Deepfake Goes Politics: How the Artificial Intelligence Can Influence the Civil Society?

Semykin Artem

Cybersecurity and Cybercrime course 2021/2022 a.y.
Master program in International Relations
University of Bologna

Abstract: New digital technologies make it increasingly difficult to distinguish real media from fake ones. One of the recent developments describing this problem is the appearance of defects that are hyperrealistic videos created with the use of artificial intelligence (AI). It is used to combine and overlay existing images and videos onto original images or videos using a machine learning technique called a "generative-adversarial network" (GAN). The combination of the existing and the original video leads to a fake video that shows a person or people performing an action that has never actually happened. Combined with the reach and speed of social media, plausible deepfakes can quickly reach millions of views from people and have a negative impact on the whole society. The use of deepfakes in such sphere as politics (especially during the war events or elections) can lead to unwanted and unexpected consequences. This article provides a comprehensive overview of deepfakes and their spread into the politics in recent days and suggests possible ways to detect fake videos in order to save stability and consciousness of society. Artificial Intelligence (AI) is the cause of the problem, but the problem can not be solved fully only from the technical point of view. The more useful and preventive decision should be found in cross-sectoral spheres: education, civil society, public policy.

Key words: deepfake, politics, artificial intelligence, machine learning, generative adversarial networks, public policy, cybercrime, cybersecurity, civil society

General Information about Deepfakes

Deepfake technology has its roots in the distant 1990s. At that time, only special effects experts in the film industry had such tools. Subsequently, the technology was modified in the Internet community, and the software for creating deep fakes was available to download in the open access. Recently, deep fake technology has attracted a lot of attention due to its use in financial fraud, practical jokes and fake news. Deepfake uses the capabilities of artificial intelligence to synthesize a human image: it combines several images in which a person is captured from different angles and with different facial expressions, and makes a video stream from them. Analyzing photos, a special algorithm learns how a person looks and can move. Two neural networks work at the same time. The first of them generates images, and the second is responsible for finding differences between them and real samples. If the second neural network detects a fake, the image is sent back to the first for improvement. Deepfake works using open machine learning algorithms and libraries, which allows to achieve the highest quality of content. The neural network receives images from the library and learns using videos on free video hosting sites. Artificial intelligence, meanwhile, compares fragments of the original portraits with what is on the video, and as a result, a believable material is obtained.

If we try to follow the contemporary history of the development of deepfakes, it is better to take into account the timeline which was made by Deepware company[1]. On the timeline in attachments (Attachment 1) there are the important events, which were described by authors plus added appearances of deepfakes in politics which will be described later. The events highlighted in blue are about deepfakes in politics.

How Are Deepfakes Created?

The process of creating a deepfake has changed as various applications and free software appeared in the public space, but the basic concept of more complex deepfake videos follows the same principles. There is usually an auto-encoder and a generative adversarial network (GAN). An auto-encoder is a computer's way of seeing a face and determining all the ways to "revive" it. He processes how this face blinks, smiles, grins, and so on. A generative adversarial network is a system by which images from an auto-encoder are compared with real images of a target person. It rejects inaccurate expressions, causing new attempts, and the cycle continues indefinitely, gradually approaching the "ideal" in the creation of a person. As a result: the first neural network creates an image of a person's facial expression, the second tells her if the expressions look fake, and they argue until all the images become almost perfect. Figure 1 shows the creation process in a generative adversarial network.

Figure 1 – Work of Generative Adversarial Network (GAN)

Deepfake works using open machine learning algorithms and libraries, which allow to achieve the highest quality content. The neural network receives images from the library and learns using videos on video hosting sites. Artificial intelligence, meanwhile, compares fragments of the original portraits with what is on the video, and as a result, a truthful material is obtained.

