Conundrum of Deepfakes An Overview and analysis of recent advancements

Kunal Ghanghav

March 30, 2021

Conundrum of Deepfakes: An Overview and analysis of recent advancements

Kunal Ghanghav **(20103060)**
kghanghav20@iitk.ac.in

KEY WORDS: Deep Fakes, Deep Fake Techniques, Artificial Intelligence and Deep Learning.

ABSTRACT:

Recent advancements in the field of artificial intelligence gave rise to deep fakes which are nowadays more often associated with "Fake news" or "False Information". Deep fakes involve manipulating the available data to fabricate false data in the form of audios, videos, or photos. Technique can potentially be used with malicious intentions in order to jeopardize one's image or to establish dominance in global or local politics. There is significant difference of amount of expertise available for generating deep fakes and detecting them. More research is going on in developing new algorithms and techniques to make deep fakes indistinguishable from real data. While this is good news from the research perspective, at the same time, it is a matter of major concern when it becomes available to those with malicious intentions. This paper mainly emphasizes on the societal impact of deep fakes and presents different analogies to demonstrate how it has a potential to interact with people's mentality in order to understand why it is important to address it as a problem. Paper also overviews a few generation techniques as well as the advancements in the field of detection. Paper also talks about the possible non-algorithmic solutions to tackle the problem.

INTRODUCTION

The word "Artificial intelligence" was first used in the year 1956 by John McCarthy who is known as the "father of Artificial intelligence" [1][2]. Right from the start, the field of Artificial intelligence was able to grab tremendous interest of researchers from all over the world and the field has been improving continuously. According to a recent survey, researchers are even predicting that AI will outperform humans in many activities in the coming 10 years [1]. New advancements and findings in the field of AI are being introduced to the world with each passing day. So, the prediction seems on point as of now.

It is evident that several processes of economics and business are improved using artificial intelligence [18]. AI is being used in conflict resolutions by peace keeping organizations in situations where complexity of perspectives is so high that it becomes humanly difficult to arrive at a decision which is win-win for the conflicting sides [19].

DEEPFAKES

Deep fake is one of those above-mentioned advancement in the field in recent years. In simplest terms, deep fakes are the outcomes of the process of creating data which does not exist in reality. They can be in the form of audios, videos or photos. Deep fakes lie under one of the subfields of AI called "Deep learning".

Using deep learning, we are currently able to generate models which can learn complicated and humanly incomprehensible details in the data to give us the most accurate results. With the development of new deep learning algorithms over the years, we can get to an accuracy of more than 99% in some of the classification problems. Slowly we are approaching similar level of accuracies in case of deep fakes which essentially means that deep fakes are getting more and more difficult to distinguish from the real data [7].

There are a few worthy applications of deep fakes such as, restoring the voices of historical figures to educate children. One company even claims that they can restore the speech of individuals who have lost their voices in some unfortunate incident from the available samples of their voices [5].

Irrespective of the original motive behind creation of deep fakes, they are now being associated more with creation of fake audio or video clips or photos of socially famous people in order to support or oppose one’s argument or in short, we can say, to spread “fake news” [6]. “Fake news” are the fragments of information that are fabricated by manipulating true information as per one’s motive and reported to people as a truth or a fact, hence attempting to mislead them into believing the false information [8].

Unlike the past century, spreading or conveying one’s viewpoint on things to the world has now become a lot easier with the introduction of social media platforms like Facebook, Instagram, twitter, YouTube [5]. Unlike before, news reporting or “the journalism” is not limited to a few organizations. Nowadays, with the evolution of social media individual journalism is also blooming in the society [6]. Spreading false information for malicious intentions is not new to the world but the introduction of deep fakes for creating fake news and availability of resources to spread them, has brought in new dimensions of severity and social impact to the problem [6].

IMPACT OF DEEPFAKES

Artificial intelligence has been an asset to the social media companies from time even before the deepfakes were introduced to the world. Different platforms have been using it for different analytical purposes such as studying the activity of a user on the platform in real-time in order to recommend content which user is finding interesting statistically, to increase the engagement of user on the platform [20] i.e., the recommendation engines [9]. There is an argument that the recommendation engines may cause polarization of opinions, which may lead to conflicts in the society [10]. Deep fakes can speed up this process of polarization significantly. To demonstrate how this could happen, let us first discuss how use of AI may lead to polarization. Let us consider currently hot topic in India, “The farmers’ protest”.

