

Dangers of deepfake technology

Erencan Tatar - 12681598

May 30, 2022

Four weeks after Russia invaded Ukraine, Sky News released a video of president Volodymyr Zelensky. Not unusual, but this time Ukraine's president supposedly appeared declaring he decided to return to the Donbas region and advised civilians to lay down arms before the Russian military and return to their families. This first appearance of deep fakes in any armed conflict was removed after already hundreds of thousands of people had seen it on social media and on the national tv channel Ukraine-24. Other forms of misinformation and propaganda had already been used during this war like spam bots and fake news articles, however deepfakes have a greater impact since they directly tackle the visual and auditory senses humans are usually naively inclined to believe. The ability for anyone to produce realistic-looking and sounding videos of people undertaking acts they never did in reality brings opportunities for deception. This essay firstly covers the advantages and disadvantages of deepfakes, which will be followed by the ethical aspect deepfakes have on identity. At last, the implication on the individual and one's view on reality will be covered, to ultimately answer the question:

Do the moral and ethical reasons against deepfakes outweigh the benefits such that restriction is necessary?

I claim the creation should be monitored such that responsible synthetic data creation is evoked and that access to the technology is limited. The applications for deepfakes to be used in information warfare, blackmail, sabotage, and ideological influencing are the main factors for restriction of the technology.

The technology

Deepfake technology lets people produce synthesized, nearly indistinguishable, videos of people and thereby create events that never occurred. A useful example of applying deepfake technology is in the media and film industry, where it can fill the role of CGI to recreate the likeness and facial expressions of the actors. In other scenarios, characters need to be older or younger which can also be done. The same technology can also be used to cross language barriers, where characteristics of someone's voice can be extracted to make them talk in different languages. Take for example the David Beckham Malaria Campaign, where it was shown he spoke in nine different languages in order to reach a greater audience. This can again be used in the film sector where movies can be translated with the voice of the original actors. A different use case is the story of the Canadian startup Lyrebird which uses generative AI to create personalized synthetic voices for those who have lost the ability to speak due to illness. In other words, deepfake technology can shatter language barriers, make video content more creative, and also give people back their own lost voice.

Conversely, deepfakes present plenty of opportunities for malicious use, misinformation, and fraud. The hype of deepfake technology only really started after its use in deepfake pornography using celebrity pictures in late 2017. The open-source technology can be used as blackmail by anyone with enough training images. Research by Deeptrace (2019) found that at the time 96% of the deepfakes videos online contain non-consensual deepfake pornography. Deepfakes can also be used to sway elections or cause international conflict by using misinformation coming from what is assumed to be trustworthy sources. One recent study showed that exposure to a deepfake depicting a political figure significantly worsened participants' attitudes toward that politician (Dobber, Metoui, Trilling, Helberger, & de Vreese, 2021). The fact that social media content can be targeted to specific demographics was exploited by this study to show that this effect is amplified when microtargeting the deepfake to groups most likely to be offended. Concern about deepfakes rises when criminals find ways to use the technology for real-time impersonation or voice-spoofing which enables them for example to imitate a CEO in order to make illegitimate transactions. We can not assume that the person we see in a video is who we assume them to be, even if we seem to be interacting with them in real time. This brings forth the question of how to deal with the highest form of affirmation in court, namely visual evidence. To combat this various types of digital evidence have been established as trustworthy evidence to counter this such as electronic seals, time stamps, and electronic signatures (Mason, 2016).

Ethics of identity

Deepfakes challenge the definition of the human face, which is both the reason they are valuable from an entertainment standpoint, but also the reason why they are dangerous. One can justify the use of the technology by arguing that it is a technological driven tool but they however have the potential to be used as a weapon through misinformation. Furthermore, there are edge cases for which it is not directly clear in which category they fall. There are deepfake cases for which there is a fairly blurred line between the positive and negative sides, such that the term ‘greyfakes’ had been introduced for it. An example of such a greyfake is the use of synthetic resurrection in the entertainment sector. CGI combined with the power of AI may lead to even higher degrees of realism, which can result in a case where a deceased actor’s synthetic twin can continue in place of them on screen. Disney is already fighting for legislation that approves the use of deepfakes and synthetic media in New York, making the case that such restrictions on the technology could stifle creative freedoms. Yet the ethical debate on synthetic resurrection does not provide a clear outcome. The possibility of a synthetic afterlife could see figures such as Tupac, Micheal Jackson, or JFK becoming puppets for companies, and making them do and say basically anything they want, risking destroying their reputation even after death. It is not illegal for people to portray themselves as a different individual in the case of a lookalike, twins, or people who used plastic surgery to make their physical appearance resemble that of someone else’s. Even stronger, they can say and claim anything as long as there is no portrait right infringement, which can not be the case for the ordinary person since most victims of deepfakes do not own a copyright interest in their own image. This results in the fact that we own our digital faces only to a certain extent. The difference between creating truly new unique digital faces and using an existing person’s face can be measured using the face information used in the technology already. So the case that someone’s status or position might be at risk due to your deep fake lookalike videos is easy to make, and thus regulating in these cases would be manageable. However, without good restrictions and laws, we become slaves to the technology and go from humans to just data blurring the line between human and synthetic faces.