[1] Deepfakes Timeline. URL: https://deepware.ai/deepfakes-timeline/

According to a technology report from the Massachusetts Institute of Technology, a device that allows to use the deepfakes may be "an ideal tool for fake news providers who want to influence everything from stock prices to elections"[2].
In fact, "AI tools are already being used to put photos of other people's faces on the bodies of porn stars and put words in the mouths of politicians," writes Martin Giles, San Francisco bureau chief of MIT Technology Review in a report[3]. He said that GAN networks did not create this problem, but only aggravated it. Although image manipulation has a significant history, often used as propaganda during conflicts, the easy availability of digital tools, the highly realistic nature of the forged content and the availability of new media channels for spreading disinformation have turned deepfakes into a viable attack mechanism.
Early-generation deepfakes have already been used to reproduce audio and visual similarities of public figures such as politicians, celebrities, and CEOs. The plausibility of these simplified examples is low, and viewers can determine whether the video is original or fake.

In addition to political information wars, deepfake also creates risks in the field of information security of the corporate sector. Advances and the growing availability of artificial intelligence technologies allow attackers to create highly realistic digital copies of managers in real time by superimposing facial structures and using voice patterns to simulate real voices. As deepfake technologies become more plausible, this new, very convincing threat is already beginning to affect many organizations. Not only do popular mobile apps like Snapchat and Zao allow people to easily create fake content, attackers will be able to buy and sell very convincing fake technologies or services on the dark web and use bots to create fake content on social networks. It will be especially difficult for organizations using low-quality audio and video streaming services to identify this threat, because the drawbacks in even the most simplified deepfakes will stay unnoticed.

Deepfakes turned out to be a good tool in the hands of scammers. The first case of the use of artificial intelligence in phishing, social engineering with the use of automated voice, was used to conduct a high-profile fraud in early 2019. The attackers, copying the voice of the general director of the energy company, were able to convince employees of the organization to transfer $243,000 to a fake supplier. The attackers used social engineering techniques to make the employee to call to the CEO, and since the voice on the other end of the phone was the same as the CEO's, the employee continued to transfer money.

In next part of the article there will be the observations of cases in which deepfakes were used in political context. There will be described fake calls between Russian pranksters and members of the European Parliament in 2021, South Korean presidential elections in 2022 and Russian-Ukrainian crisis of 2022.

Members of the European Parliament vs. Russian Pranksters Vovan and Lexus: Deepfake or Not?

On April 21 2021 parliamentarians from the UK, Latvia, Lithuania, and Estonia had all arranged video calls with an anonymous claiming to be Leonid Volkov, chief of staff to

---

[2] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), p. 1146–1151. DOI:10.1126/science.aap9559
[3] Giles M. (2018). The GANfather: The man who's given machines the gift of imagination. MIT Technology Review, URL: https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

Alexei Navalny, imprisoned leader of Russian opposition. Politicians fell victim to a deepfake: a pranker used "deepfake" technology created specially to trick them. But was it really so?
Those tricked include Rihards Kols, who chairs the foreign affairs committee of Latvia's parliament, as well as MPs from Estonia and Lithuania. Tom Tugendhat, the chair of the UK foreign affairs select committee, has also said he was targeted.

"Putin's Kremlin is so weak and frightened of the strength of @navalny they're conducting fake meetings to discredit the Navalny team," Tugendhat posted in a tweet, referring to the Russian opposition leader Alexei Navalny. "They got through to me today. They won't broadcast the bits where I call Putin a murderer and thief, so I'll put it here."[4]
Latvian politician Rihards Kols posted on Facebook on April 22nd that he had been tricked into taking a call in March with an unknown prankster, and shared two pictures (Figure 2) supposedly showing the real and fake Volkovs.

Figure 2 – Comparison between Original and "Fake" Images of Leonid Volkov

Volkov then reposted the images that same day, ascribing the prank to Vovan and Lexus, popular Russian prankers, but also suggesting that AI was used. "Looks like my real face — but how did they manage to put it on the Zoom call?" he wrote, according to a Facebook. "Welcome to the deepfake era."[5]

The reality is a bit different. Kuznetsov and Stolyarov, known as Vovan and Lexus - a pair of self-proclaimed "pranksters" who have fooled Western politicians and celebrities.

[4] European MPs targeted by deepfake video calls imitating Russian opposition in The Guardian, 2021. - URL: https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition
[5] Ibid.

