - a movie/series/book/etc. that includes relevant aspects to the topic selected. The scientific resources are used to capture the scientific perspectives and dimensions of the topic studied, while the non-scientific resources are used to capture the social and cultural perspectives and dimensions of the topic selected and position this topic in the ongoing societal discourses. Afterwards, write a small paragraph describing the topic selected and resources you plan to use and send them to the teacher for verification. In case of unclarities, this process can repeat until it is clear. In case the topic and approach are clear, the student can send a document of max. one page with the outline of this assignment for verification and feedback from the teacher. Write a critical essay of around 2500 words (exclusive references) on how the AI method/application produces certain consequences in the context selected to certain stakeholders following the requirements presented below.

Introduction

introduce the aim and motivation of the essay considering the following setting: Introduce the aim of the essay using a small ethical dilemma/scenario as a fictional setting that illustrates the context. You can consider this an instantiation of consequences created by applying AI in a certain domain or problem. The purpose of this ethical dilemma/scenario is to capture the interest and attention of the reader and state the need to tackle this topic. For this include the following scenario elements: setting, action, and impact. On this behalf, you can use the article by Wright et al. (2014) describing ethical dilemma scenarios available here: Wright, D., Finn, R., Gellert, R., Gutwirth, S., Schütz, P., Friedewald, M., ... & Mordini, E. (2014). Ethical dilemma scenarios and emerging technologies. Technological Forecasting and Social Change, 87, 325-336.

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 that 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 the screen time of Peter, 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 knowlegdeable about investments. After looking at the website of the investment, he decides it looks good and invest all his money in this opportunity.

Unfortunately, the video was a deepfake video. Scammer Kris created this video using freely available software, and was able to use a lot of video and speech from Michael's television appearances. Bert 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 these scams. 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, scammers are getting away with this and trick people in their fake investments.

Issue

> summarize the content of the four resources used and define the issue discussing the following elements: • Discuss how and where AI is applied to tackle the chosen problem/domain. • Discuss why and how consequences of AI application are created reflecting on who are the stakeholders experiencing/impacted by these consequences. • Explain why applying AI in this setting can be considered an issue or a challenge from an ethical perspective and argue if this represents a new or an old (well-known) problem in this domain. At the same time, reflect if this issue appears only due to the application of AI or also other (societal) factors are involved or take part in this process.

Sources

Deepfake videos are pervasive in the media. Recently, hyper-realistic videos have emerged that depict someone say and do 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 yield better accuracy as compared to models trained from scratch.

Several news anchors of VRT.nws, a public news outlet in Belgium (9), 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. Scammers 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 scam.

Deepfake videos appear in popular media as well. The BBC TV show "The Capture" (10) 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 data available (datasets of video, text, speech, ...), this gives scammers powerful tools to generate convincing deepfake videos.
- 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 :

- victims who get tricked into fake investments and lose money. 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.
- 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.
- scammers creating and distributing the deepfake videos. They can make a lot of money with relatively little effort by abusing people's trust.
- 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.
- companies affiliated with those whose identity was abused. They endure reputation damage as well.
- governments who need to invest time and effort in **i)** regulating social media companies, video generation software, and **ii)** penal laws to punish offenders.
- 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, abuse of identity
scammer	financial gain
social media companies	financial gain, subject to more regulation
companies affiliated with victims	reputation damage
governmental bodies	regulation, punishment
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. Financial information like credit card details fall into the wrong hands. Other personal information may be given as well.
- **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.

TODO is this new or old problem?

Solution

consider as a starting point (i) either existing solutions to tackle the identified issue and propose ways to enhance them or (ii) propose your own solution(s) to tackle the identified issue. In both cases, include the following aspects and relate them to the influence on impacted stakeholders: • Ethics: discuss ethical aspects and frameworks involved in building and applying the existing/proposed solution(s). On this behalf, you can use the article by Mittelstadt et al. (2016) describing debates on ethical aspects surrounding the application of AI algorithms, available here: Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679. Note that you can use this recommended article, or another article found by you, to discuss the ethical aspects. Do not forget to mention it in the references of the assignment. • Regulation: discuss how this issue could be regulated or how is already regulated considering existing applied regulation frameworks. • Society: discuss the societal perception and mechanisms of tackling this issue. You can think here of mechanisms such as governmental programs, NGO initiatives, dedicated standards (e.g., IEEE and ISO) etc. • Personal reflection: provide your own perspective on the identified issue and existing/found solutions to tackle it.

Conclusions

provide concluding remarks for the issue identified and existing/applied solution using the four resources found by yourself and relate these findings to your personal assessment following the structure below: • Discuss one or two concluding remarks. • Discuss the findings on the existing/proposed solution considered to tackle the issue identified. • Provide an answer and own reflection to the following question: In your opinion, is the existing/proposed solution suitable to tackle the identified issue or should it be considered a total change in approaching it?

References

1. NWS V. CHECK - Alweer VRT-gezichten misbruikt in valse nieuwsberichten: Zo herken je online oplichting [Internet]. vrtnws.be. 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. Available from: https://www.vrt.be/vrtnws/nl/2024/01/16/deepfake-reclame-game/
3. Enforcing Against Manipulated Media [Internet]. Meta. 2020. 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. 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. NWS V. VRT NWS: nieuws [Internet]. VRTNWS. 2024. Available from: https://www.vrt.be/vrtnws/nl/

10. *The Capture* (TV series) [Internet]. Wikipedia. 2024. Available from: https://en.wikipedia.org/w/index.php?title=The_Capture_(TV_series)&oldid=1213822347