POLITECNICO
MILANO 1863

Can we believe our eyes?

Analysis on deepfakes: risks, benefits and strategies

Giovanni Ploner (ID: 970997)
MSc in Automation and Control Engineering
Ethics for Technology (5 CFU)

September 1, 2022

Abstract

In the present work, it is addressed an ethical study and a cost and benefits analysis on the topic of "deepfakes". The first section introduces the subject, giving some technical details; the two following sections are dedicated to show the threats and upsides that come after deepfakes; then possible countermeasures are presented along reasons to defend the benefits of such technology without underestimate the risks. Finally, before the conclusive resume, it is also presented an alternative "kantian" perspective and strategy on the same matter

Keywords: *Deepfakes, Artificial Intelligence, Privacy, Legislations, Education*

1 What is a "deepfake"?

"But these men were the first and they will remain the foremost in our hearts. For every human being who looks up at the moon in the nights to come, will know that there is some corner of another world that is forever mankind". With the above mentioned words President Richard M. Nixon concluded his nation address after the tragedy of Apollo 11 mission in 1969.
Nevertheless, this speech was never really pronounced, since no disaster happened during the first human landing on the moon: in fact, this is a clear example of a "deepfake" (*portmanteau* of "deep learning" and "fake") video, realized by a group of researchers from Massachusetts Institute of Technology [1]
Deepfakes can be defined as synthetic videos (or audio) in which a person's gestures and voice are mimicked and replaced with someone else's likeness. Although manipulation of people's images and identity is a long known issue - a clear example is the enormous amount of forged lithographies of President Lincoln immediately after his death - deepfakes are a quite recent case, having drawn public attention thanks to image reconstruction developments at the beginnings of the last decade.
Nowadays, the algorithms used to created deepfake material are two: the "oldest" one that make use of particular neural networks called *autoencoders*, that are used also in common smartphone applications, and the more advances one with the Generative Adversarial Networks, developed by the American computer scientist Ian Goodfellow [2]. Goodfellow's researches led to a significant quality improvement of deepfake material, making the identification of manipulated videos even more challenging.
It is worth noting that the term "deepfake" has its roots in the nickname of a Reddit (one of the most popular web forums) user, who exploited AI techniques in order to create and publish on the social media fake pornographic material of famous actresses and celebrities in 2017, superimposing their faces on preexisting videos.
This hideous semantic origin leads us directly to the ethical discussion of this technology's moral ambiguity, along with a cost & benefits analysis aimed to show how humanity may actually profit from it and counterbalance its threats.

2 A subtle menace

In the present section, we look through the dangers and threats that might come with deepfakes' spread.
First, remembering the story behind their name, it is important to underline that deepfakes are largely used to manipulate pornographic contents. A study performed by the Italian-founded company DeepTrace Technologies [3] shows that 96% of deepfakes are pornographic videos, with more than 135 millions of views on pornographic websites. Main victims of these fakes are women, both celebrities and people outside the public spotlight, who see their reputation ruined; on that topic, it is