Over the years, the couple has been tricking phone conversations with people like Justin Trudeau, Elton John, Bernie Sanders, Lindsey Graham and Boris Johnson, each time trying to catch these figures by surprise and extract potentially embarrassing statements from them. As they said to The Verge "Our job is to prank high-ranking officials and celebrities and do a lot of fun posting it on social networks"[6]. Speaking to The Verge, the hoaxers say their imitation Volkov was created using effects no more sophisticated than makeup and artfully obscure camera angles. The couple says they managed to organize various meetings with European politicians and even a live interview on Latvian television with tricks and lies. They did this by cold calls and emails to their targets from fake addresses, using a real photo of Volkov as their digital avatar.

Kuznetsov and Stolyarov were surprised that the prank was described as a deepfake, not least because the image they used, which Volkov himself called fake, is taken from a real video. "It was his real photo, but he denied it was him," Stolyarov says, before Kuznetsov adds that perhaps Volkov didn't like the photo because he looked "too fat."[7]

It is not clear whether the politicians told about deepfakes for this trick because of sincere confusion or for more selfish reasons. Of course, it is less embarrassing to be deceived by a sophisticated fake artificial intelligence than by a couple of pranksters with a convincing manner of writing. But this incident really shows that the fear of deepfakes has the same effect on disinformation as the technology itself. For those who were deceived by Vovan and Lexus the accusation of new technologies may have been just a matter of saving face.

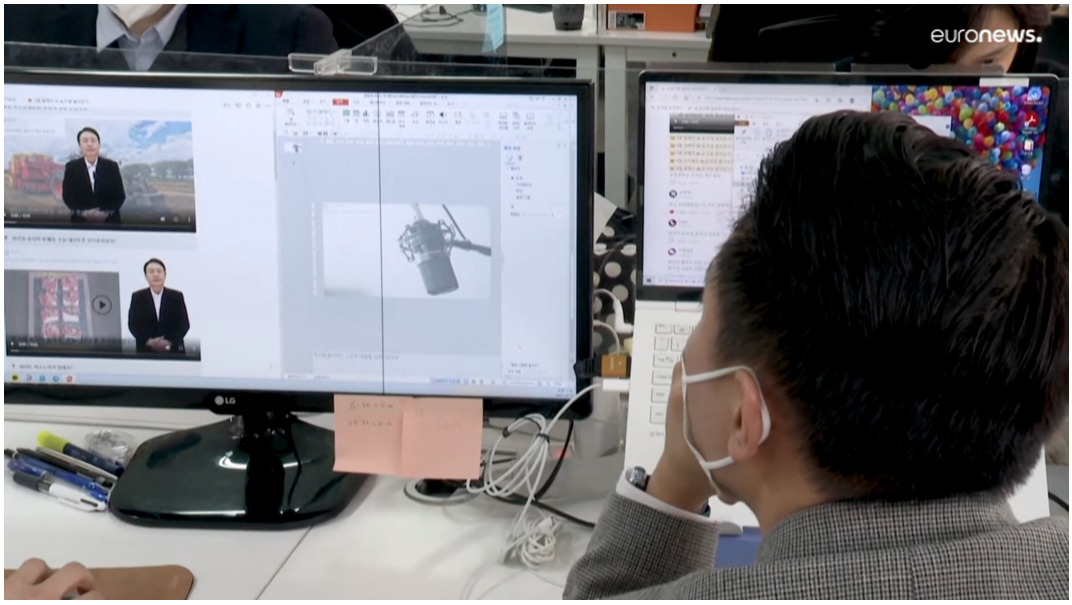
South Korean Virtual Candidate Asked for Votes: Democracy Built on Deepfake

The presidential candidate of South Korea from the conservative party " People Power Party" Yoon Suk-yeol received a digital deepfake double, whose task is to attract votes of young people.

A deepfake avatar named AI Yoon was created by IT specialists from the election team of Yoon Suk-yeol. They processed 20 hours of speeches by the 61-year-old politician, specially recorded for this purpose. The digital person looks like a real candidate, but at the same time uses sharp phrases and jokes that are quickly becoming Internet memes among young voters. Since the start in early January, Yoon has received millions of views on his website, where he answers questions from voters. The character's answers are created by the election team.