Let us say, I do not have enough information with me to support or oppose farmers’ protest. So, I choose to remain Neutral. Then, on some day, I come across some post or video on Facebook or YouTube or such platform, which is in support or opposing the farmers. I, being curious, read or watch that post or video. I find it convincing. I have some opinion about farmers’ protest now but still I choose to be silent. I forget about it and exit. Next time I open the platform, I see few more such posts or videos, I read or watch them as well. They match my opinions now. They give me false sense of validation regarding my newly formed opinion. Means I end up thinking the opinion that I have, must be correct. Slowly I am becoming a supporter or opposer of farmers’ protest. Recommendation algorithm of the platform has recorded all this activity. Now, it is slowly understanding where I get more engaged. It then starts to recommend me more and more such videos. I engage in most of them. More and more validation for me about my opinion. I feel a need to express how correct I am. I express this feeling on the same platform and some other neutral person in my network comes across this post supporting or opposing farmers’ protest and now he is also trapped in the algorithm. Same thing repeats and I end up surrounded by people on social platforms who share same opinion as me about farmers’ protest which strengthens my opinion more and more. Now, let us consider a neutral person not belonging to my social network, has come across a post which is supporting an opinion which is opposite of my stand, on farmers’ protest. Now, sooner or later, the person and those in his network are more likely to share an opinion opposite to mine and my network, through same process. This is called polarization. Now, we have two groups which have different opinion about farmers' protest. Then, it is just a matter of time when this difference of opinions turns into conflict. If recommendation algorithm hadn't meddled in, chances of any of these conflicts occurring would have reduced significantly. This is a hypothetical situation, but we cannot deny that this has not happened already about this issue or will not happened about any issue in future [10] [11].

The process of polarization or formation of opinion bubbles [6] is likely to take time and it is not certain that it will happen for each and every issue, it has dependencies such as extent of social awareness in the society but use of deep learning for generating deep fakes to spread fake news on these platforms, may bypass these dependencies by creating content so accurate that even a socially aware person gets fooled by them and becomes a victim of the situation unknowingly. The use of deep fakes is likely to trigger this process, even in situations where chances were less, and they could even accelerate the time frame within which any such situation would have occurred under normal use of AI in recommendation engines. This is just one of the possibilities.

There is also an argument that use of deep fakes would affect the society, even if it is not leading to situations like polarization or formation of "individual perception bubbles" [6]. Let us Consider the time of presidential election for some country. Under a disinformation campaign by a foreign country which aims at bringing a political crisis to the target country in order to establish the political dominance, deep fake audio recording is generated for one of the candidates competing for elections. In this fake audio, the candidate is framed for accepting a bribe of 50 million from a company with an agreement that once he or she gets elected as the president he or she will protect the business interest of that company. Now, just before the voting day, the audio is released and presented as audio leak to the public. The result of the election changes dramatically and the targeted candidate looses in election. Later, it is found that the audio was fake. But damage has already occurred [3]. All this due to the said, "advancement" in the field of AI.

Motivated biases are those biases which cannot be changed under any situation [15]. Disproving motivated biases is nearly impossible [16]. Deep fakes have the ability to interact with motivated biases and reinforce them further, ultimately leading us to a limited way of thinking and inability to learn anything new [3].

The Ultimate attribution error is the error in our thought process because of which we try to reason one's activity or associate explanation to one's activity based on prejudiced image of that person in our mind [12]. The Ultimate attribution error is a kind of motivated bias. To understand what exactly does, "Ultimate attribution error" mean, let us take another example. Suppose a student arrives late to school. The effect of this lateness of student on teacher will depend on how the teacher perceives the student. If the teacher perceives the student as not so clever, lazy, and inattentive based on what he or she has observed about that student, he or she will consider that this is how that student is. On the contrary if the teacher thinks that the student is studious, clever based on his marks or his behaviour in class, then the teacher will try to associate reasons to the student's lateness and in that case the lateness will have no effect on how the teacher perceives the student [3]. This is how ultimate attribution error works.

If person A is prejudiced to dislike and mistrust person B and A comes across any genuine deep fake about B. The deep fake acts as a validation to the dislike and mistrust that A has for B. However, if A is prejudiced to like C due to his motivated bias and A comes across such a deep fake about C, and A has heard about deep fakes already, then A is more likely to explain away this situation in favour of C assuming that it is fake [17]. Therefore, denying any real or fake audio or video as fake, always works for people about whom people generally have such prejudices. In both the cases, the walls of prejudices are only becoming stronger for A. In this way deep fakes interact with motivated biases. Whenever any such situation appears in the society, not having any agreed upon reality, deep fakes are likely to worsen the condition of divisions among people.