Implications on society

A good starting point for understanding the implication of deepfakes on individuals is immersive virtual reality (VR). From this field, we learn that compared with watching scenes of another person, watching your own doppelganger causes encoding of false memories in which participants believe they actually performed the deep fake activity (Segovia & Bailenson, 2009), more exercise behavior after watching a positive health outcome (Fox & Bailenson, 2009), and brand preference for products used by the virtual self in the deepfake (Ahn & Bailenson, 2014). These studies show that even with relatively cartoonish animated

doppelgangers, people still could be motivated or influenced by watching itself succeed (in VR). A study by (Wu, Ma, & Zhang, 2021) went even further with altering self-perception and showed, contrary to intuitive predictions, positive effects on women viewing themselves within a blended deepfake celebrity image. On the contrary, deepfakes can have a devastating impact on the self-identity of a victim's life in the case of nonconsensual deepfake pornography where they could be used to extort, humiliate, or harass. In consequence consumers of that content will view the victims as plain objects, promoting this content, and in turn harming their own image of reality. Deepfakes are not the only way to cause doubting among people but the impact of deception by deepfake has the potential to be greater than that of verbal deception because of the primacy of visual communication for human cognition (Posner, Nissen, & Klein, 1976), (Koppen & Spence, 2007).

Deepfakes deceive people into believing online content, but at the same time, it initialized a spark among individuals that online information can be altered and that we should be doubtful about the information we trust. The philosopher Rene Descartes made a case for doubting, which he thought was the beginning of knowledge and the end of false belief. Another philosopher by the name of David Hume realized that our beliefs inevitably rely on the beliefs of others. How can we ensure that the reality we perceive matches beyond our reach, in the past and in the future? Hume argued that the only way was by delegating our beliefs to trusted sources. Doubting is still a powerful instrument in order to guard beliefs, however one can only doubt what one can think. The same holds for the people and organizations that the general public label as trustworthy, they in turn deal with the same dilemma of who and what to trust. This leads to the fact that we can not let people blindly create synthetic data and let society or organizations figure out what is real and fake. This will only create disruption in the media, trust in the media, and our own trust and beliefs.

Conclusion

Like with every new technology there are downsides and it is our responsibility to choose how to interact with it. At its core, deepfake is deception with synthesized data in order to intentionally mislead others. Although having positive aspects, the far more significant negative impact the technology can have on self-image, the view of others, and the decreasing trust in online content and other new knowledge makes it dangerous for society. Provided that the output is not made for malevolent ends and does not misuse source material, the claim that it is morally acceptable for people to use the technology can be made. However, we can not guarantee that people behave this way and therefore need licensing and responsibility, and thus complete restriction for the general public is necessary. Meanwhile, we know that misinformation will reach us nevertheless, may it be via deepfakes or any other new medium. It is important to recall that humans over time had to adapt to new forms of deception (Reeves & Nass, 1996). For example, the effectiveness of email spam decreases when awareness

among people develops (Wang, Irani, & Pu, 2013).

It is best to teach the public how to view and process online data, but the issue that may arise is that it is assumed that video and audio data of people are trustworthy-, when in fact we should understand that it is the information source that should be trusted. Teaching the public to be critical and restricting access are the best actions to take.

References

- Ahn, S. J. G., & Bailenson, J. (2014). Self-endorsed advertisements: When the self persuades the self. *Journal of Marketing Theory and Practice*, 22(2), 135.
- Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1), 69–91.
- Fox, J., & Bailenson, J. N. (2009). Virtual self-modeling: The effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychology*, 12(1), 1–25.
- Koppen, C., & Spence, C. (2007). Seeing the light: exploring the colavita visual dominance effect. *Experimental brain research*, 180(4), 737–754.
- Mason, S. (2016). *Electronic signatures in law*. University of London Press.
- Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: an information-processing account of its origins and significance. *Psychological review*, 83(2), 157.
- Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10, 236605.
- Segovia, K. Y., & Bailenson, J. N. (2009). Virtually true: Children’s acquisition of false memories in virtual reality. *Media Psychology*, 12(4), 371–393.
- Wang, D., Irani, D., & Pu, C. (2013). A study on evolution of email spam over fifteen years. In *9th ieee international conference on collaborative computing: Networking, applications and worksharing* (pp. 1–10).
- Wu, F., Ma, Y., & Zhang, Z. (2021). “i found a more attractive deepfaked self”: The self-enhancement effect in deepfake video exposure. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 173–181.