[6] 'Deepfake' that supposedly fooled European politicians was just a look-alike, say pranksters. The Verge, 2021. - URL: https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus
[7] Ibid.

Figure 3 – The Creation of the Deepfake AI Yoon



Baik Kyeong-hoon, director of the AI Yoon team: "The image synthesis technique makes it easy to create huge amounts of content, you just need the staff to continue to devote time to this. Deepfake technology will become even more widespread in the next election. It's inevitable."[8]

With neatly-combed black hair and a smart suit, the avatar looks near-identical to the real South Korean candidate but uses salty language and meme-ready quips in a bid to engage younger voters who get their news online. The avatar politician has also used humour to try and deflect attention from Yoon's past scandals, for example claims he received inappropriate fruit gifts from a construction company when he was a senior prosecutor. "I am not beholden to persimmons and melons. I am only beholden to the people," AI Yoon said -- although his campaign was later forced to acknowledge he had accepted some gifts.

This approach has paid off — AI Yoon statements often made headlines in the South Korean media. And finally on the 9th March 2022 Yoon Suk-yeol won the presidential elections with 48,59% of votes. "If we had only produced politically correct statements, we would not have this reaction," Baik said.

Can Deepfakes Change the Course of War: Russian-Ukrainian case

In the third week of Russia's invasion in Ukraine Volodymyr Zelensky, the current Ukrainian president, appeared in the video, dressed in a dark green shirt, speaking slowly and deliberately, standing behind a white presidential podium with the coat of arms of his country (Figure 4). Except of his head, the Ukrainian president's body barely moved when he spoke. His voice sounded distorted and almost hoarse as he seemed to be calling on Ukrainians to put down their weapons in the weeks-old war against Russia.

[8] A deepfake double was created for the presidential candidate of South Korea. Bird in Flight, 2022. URL: https://birdinflight.com/ru/novosti/20220217-deepfake-candidate.html

Figure 4 – Deepfake of Volodymyr Zelensky (in comparison with original)

"I ask you to lay down your weapons and go back to your families. This war is not worth dying for. I suggest you to keep on living, and I am going to do the same," he appeared to say in Ukrainian in the video, which was quickly identified as a deepfake.

In addition to the fake Zelesnky video, which went viral, there was another widely circulated deepfake video depicting Russian President Vladimir Putin supposedly declaring peace in the Ukraine war (Figure 5).

Figure 5 – Deepfake of Vladimir Putin

In just three hours, more than a thousand Facebook users shared a video in which Russian President Vladimir Putin declares the end of the war between Russia and Ukraine.

However, the video was edited from Putin's speech of February 21, in which he explains the decision to recognize the independence of the Donetsk People's Republic and the Lugansk People's Republic.

In the video, which lasts almost a minute and a half, Putin said: "Negotiations with the Ukrainian side have just started. And they were successful for the Russian side. I will inform you shortly. So, we have reached peace with Ukraine. With Ukraine in its world-recognized borders with Donetsk and Luhansk oblasts (territories). We have agreed that we will set up a large foundation together with the USA and the EU for the restoration of infrastructure in these Ukrainian regions. We also signed a five-year roadmap on restoring Crimea's independence as a republic within Ukraine. In the negotiations, I was guided by one principle: to preserve peace and the lives of the Slavic peoples. Russian language in Ukraine will remain — and there will be no oppression, just like the oppression of the Russian population. It is clearly written in peace agreements"[9] In reality, Putin never said this. As is the case with previous deepfake videos of other people, these audio and mouth movements were manipulated through the use of artificial intelligence programs, and the original video is Putin's address to the Russians on February 21, when he explained the decision to recognize the independence of the Donetsk People's Republic and the Lugansk People's Republic[10].

