DEEPFAKES: A BRIEF DISCUSSION WITH EXAMPLES

INTRODUCTION

Since the invention of photography in the 19th century, visual media have enjoyed a high level of trust by the public, and unlike audio recordings, photos and videos have seen a widespread use as evidence in court cases (Meskin and Cohen, 2008[i]), and it is widely accepted that visual media are a particularly effective propaganda tool (Winkler and Dauber, 2014). Consequently, the incentives for creating forged visual documents have always been high. Extensive use of manipulated images for political purposed is documented as early as the 1920s by the Soviet Union (Dickerman, 2000; King, 2014).On the other hand, video manipulation took skilled experts and a significant amount of time to create, since every frame had to be changed individually. The technology for manipulating videos was perfected in Hollywood in the 1990s (Pierson, 1999), but it was so expensive that only a few movies made full use of it. Consequently, creating manipulated videos for the purpose of political propaganda was rare. However, a technology known as deepfake that allows manipulation of entire videos with limited effort and consumer-grade computing hardware has recently become available. It leverages modern artificial intelligence to automate repetitive cognitive tasks such as identifying the face of a person in every frame of a video and swapping it for a different face, thus making the creation of such a manipulated video rather inexpensive. Thus, the fundamental change does not lie in the quality of manipulated media, but in the easy accessibility. Given moderate technical skills, input video material, and consumer grade computer equipment, almost anyone can create manipulated videos today (Hall, 2018; Beridze and Butcher, 2019). However, so far deepfake

videos have not played the prominent role in politics that many initially feared (Chesney and Citron, 2019a), even though the year 2020 has seen a massive amount of other misinformation, especially connected to the COVID-19 pandemic and the US presidential election. In this papwer we investigate possible explanations for this development and discuss the likely future developments and identify areas in which deepfakes are expected to be effective and thus likely to appear in the future. We review the deepfake technology, present the latest developments in an accessible manner, and discuss its implications in the context of historical media manipulation. We focus only on the manipulation of visual media in the sense of creating or changing media directly such that they show contents that does not match physical reality, and we exclude misinformation created by incorrect labelling, suggestive editing, and similar techniques

DEEPFAKE EXAMPLES

To date, a whole subgenre of deepfake parody videos has established itself on video platforms like YouTube, with switching the main actors of a movie being the most common manipulation. For example, a short film named Home Stallone (Face, 2020)—a reference to the 1990 comedy movie Home Alone—features an 8-year-old AI generated Sylvester Stallone instead of the real actor, Macaulay Culkin. Many similar videos are available on the internet.

In 2020, several cases of politically motivated deepfake videos made appearance on the news. On February 7, 2020, during the Legislative Assembly elections in Delhi, two video messages surfaced, showing the leader of the Delhi Bhartiya Janata Party (BJP), Manoj Tiwari, addressing his potential voters in English and Haryanvi. According to VICE India, the videos were shared in over 5,800 WhatsApp groups, reaching about 15 million people (Vice, 2020). Surprisingly, Manoj Tiwari does not speak Haryanvi, nor has he ever recorded the video message

in English or any other language. An Indian PR company called “The Ideaz Factory” had taken an older video message of Manoj Tiwari where he spoke about a totally different topic–in Hindi. Then they trained an AI with videos of Manoj Tiwari speaking, until it was able to lip synchronize arbitrary videos of him. Then they used a voice artist to record the English and Haryanvi and merged both audio and video (Khanna, 2020).

On April 14, 2020, the hashtag #TellTheTruthBelgium caught media attention. A video showed a nearly 5 min speech of Belgian premier Sophie Wilmès, depicting the COVID-19 pandemic because of environmental destruction. The environmental movement Extinction Rebellion used deepfake technology to alter a past address to the nation that Sophie Wilmès held previously (Extinction Rebellion, 2020; Galindo, 2020).

Another example is an appeal addressed to Mexican president Andrés Manuel López Obrador. In the video published on October 29, 2020, Mexican author, and journalist Javier Valdez prompts president López Obrador and his administration, to fight harder against corruption and organized crime. Javier Valdez was murdered on May 15, 2017, as his digital alter ego explains in the beginning of the video. Presumably he was killed because of his investigations into organized crime. For 1 min and 39 s, the “Defending Voices Program for the Safety of Journalists” brought Mr. Valdez back to life using deepfake technology to demand justice for killed and missing journalists. The Defending Voices Program for the Safety of Journalists is a cooperation between the Mexican human rights NGO “Propuesta Cívica” and the German journalists association “Reporter ohne Grenzen” (Journalists without Borders) (Reporter ohne Grenzen, 2020).