quite illustrative the circulation of fake sexual videos of Indian journalist Rana Ayyub after a BBC interview in which she expressed her critical views on India's ruling party [4].
However, despite the clear level of violence and violation of dignity of these fake videos, the creation of such material is not unanimously condemned: the controversial decision of the pornographic film studio Naughty America to realize and sell custom sexual deepfakes to their premium users caused a stir among the sector's industry.[5]. Indeed, such position can be morally defended resorting to the so-called *pervert's dilemma* and the levels of abstractions method formulated by Carl Öhman [6]: according to a possible interpretation of the Swedish professor's conundrum, it can be argued that if sexual deepfakes are to be condemned *in toto* than private fantasies also deserve the same treatment. However, I beg to differ and underline that, *de facto*, pornographic deepfakes are utterly harmful and too risky to be compared to private fantasies, even if created just for personal use.
Deepfake riskiness lies not only in videos but also audio plays a major role, especially on the topic of frauds. On this regard, two recent scam cases, first one within a British energy company [7] while the latter in a U.A.E. based enterprise [8], didn't pass unnoticed to the public. In both situation, artificial intelligence was used to simulate other people voice misleading the victims through fake audio messages and phone calls and demanding fraudulent transfer of money. Moreover, this kind of manipulation further worsen the social problem of online scams, whose reports, just in the U.S., have more than doubled between 2014 and 2021, according to a Federal Trade Commission's study [9].
Another, non-secondary issue is strictly related with the instrumental use of deepfakes video for political purposes: the creation and spread of such manipulation not only can discredit public officers internationally, but also undermine democratic institutions themselves, by worsening social and civil clashes. On this regard, it is worth underlining the attempt to damage the public figure of then-candidate Emanuel Macron during 2017 French presidential electoral campaign few days before the election day by spreading fake videos of him admitting to be corrupted. Without the existence of a clear French media law [10], along with the relatively not interesting content of the material, the malicious effort could have succeeded in overturning the election's results, as presented in a 2019 American analysis [11]. Thereafter, this synthetic videos may realistically also represent a cyber-war weapon to deceive public opinion of countries seen as "enemies".
Conversely, it cannot be neglected that such deepfakes are a convenient instrument for mischievous politicians to exploit the "liar's dividend". Borrowing K. Schiff and N. Bueno's words this "dividend" can be defined as "false allegations of fake news, whereby politicians or other public figures claim that real news stories are 'fake news'" [12], and thus it is a valid shelter from press criticism and questions.
Hence, generalizing, the main threat represented by deepfakes is of epistemic nature, i.e. it weakens our ability to distinguish reality from fakes, making video (and audio) documents that we see everyday less reliable and informative, as clearly described in Fallis' analysis [13]. Moreover, the rapid spread that a deepfake can have thanks to social media makes it even more dangerous than classical counterfeit documents.
Nevertheless, ethics of deepfakes is much more shaded than it might appear, as will be discussed in the following section.

3 Potential upsides

Despite the numerous risks which deepfakes may expose us to, society can benefit from this kind of technology.
Covid-19 pandemic has forced almost any company to improve their online facilities and

adapt themselves to smart working and it is in such context that synthetic material became useful: in fact, many enterprises started using videos created with Machine Learning as training tools for their employees; on this topic, the American startup Synthesia has become famous for focusing its value proposition in producing this kind of tutorship [14]. Virtual teaching avatars made with deepfake techniques, indeed, can provide the wherewithal to follow lessons remotely and in any language, without even using a camera.
Another important impact that this new Deep learning might have is the one on the movie and tv-shows industries: thanks to the face-swap technique an entire movie can be shot without even the physical presence of the actor (as done with Peter Cushing while shooting *Rogue One: A Star Wars Story* in 2016 [15]) so that the budget movie making is considerably lower and even deceased actors can "reappear" on screen.
However, the benefits are not just for moviemakers but also for the public: AI-based company Flawless made its core business on refining synchronisation between actor's lips movements and dubbing with neural networks and machine learning techniques [16]. The drastically improved visualization allows international fans to enjoy the movie with more quality and entertainment.
Regarding entertainment, it is worth mentioning that also satire, especially the political one, benefits from deepfakes: in fact, this kind of comedy strongly relies on mimicking and impersonating famous figures, making the performance more realistic and strengthening the satirical effect. Comical deepfakes are already a standard in the U.S., such as the comedy show *Sassy Justice* [17], which also educate the public on this technology; nevertheless, they are rapidly spreading also in Europe and Italy for the same purpose[18].
Artifical Intelligence, furthermore, can deeply sustain the health system and strengthen fight against diseases. Artificial voice cloning through neural networks has allowed many people affected by A.S.L (*alias* Motor Neurone Disease) to "speak" again with their voices synthesized thanks to a speech-generating device [19].
Surprisingly, deepfakes can also be a good instrument to protect someone's privacy: they have been used to protect the identity of the interviewed during the shooting of *Welcome to Chechnya*, documentary on the 2010 anti-LGBTQ purges in southern Russia [20]. Furthermore, as proposed by a group of Chinese medical researchers, face-swapping technology might be used to promote medical video data sharing without violating patient's privacy [21].
Hence, as described in this section, despite carring many risks, deepfake technology is not to be considered as morally wrong *per se*, and possible counteractions to limit its damage exist. The next paragraph addresses them describing a "consequentialist" strategy