None of the fake videos with Zelensky or Putin came close to the high production performance of TikTok Tom Cruise, for example (firstly, they had noticeably low resolution, which is a common tactic to hide flaws). But experts still consider them dangerous. That is because they show the lightning speed at which high-tech disinformation can now spread around the world. The second reason is that in time of war events people lose their ability to think rationally. Emotions and fears are so strong that affects people consciousness. If currently the only mission is to survive or save your own family, you will probably believe in anything that make your mental health handle easier the stress. As they become more common, video fakes make it difficult to distinguish fact from fiction invented via the Internet, and even more so during the war that is unfolding on the Internet also and is full of disinformation. Even a bad fake has a risk to instability in society. "As soon as this red line is erased, the truth itself will cease to exist," said Wael Abd-Almagid, associate research professor at the University of Southern California and founding director of the school's laboratory of visual intelligence and multimedia analytics. "If you see something and you can't believe it anymore, then everything becomes false. This does not mean that everything will become true. We'll just lose confidence in everything and everything."[11]

Now it has become easier to make higher-quality deepfakes, but perhaps more importantly, the circumstances of their use are different. The fact that they are now being used in an attempt to influence people during the war is especially detrimental, simply because the confusion they sow can be dangerous. Under normal circumstances, according to Siwei Lyu, director of the computer vision and machine learning lab at

---

[9] Fake: Putin did not announce peace between Russia and Ukraine. RE:Baltica/RE:Check, 2022. URL: https://ru.rebaltica.lv/archives/3747 (In Russ.)

[10] Il deepfake di Putin che dichiara la pace con l'Ucraina. La Stampa, 2022. URL: https://www.lastampa.it/esteri/2022/03/18/video/il_deepfake_di_putin_che_dichiara_la_pace_con_lucraina-2876409/?fbclid=IwAR1cXgr-3O8duuj9rg5OW9jNl08vattXK3lSvSjvtNv3yepJzfVfE9W2d3A

[11] Deepfakes are now trying to change the course of war. CNN Business, 2022. URL: https://edition.cnn.com/2022/03/25/tech/deepfakes-disinformation-war/index.html

University at Albany, deepfakes may not have much impact other than attracting interest and attracting attention online. "But in critical situations, during a war or a national catastrophe, when people really can not have an ability to think very rationally, and they have a very short attention span, and they see something like that, that is when it becomes a problem".
Suppression of disinformation in general became more difficult during the war in Ukraine. The Russian invasion of the Ukraine was accompanied by a flood of real-time information that hit social platforms such as Twitter, Facebook, Instagram and TikTok. A lot of it was real, but some of it was fake or misleading. The visual nature of what is shared, along with how emotional and intuitive it often is, can make it difficult to quickly determine what is real and what is fake.

Nina Shik, the author of the book "Deepfakes: The Coming Infocalipsis", considers deepfakes like those made with videos of Zelensky and Putin as signs of a much more serious problem – the disinformation on the Internet, which, in her opinion, social media companies are not doing enough to solve. She argued that responses from companies like Facebook, which quickly claimed to have deleted Zelensky's video, are often a "fig leaf." "You are talking about one video," she said. But a more serious problem still remains.

From this episode, it is tempting to conclude that deepfake algorithms are simply not powerful enough yet, that the terrible moment when deepfakes wreak havoc on political arenas is still in the distant future if it ever comes. However, a closer analysis of the last few years shows a different story: the destabilizing consequences of deep falsifications in politics have already manifested themselves — not in one widely publicized scandal or a stream of numerous incidents, but in an ominous stream of inconspicuous but impressive incidents that have largely escaped attention. These deepfakes of Zelensky and Putin was quickly suppressed, but there are other videos from previous years whose authenticity is unclear — even today we do not know whether they are deepfakes or real. The difficulty of exposing deep forgeries is not just a matter of technological complexity.

The creator of Zelensky's deepfake remains unknown, but the Ukrainian authorities have been warning citizens for several weeks about the possible dissemination by Russia of fake information about the war in the media. Predicting the use of this tactic in early March, Ukraine's military intelligence distributed a video explaining how the falsified videos could be used to incite anxiety and confusion among its citizens. After its publication, many Ukrainians discovered that the video was a scam. Despite the fact that Zelensky's lipsync was fine, the rest of the video was not done very well, Zelensky's accent, head shape and voice do not match.
This deepfake may not have been as convincing, but future versions will be harder to distinguish. Paying close attention to subtle details (such as the mouth, face, and body movements) is a key way to understand whether a video is real or not. In addition, most deepfake creators will rely on pre-existing footage as a template to create their media, which means that finding the source of the original video clip can help eliminate any looming uncertainties.