While many applications of deepfake technology are humorous or benign, they harbor an inherent possibility of misuse. The above examples are not malicious by intent, the

case of Manoj Tiwari shows that it is indeed possible to mislead voters by using synthetic media. The other two examples were published accompanied by a disclaimer, stating that deepfake technology was used. It has been widely suspected that deepfakes would be used to influence the 2020 United States presidential election. However, video manipulation has relied more on conventional techniques (Politifact, 2020; The Verge, 2020). Instead, deepfake videos that warn about threats to democracy have been released (Technology Review, 2020). Thus, so far it seems that outright fabrications rarely make it to the public sphere where they can be debunked, but it is possible that such videos are circulated in closed groups with the intent of mobilizing supporters. Videos involving foreign politicians seem to be especially viable since it is harder for people to judge whether the depicted behavior is believable for that person (Schwartz, 2018). For now, deepfakes do not primarily target the political sphere.

The Dutch startup Sensity. ai (Sensity, 2020), which traces and counts deepfake videos available on the internet, reports that more than 85% of all deepfake videos target female celebrities in the sports, entertainment, and fashion industries. Some of these constitute cases of so-called involuntary pornography where the face of the target person is placed in a pornographic source video. Initially, this required large amounts of images of the target person and thus, only celebrities were targets. However, recent deepfake technology based on generative adversarial networks makes it possible to target persons of whom only a small number of images exist. Thus, considering the wide availability of this technology, using deepfakes for cyberbullying has become a relevant threat. Such an event, called the Nth Room scandal (International Business Times, 2020), happened in South Korea in 2019. The event involved the production and distribution of pornographic deepfake videos using the faces of female celebrities, along with other exploitative practices. Similarly, an extension for the popular messaging app Telegram,

called Deep Nude (Burgess, 2020), became available in 2019 and reappeared in 2020. Given a photo of the target person, it essentially replaces the clothed body with a naked body, thereby creating the likeness of a naked picture of the target person. The program is relatively unsophisticated and was apparently only trained to work for white women. However, its design essentially removes all remaining barriers from the creation of potentially harmful contents. Since it relies on external servers, it is conceivable that more sophisticated programs that deliver high quality images or videos will appear in the future.

A different application of deepfake technology is the impersonation of other individuals in online communication. In 2019, criminals used audio manipulation to impersonate a CEO on a phone call and order an employee to transfer a large amount of money to a private account(Stupp, 2019). Recently, a software that allows impersonation in video calls using deepfake technology was released (Siarohin et al., 2019). Clearly such a technology has a wide range of possible criminal applications, the most concerning among them being the use by sexual predators to impersonate minors (Foster, 2019; Hook, 2019).

Individual Images

Individual images are relatively easy to manipulate via the technique of retouching. The technique was already in widespread use in the 1920s, most prominently in the Soviet Union (King, 2014). The possibility of manipulating photos in general has been commonly known for a long time, the effort involved in the creation of manipulated images was comparatively high until recently. Furthermore, experts were typically capable of detecting such manipulations. Consequently, photos, unlike audio recordings, generally retained public confidence. Even today, the term “photographic evidence” is somewhat commonly used, even

though this confidence seems to be in decline. The manipulation of images has been available to the public since the 1990s. In fact, it is so common today that the term “to photoshop,” named after the Adobe Photoshop program (Adobe, 2020), is being used as a verb for the act of manipulating images. Typically, this refers to minor manipulations that make a person conform to some beauty ideal, but larger manipulations for commercial or entertainment purposes are plenty. Naturally, these techniques can be and have been used for the purposes of propaganda. With the help of image processing software, the task requires a few minutes of work by a skilled user. Now, manipulations such as removing a person can be facilitated further for both individual images as well as larger numbers of images via the use of artificial intelligence (AI). For example, NVIDIA Image Inpainting (NVIDIA, 2020) is a simple tool for removing persons or objects from photos.

1. The effort required by the user is very small. First, the part of the image to be removed must be specified.
2. Then, the system determines how to fill the created gap by analyzing the remaining picture.