4 A consequentialist strategy

Having shown the costs and benefits of deepfakes, a consequentialist judgement on the matter would label them certainly as a possible threat but, at the same time, as potential sources of social upsides, which cannot be ignored and cannot lead to judge them as totally unethical.
Nevertheless, a utilitarian pursuit of pleasure (i.e. benefit) maximization requires countermeasures to erase, or at least minimize, the security costs that come with this synthetic AI creations.
The first line of action has to be the technological response, driven by research and tech companies with the aid of governemnts: in fact, the improvement of deepfake detection algorithms can become a discriminating factor to immediately recognize and label some documents as manipulated, thus weakening, or even solving,

the epistemic dilemma described in section 2. Many examples of neural-network-based detectors are being used or developed in these years, even with the support of United Nations, as for the case of *FakeSniff* [22].
Another crucial point, but in a completely different context, is education; technology is rapidly changing the international and national societies and an education system not able to keep up with innovation will unlikely give its pupils the necessary skills to orient themselves in such world. That doesn't mean only raising in quantity and quality technical subjects and study programs but also teaching how to properly inform themselves, how to distinguish opinion from facts and verify them, how to avoid impulsive judgment and rely more on deep and reasoned research. The words, indeed, of Silbey and Hartzog resume distinctly the point: "deep fakes may be annoying — they might even be amusing — but they will not be quite as disruptive to our children's hopeful future if education trains them to be curious, collaborative, skeptical, and productive" [23].
Eventually, also lawmakers have to take part in the fight against dangerous deepfakes. Already many countries, such as U.S., Singapore and the European Union's states, adopted in the recent years laws that strictly punish fake material diffusion and identity appropriation. Rules to protect privacy from fake pornographic material ought to be fostered, allowing a total removal of such contents; verily, a model that could be followed as example may be Art.17 of "Eu General Data Protection Regulation" (*alias* GDPR) [24]. Moreover, celebrities and public officers' image rights has to be defended against unauthorized use.
However, laws, even the most draconian ones, do not have the ability to completely prevent deepfake's phenomena and deter the creators from making and spreading counterfeit material, and lawmakers ought to make a thorough work avoiding approving bills that limit or erase the potential benefits of AI-aided video manipulation. Hence, the first two fields of action remain the core point of a reliable counteraction to deepfakes.

5 Possible deontological objections

Judging deepfakes, someone with a more deontological perspective may object the analysis here described and advocate for more harsh restrictions of the phenomenon, evaluating it as completely unethical.
A modern-day Immanuel Kant, epitome of duty ethics, certainly would condemn *in toto* the deep-learning-aided manipulation of videos: in fact, since they are counterfeit documents, they ought to be seen as lies, which are strictly judged as immoral. Moreover, as interestingly noted in de Ruiter's work [25], the very same deepfake's existence is a violation of Kant's categorical imperative "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means. " [26]: in fact, creating a deepfake means using a person's appearance and identity as a mean, both in case of mischievous ends (i.e. pornography, frauds, etc.) and for pure entertainment.
In addition to this point, from a deontological standpoint surely deepfakes have to be banned by every country as actual weapon of (cyber)war: not doing so would violate Kant's articles to ensure a perpetual peace [27].
Nevertheless, a complete ban of manipulated videos *ex ante* is intricate to implement, and more realistically a true duty ethics follower would support a counteraction based on prevention. Possible measures to reach this goal may be a reinforcement of U.N. (seen as the kantian "federation of free states" [27]) cybersecurity office in order to have a global action against the phenomenon and also more strict regulation on social media companies' accountability for what it is published on their plat-

form.
Despite the soundness of these arguments, they remain of difficult application, especially on a global scale, and they might have a lower effective impact with respect to the proposals described in section 4.

6 Conclusions

In this article a costs and benefits analysis on the subject of deepfake has been performed sustaining the position that it cannot be judged as an unethical phenomenon itself. In fact, despite the presence of clear threats both on an individual and social scale, AI-aided manipulation can become an instrument profitable for the community by improving aspects such health, work condition and even privacy protection.
Indeed, measures to properly regulate and control its most critical aspects ought to be taken and they have to act significantly on advanced research education and legislation. Clearly, The strategy proposed in this paper is far from exhaustive and further work and development on the matter has to be performed.
However, all useful and disruptive inventions, from the steam engine up to the internet, have shown us that no impactful change can ever occur without any social cost, but, at the same time, it is not possible to recover from them ignoring the risks nor rejecting innovation.