Discussion till now mainly focused on the societal impact of deep fakes that have high chances of becoming truth in near future. Even though we witness the involvement of deep fakes in our day-to-day life in the form of funny videos of political leaders [22], face swaps [23], face re-enactment utilities on different platforms, platforms which make your selfies dance [21] and so on, but there is certainly not enough awareness of what could be the future of deep fakes if used without caution or not made available for use with caution by the developers.

GENERATING DEEPFAKES

If deep fakes are the problems of near future, then there is a chance that the solution also lies in the process of their creation. So, it is important to understand how deep fakes are generated at the algorithm level in the first place.

The creation of deep fakes mainly relies on Deep neural networks (DNN) and generative adversarial networks (GANs). Among many, Face swap is one of the techniques of generating deep fake images. There are primarily two notable techniques by which Face swap is implemented. First by Feature Selection Networks (FsNets) and Second by Open-Set Identity Preserving Face Synthesis [27].

Face swaps by Feature Selection Network (FsNets) use DNN [25]. They are based on three dimensional morphable models (3DMMs) which model a face into two components. First, a mesh consisting of mean face and other consisting of Two matrices; one for shape and other for texture of face [24]. The two matrices mentioned, describe various modes of variation of the face component from mean.

The FsNet architecture separates the network into two parts, Encoder-decoder network and Generator Network [25]. Based on the variational autoencoder [26], first part converts the source face into latent variable which defines the source face without the geometry of the face and the appearance of non-face part [24]. The generator part uses this latent variable and non-face part of other image, to generate the required face swap image.

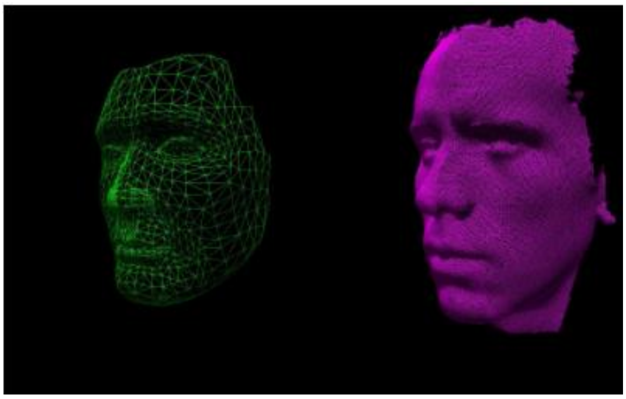

Fig. 3DMM [24]

Open-Set identity preserving face synthesis is another notable face swap technique. It uses framework based on Generative Adversarial Networks (GANs) [28]. It separates target and the source face into two components: identity and attributes of the face. Then it combines the identity component of one face with the attribute component of other face to generate a face swapped image.

DeepFaceLab (DFL) is another method of generating deep fake images. It is an open-source project which abstracts the method of face swapping into three components. Extraction, training and conversion [29]. The extraction part consists of Face detection, Face Alignment and Face Segmentation. For face detection, DFL by default uses Single shot scale-invariant face detector (S3FD) [30]. It also provides facility to use other face detection techniques as well (i.e RetinaFace [31], MTCNN [32]). In the Face alignment step, DFL extracts the facial landmarks. DFL provides two algorithms to implement Face alignment a) Heatmap-based facial landmark algorithm 2DFAN [33] (for faces with normal posture) b) PRNet [34] (for 3D tilted faces). Next, Face segmentation deals with face images that have obstructions such as fingers, hairs or glasses in the view of the normal face. It is optional but provides extra robustness to the DFL technique.

Training stage consists of three layers Encoder, Inter layer and Decoder layer. It uses two types of structures at this stage, DF and LIAE. There are implementational level changes in both the structures at intermediate and the decoder layer [29]. In conversion stage, we perform the face swap. Output can optionally be fed to the discriminator which can analyse whether generated result is recognizable as fake or not and provide feedback to the training stage so that more accurate deep fake results can be obtained.

There are many other ways with which deep fakes can be generated such as GUI based desktop apps like FakeApp [35], OpenFaceSwap [36], ZAO [39] etc., Websites like deepfakesweb.com [37] and thispersondoesnotexist.com [38].

SOLUTIONS

Even though some implementation like thispersondoesnotexist.com [38] have shown impeccable results, deep fakes are still at their primitive stage of development. Studying the possible solutions and presenting them Infront of research community for analysis, has become very important if we want to battle this coming wave of deep fakes. With this, different researchers with their expertise may specify shortcomings in some solution or may even start working on correcting the noticed shortcomings, as their own research problem. This will eventually lead to an optimal solution and possibly we can avoid the severity of impact of deepfakes at different levels such as political, societal or media [6].