The manipulation is never perfect, mostly since the part of the image specified for removal is replaced by an artificial background created from scratch. There are commercial services that perform essentially the same function but produce better results. However, they make use of human intervention and charge for each image to be which makes this approach limited to small numbers of pictures. Similarly, instead of removing a person, it is also possible to replace them with a different person, typically by replacing the face. Automated tools that can do so, e.g., a

program called Reflect by NEOCORTEXT), are available, but they have similar limitations when working on single images. Conversely it is much easier for AI systems to remove a person from a video if the person is moving. The video will contain images of the background, and all the AI must do is correct background from one video frame and place it over the person in a different frame, effectively removing the person. On the other hand, today the manipulation of an individual image carries little weight for political purposes. With the proliferation of digital photography and smartphones that are carried constantly, the number of pictures taken per year has exploded to an estimated 1.2 trillion pictures taken in 2017 (Business Insider, 2017). Thus, for important events, it is usually no longer possible to establish a narrative by censoring or manipulating individual images when thousands of original images exist. Doing so would require manipulating many images automatically. On the other hand, if many images or videos of an event is available, it becomes possible to create believable manipulations using artificial intelligence automatically, assuming most of the material can be accessed and changed. Simple versions of such systems have been used for a while to, e.g., censor pornographic content in online platforms). Similarly, Google Street View automatically detects, and blurs faces and license plates for privacy reasons. But the same technology could be used for far more malicious purposes such as automatically removing or replacing people or events from all accessible video documents. Thus, if the same entity has access to many or all images of an event, it could conceivably use advanced AI to manipulate or censor most images taken of a given event. Because today most photos are taken by smartphones that are permanently connected to cloud computing servers, in a country that has central control of the internet this is technologically feasible. According to the University of Hong Kong, some variants of this technology have been implemented in China (University of Hong Kong, 2020). However, very few other governments

have both the technological capability as well as the constitutional right to implement such a system. Thus, in the following we focus on videos by individuals or small organizations rather than state actors.

Video Alterations

Video traditionally requires a much higher amount of effort and technical skill than photo manipulation. The large number of frames (i.e., images) that need to be manipulated, together with the need for consistency in the manipulation, creates high technological barriers to successful video manipulation. However, just as in the case of photo manipulation, the cost of doing so can be reduced greatly due to the use of new machine learning techniques.

Until recently, direct manipulation of the video material was rare, and a more common way of video-based misinformation was mislabeling, i.e., videos that claim to show something different from what they show, or suggestive editing, i.e., cutting authentic video material in such a way that it misrepresents the situation that was filmed (Matatov et al., 2018). To separate such techniques from deepfakes, the term shallowfakes has been suggested (European Science Media Hub, 2019). A third case is the manipulation of the actual video contents at a level that does not require AI. Such videos are sometimes referred to as cheapfakes. A video showing a politician that was reencoded at reduced speed, thus giving the impression of slurred speech, is the most well-known example of a cheapfakes (Donovan and Paris, 2019). While such techniques can be very effective, they are not new and thus they will not be discussed here as we only focus on AI based manipulation of video contents. To a large degree, the technology for directly manipulating video material was developed for and by the movie industry, mainly in Hollywood. Milestones include Jurassic Park (1993), that added

believable computer-generated dinosaurs to filmed scenes, and Forest Gump (1995), where Tom Hanks is inserted into historic footage of John F. Kennedy. Avatar (2009) showed that given a large enough budget, almost anything can be brought to the screen. However, except for political campaign advertisement, which does not fall under most standard definitions of misinformation, there are very few known cases of the use of this technology for propaganda. A likely reason is the fact that the cost of deploying this technology was very high. Movies containing many high-quality CGI effects typically cost more than US$ 1 million per minute of footage. However, this has radically changed due to the introduction of AI based methods for video manipulation. In the following, we will discuss the methods that make this possible.

TECHNOLOGY OF DEEPFAKES

In 2017, users of the online platform Reddit presented videos of celebrities whose faces were swapped with those of different persons. While the effect was novel, the software that created these images relied on a technology that had been developed a few years earlier. In a landmark paper in 2012, deep learning, a refinement of artificial neural networks, was established as a superior technology for image recognition (Krizhevsky et al., 2012). From there, an immense body of work has emerged in the recent years, proposing both refinements of the method and extensions to other application areas. In addition, this development has had a considerable impact outside the scientific community and brought the topic of artificial intelligence to the attention of politics, industry, and media (Witness Lab, 2020). While the mathematical foundations had been known for decades, the 2012 paper demonstrated that when trained with many suitable input images, convolutional neural networks (CNNs) can categorize

the contents of a picture with high accuracy. A CNN can be trained to recognize specific people and to reliably tell them apart on a wide range of images. The prerequisites for doing so are a powerful computer and many images from which a CNN learns. Once trained, CNNs and other neural networks can be inverted. This is done by specifying an output, and then performing the inverse mathematical operations for each layer in reverse order. Strictly speaking, not all operations can be inverted, but this does not limit the applicability of the concept. The result is an image being created from the abstract features that the network has learned. The original input layer then acts as the output layer. It produces an image of the same resolution as the images that were originally used to train the CNN. A neural network requires a fixed image resolution, but images can easily be scaled. We call a CNN that has been trained to recognize a specific person a detector network, and the inverted version of it a generator network. The deepfake technology relies on combining both types.