References

[1] moondisaster.org, last access Aug.2022

[2] Goodfellow, I. et al. (2014) Generative adversarial nets. *Advances in neural information processing systems*, 27. DOI: doi.org/10.1145/3422622

[3] Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. (2019) *The State of Deepfakes: Landscape, Threats, and Impact.* Available at regmedia.co.uk

[4] Ayyub, R. (2018), 'I was a Victim of a Deepfake Porn Plot Intended to Silence me', *Huffington Post*, November 21, www.huffingtonpost.co.uk, last access Aug.2022

[5] Levesley, D. (2019), 'Bespoke "deepfake" porn means you can put your face on your favourite pornstar', *GQ*, August 23,www.gq-magazine.co.uk, last access Aug.2022

[6] Öhman, C. (2020). Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography *Ethics and Information Technology* 22,pp. 133-140, DOI: doi.org/10.1007/s10676-019-09522-1

[7] Supp, C. (2019), 'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case', *The Wall Street Journal*, August 30, www.wsj.com last access Aug.2022.

[8] Brewster, T. (2021), 'Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find', *Forbes*, October 14, www.forbes.com last access Aug.2022.

[9] Federal Trade Commission, (2021), *Consumer Sentinel Network: Data Book 2021*, www.ftc.gov, last access Aug.2022.

[10] Caruso Cabrera, M. (2017). 'In France, strict election laws mean there's near silence on massive campaign hack', *CNBC*, May 6, www.cnbc.com, last Access Aug.2022

[11] Chesney, R. & Citron, D. (2019). 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics' *Foreign Affairs*, January/February, Available at alinstitute.org

[12] Schiff, K.J., Schiff, D. & Bueno, N.S. (2020), The Liar's Dividend: How Deepfakes and Fake News Affect Politician

Support and Trust in Media, *2020 APPAM Fall Research Conference*, Available at osf.io.

[13] Fallis, D. (2021), The Epistemic Threat of Deepfakes, *Philosophy & Technology* 34, 623–643. DOI: doi.org/10.1007/s13347-020-00419-2

[14] Simonite, T. (2020), 'Deepfakes Are Becoming the Hot New Corporate Training Tool', *WIRED*, July 7, www.wired.com, last access Aug.2022

[15] Suciu, P. (2020), 'Deepfake Star Wars Videos Portent Ways The Technology Could Be Employed For Good And Bad', *Forbes*, December 11, www.forbes.com, last access Aug.2022

[16] www.flawlessai.com, last access Aug.2022

[17] Itzkoff, D. (2020), 'The "South Park" Guys Break Down Their Viral Deepfake Video', *New York Times*, October 29, www.nytimes.com, last access Aug.2022

[18] (2019), 'Il video "deepfake" di Matteo Renzi trasmesso da "Striscia la notizia"',*il Post*, September 24,www.ilpost.it, last access Aug.2022

[19] www.projectrevoice.org, last access Aug.2022

[20] RD. (2020). '"Welcome to Chechnya" uses deepfake technology to protect its subjects', *The Economist*, July 9, www.economist.com, last access Aug.2022.

[21] Zhu, B., Fang, H., Sui, Y. & Li L. (2020). Deepfakes for Medical Video De-Identification, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ACM, DOI: doi.org/10.1145/3375627.3375849

[22] Casale, P., Osin, V., Raguckaja, G. & Violatto, G. (2020). Collective human action against deepfakes, *Freedom from Fear, Vol. 2018, Iss. 15*, January 08, p. 88-91, DOI: doi.org/10.18356/ae65ce54-en

[23] Silbey, J.M. & Hartzog, W., (2019), The Upside of Deep Fakes. *Maryland Law Review, vol.78*, Northeastern University School of Law Research Paper No. 356-2019, p.963, Available at scholarship.law.bu.edu

[24] European Commission (2016). *Eu General Data Protection Regulation*, Art.17, April 26, Available at iapp.org.

[25] de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy of Technology* 34, p 1311–1332. DOI: doi.org/10.1007/s13347-021-00459-2

[26] Kant, I. (1785). *Groundwork of the Metaphysic of Morals*. Translated by Gregor, M. (1997). Cambridge press. p.38, Available at cpb-us-w2.wpmucdn.com

[27] Kant, I. (1795). *Perpetual peace: a philosophical essay*, Translated by Campbell Smith, M. (1903), George Allen & Unwin LTD. Available at oll-resources.s3.us-east-2.amazonaws.com