While deepfakes are trying to outsmart human intelligence, the present awareness among research community has led them to start approaching the problem of deep fakes in the reverse manner meaning use of deep learning to detect the deep fakes which are themselves generated using deep learning. Researchers believe that deep fakes generated using any complex neural architecture always leave small traces at pixel levels, that are not easily detectible to human [40]. However, available sophisticated algorithms can be employed to detect those tell-tales. A research team from UC Riverside and UC Santa Barbara has developed techniques to detect the scaling, splicing or rotation which are types of digital manipulations often associated with deep fakes [41]. Another research tried to exploit the resolution inconsistencies that occur while creating face swaps [42]. One of the research team has developed a Model based on Convolution Neural network (CNN) along with Long Short-Term Memory Network (LSTM) architecture which identifies the inconsistencies in video frames occurring due to face swaps and have reported accuracies of more than 99.5 % on training sets, more than 96.9 % on validation sets and more than 96.7 % on Test set which is very impressive given currently available research in the field of deepfake detection [43].

In addition to research at algorithmic level, legal and legislative remedies can also be employed which will restrain those who are developing deep fakes with malicious intention such as defamation, spreading fake news for disturbing order in society, causing emotional distress to certain people, blackmailing and list goes on. Making people aware about deep fakes is one of the most important solution as of now. Deep fakes tend to have most impact on the mindset of the society leading them into mistrust about the information from any source including a few genuine sources.

Some of the solutions talk about digital fingerprinting of each and every media item at the source itself which makes it easy to track [44]. Life logging is one more solution. It involves private companies tracking and logging, each and every activity of a few public figures which are prone to be a victim of deep fakes with their consent [45].

CONCLUSION

It is quite evident that deepfakes are going to have different impacts at different levels. There are more negatives of this technology than the positives in the current situation. More progress is going on in the field of deepfake generation compared to their detection. There is very limited amount of expertise available in the field of detection which needs to be improved in the coming years in order to avoid the severe consequences of deepfakes. If used with caution in a ethical way and for solely research intentions, deep fakes technology has a potential to revolutionize the world and create a better future.

REFERENCES

[1] Peart Andy. "Homage to John McCarthy, the Father of Artificial Intelligence (AI)", Web post. 29th October 2020 [Online], Available: https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence

[2] Guillen Beatriz."*The true father of artificial intelligence*", Web post. Ventana al Conocimiento (Knowledge Window), 4th September 2016[Online], Available: https://www.bbvaopenmind.com/en/technology/artificial-intelligence/the-true-father-of-artificial-intelligence/

[3] Sean Dack. "*Deep Fakes, Fake News, and What Comes Next",* The Henry M. Jackson school of internation studies, University of Washington. 20th March 2019 [Online], Available: https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/

[4] Pagin, Shaun. "The Evolution of Photoshop: 25 Years in The Making." Adobe Photoshop History – 25 Years in the Making, 1.

[5] Chesney, Robert, and Danielle Citron. "Deepfakes and the New Disinformation War: The Coming Age of Post Truth Geopolitics." Foreign Affairs 98, no. 1 (2019), 1

[6] Stamatis Karnouskos. "Artificial Intelligence in Digital Media: The Era of Deepfakes". IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY, VOL. 1, NO. 3, SEPTEMBER 2020

[7] You Won't Believe What Obama Says in This Video, YouTube, San Bruno, CA, USA, 2018. [Online]. Available: https://www.youtube. com/watch?v=cQ54GDm1eL0

[8] Fake news, Wikipedia. https://en.wikipedia.org/wiki/Fake_news

[9] P. Ducange, R. Pecori, and P. Mezzina, "A glimpse on big data analytics in the framework of marketing strategies," Soft Comput., vol. 22, no. 1, pp. 325–342, Mar. 2017.

[10] Orlowski, J. (Director). (2020). *The Social dilemma* [Documentary]. Exposure Labs Productions.

[11] Ghanghav, Kunal. "Algorithm influenced", Web post. February 06, 2021 [Online]. Available: https://kunalghanghav.blogspot.com/2021/02/algorithm-influenced.html

[12] Pettigrew, Thomas. (1979). The Ultimate Attribution Error: Extending Allport's Cognitive Analysis of Prejudice. Personality and Social Psychology Bulletin. 5. 461-476. 10.1177/014616727900500407.

[13] Fillion, Rubina Madan. "Fighting the Reality of Deepfakes." Nieman Lab, 3.