Autoencoders for Deepfakes It is possible to link a detector and its corresponding generator. Such a system is called an autoencoder because it learns to encode an image–in our case the face of a person–in some abstract way as a result of its training. Given enough training data, the network can also recognize a face in a noisy image or an unusual angle. Because it represents the face internally in an abstract manner, it can output the face without noise using the generator network. Deepfakes for face swapping are created by training two such autoencoders. One network is trained to recognize the target person whose face is to be replaced by that of the source person, while the other network is trained to recognize the source person. Then, the detector for the source person is linked to the generator for the target person, thus creating a new autoencoder. When applied to a video of the source person, the result is that the face of the target

person appears instead of the face of the source person. This creates the impression that the target person is doing whatever the source person was doing in the input video, thereby creating a substantial potential for manipulation and misinformation.

Deepfakes Software Overview

The technology described above became publicly available in 2017. Programs that implement it formed the first generation of deepfake software, which was *followed* by two further generations of increasing sophistication. We will discuss the software generations here. The first generation of deepfake software required many training images to function properly. Consequently, these programs are impractical for creating manipulated videos of an average person. For that reason, most deepfake videos that were created for entertainment purposes featured famous actors of which many images are publicly available. However, the second generation of deepfake software no longer has this restriction. This is due to the use of generative adversarial networks (GANs) (The Verge, 2020). GANs are a type of neural network like autoencoders. However, in a GAN the detector and the generator networks work against each other. The task of the generator is to create variants of images that are similar but not identical to original inputs by adding random noise. In this manner, the discriminator network becomes very good at recognizing variants of the same image, such as a face seen from many different angles, even if there are few original images to train from. Note that unlike the first generation, which contains easy to use software, most of the second generation of deepfake generators are research codes which require considerable technical skill to be used successfully. However, it would certainly be possible to create user-friendly software from them.

Typically, creating a manipulative video requires the creation of a manipulative audio track. However, compared to video, manipulating audio is a simple task. In fact, in many judicial systems it is significantly harder to establish audio recordings as evidence (Al-Sharieh and Bonnici, 2019). Furthermore, imperfect audio quality can always be masked to appear as a lack of microphone quality in the original recording. Thus, for creating convincing deepfake, audio manipulation is a minor challenge. Nonetheless, the effect can be considerable, as manipulated audio has been used successfully for cybercrime (Stupp, 2019).

The latest generation of deepfake software builds upon the second generation but extends its capabilities in several ways. It significantly increases model complexity, combines multiple generator networks in one model, and extracts features from linked time-series. This allows working in a face-features space rather than in 2D-frame space and simulating changes from frame-to-frame in a naturally looking way. They usually include audio directly to generate natural lip movements.

SUMMARY AND DISCUSSION

We have seen that while photo and even video manipulation has a long history, it is only recent technological developments that have undermined the trustworthiness of video evidence. Modern deepfake software can easily generate videos that a casual observer might perceive as perfectly real. Furthermore, automated systems can be defeated easily if the attacker can test and, if needed, modify, the manipulated video. Thus, with no easy technical solution in sight, and much like with disinformation and fake news in general, education and media literacy remain the foremost defense against disinformation. However, the impact of deepfakes should not be

overestimated. On the one hand, disinformation is widespread, and conventional disinformation has proven effective. Thus, even though the creation of deepfakes is cheap, it still requires more effort than producing texts. Consequently, they will only replace other methods of disinformation where an enhanced effect can be expected. On the other hand, just because an artificially created video cannot be distinguished from a recorded video does not mean that people can be made to believe that arbitrary event happened. Today, video content creators on video platforms such as YouTube, TikTok, and Reddit compete for attention. Due to significant monetary incentives, content creators are pushed towards presenting dramatic or shocking videos, which in turn leads them to stage dramatic or shocking situations, and staged videos are often labeled as such by viewers. Thus, regular users of these video platforms are becoming increasingly aware of manipulative videos. By the same token, increased proliferation of deepfake videos will eventually make people more aware that they should not always believe what they see.

---

i