[14] (2019). This Person Does Not Exist. [Online]. Available: https://thispersondoesnotexist.com

[15] Hewstone, Miles. "The 'ultimate Attribution Error'? A Review of the Literature on Intergroup Causal Attribution." European Journal of Social Psychology 20, no. 4 (1990): 311.

[16] Hewstone, Miles. "The 'ultimate Attribution Error'? A Review of the Literature on Intergroup Causal Attribution." European Journal of Social Psychology 20, no. 4 (1990): 315.

[17] Fillion, Rubina Madan. "Fighting the Reality of Deepfakes." Nieman Lab, 3.

[18] C. Dirican, "The impacts of robotics, artificial intelligence on business and economics,"Procedia Soc. Behav. Sci., vol. 195, pp. 564–573,Jul. 2015.
[19] D. J. Olsher, "New artificial intelligence tools for deep conflict resolution and humanitarian response,"Procedia Eng., vol. 107, pp. 282–292,Jul. 2015.

[20] R. Alharthi, B. Guthier, and A. E. Saddik, "Recognizing human needs during critical events using machine learning powered psychology-based framework," IEEE Access, vol. 6, pp. 58737–58753, 2018.

[21] Wombo ai App. https://www.wombo.ai/. Accessed: 2021-25-03.

[22] Quartz [YouTube Channel], "*Nothing is real: How German scientists control Putin's face*". Web post. Apr 7, 2016 [Online].
Available: https://www.youtube.com/watch?v=ttGUiwfTYvg

[23] Face swap-GAN, https://github.com/shaoanlu/faceswap-GAN. Accessed: 2021-25-03.

[24] The Surrey 3D face models by university of survey. Web post. Available: https://cvssp.org/faceweb/3dmm/. Accessed: 2021-25-03

[25] Natsume, Ryota, Tatsuya Yatagawa, and Shigeo Morishima. "Fsnet: An identity-aware generative model for image-based face swapping." In Asian Conference on Computer Vision, pp. 117-132. Springer,Cham, 2018.

[26] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241

[27] Teng Zhang, Lirui Deng, Liang Zhang, Xianglei Dang. "*Deep Learning in Face Synthesis: A Survey on Deepfakes*". 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology.

[28] Bao, Jianmin, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. "Towards open-set identity preserving face

synthesis." In Proceedings of the IEEE Conference on Computer Vision and Pattern
Recognition, pp. 6713-6722. 2018.

[29] Ivan Perov, Daiheng Gao, Nikolay Chervoniy. “DeepFaceLab: A simple, flexible and extensible face swapping framework”. 20th May 2020.

[30] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In Proceedings of the IEEE International Conference on Computer Vision, pages 192–201, 2017.

[31] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641, 2019.

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[33] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, pages 1021–1030, 2017.

[34] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 534–551, 2018.

[35] Fakeapp, https://www.fakeapp.com/. Accessed: 2021-25-03

[36] Openfaceswap, https://www.deepfakes.club/openfaceswap-deepfakes-software/ Accessed: 2021-25-03

[37] Deepfakesweb, https://deepfakesweb.com/. Accessed: 2021-25-03

[38] (2019). This Person Does Not Exist. [Online]. Available: https://thispersondoesnotexist.com Accessed: 2021-25-03

[39] ZAO APP. https://www.zaoapp.net/. Accessed: 2021-25-03 March 2019 [Online], Available:
https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/

[40] Villasenor, John. “Artificial Intelligence, Deepfakes, and the Uncertain Future of Truth.” Brookings.edu. 2.

[41] Jason Bunk, Jawadul H. Bappy, Tajuddin Manhar Mohammed, Lakshmanan Nataraj.“Detection and Localization of Image Forgeries using Resampling Features and Deep Learning”. 3rd july 2020 [Online]. Available: https://arXiv:1707.00433v1. Accessed: 2021-25-03

[42] Yuezun Li, Siwei Lyu. “Exposing DeepFake Videos By Detecting Face Warping Artifacts”. Computer Science Department. University at Albany, State University of New York, USA. 22 May 2019 [Online]. Available: https://arXiv:1811.00656v3. Accessed: 2021-25-03.

[43] David Guera, Edward J. Del. Deepfake Video Detection Using Recurrent Neural Networks. Video and Image Processing Laboratory (VIPER), Purdue University. Available: https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf. Accessed: 2021-25-03

[44] For an example of a company making this technology see: Newman, Lily Hay. “A New Tool Protects Videos from Deepfakes and Tampering.” Wired. February 12, 2019

[45] Chesney, Robert, and Danielle Citron. “Deepfakes and the New Disinformation War: The Coming Age of Post Truth Geopolitics.” Foreign Affairs, page 154.