In the era of news disinformation and machine learning algorithms, it is anxiously to think that we can see the development of this technology and become much more involved in future political conflicts. However, while the Internet allows for the rapid dissemination of fake images and videos, mobile phones and social networks also make it easier for world leaders to expose these falsifications by contacting their citizens directly

at any time. After news about the deepfake, Zelensky addressed to Ukrainians in his official Telegram channel, saying: "We protect our land, our children, our families. So we are not going to lay down our arms. Until our victory."

Risks and Consequences of Deepfakes in Politics

However, the fact that the videos of Russian and Ukrainian presidents were quickly marked as fake ones does not mean that it did not cause harm at all. In a world that is becoming increasingly politically polarized, in which media consumers can believe information that reinforces their biases, regardless of the apparent legitimacy of the content, deepfakes pose a significant threat. Once deepfakes enter the market of political disinformation, the problems we had now, mostly text-based false news, may look like child's play. "It's the calm before the storm" is how an industry insider described it.

Video affects the human brain in a much more direct way than text - then the phenomenon of "I saw it with my own eyes"[12]. Visual content can easily reach brain because on the psychological level people love more storytelling than raw data. And even worse, the very possibility of manipulating the video can raise doubts about anything. An authentic video can be rejected as a fake, and a falsified one can be declared authentic. Video is a good tool to create fakes in a more instinctive way, comparing to text. Over time, public attitudes can turn into "default distrust". What needs to be done? The most promising technical solution focuses on technologies that could limit videos or mark the content at the time of production with something like a watermark. If any manipulations are made with it later, this will affect the watermark. Any user can easily see that the video is not original.

Deepfakes can harm all societies and all governments equally, so hopefully everyone will be interested in preventing abuse. This is a problem that needs a global solution. If some companies do not play along and allow unmarked deepfakes to flourish, the public's trust in online visual content will suffer, even if other companies do their best to prevent it.

The strengthening of the political divide has a similar effect on how people interpret fake news, where users are clearly looking for and accepting information consistent with their previous biases, notes David Lazer, a professor of Northeastern Political Science and Computer Science. However, it is unclear how much a person loses his critical thinking skills when confronted with the media that strengthen his worldview.

“Definitely we are seeing an increase in the polarization of public opinion, and this is obviously one of the factors that can contribute to the spread of misinformation,” says Lazer. “It is likely that political polarization and the spread of disinformation go hand in hand, but this is an area of necessary research.” Director of the Northeastern Laser Laboratory, which conducts research on social influence and networks, Laser research focuses primarily on the spread of misinformation on social networks. In 2019, he co-authored a study on the prevalence of fake news on Twitter during the 2016 presidential election. According to Lazer, Deepfake technology is also “very relevant” for his research, but it is necessary to conduct additional research on various types of disinformation, ways of their dissemination and their psychological impact on media consumers. The rise of

[12] WebFeatCompleate. Comparison of Visual and Plain Text Content, 2021. URL: https://www.webfeatcomplete.com/comparison-of-visual-and-plain-text-content/

political polarization and its impact on media consumption is also a priority area of study, he adds. “We can certainly say that the polarization of many species has increased over the past 40 years, and this is worrying," says Lazer[13].

How to spot a deepfake?

Currently, there are small visual aspects that disappear if you look closely, everything from ears or eyes that do not match the fuzzy borders of the face or too smooth skin, to lighting and shadows. But it is getting harder and harder to detect "gestures" as deepfake technology becomes more advanced and videos look more realistic. The main question is how to detect a deepfake. There are several tips for ordinary users of the Internet and some professional opinions of scientists working with artificial intelligence.

The Deutsche Welle has posted a number of advises you should follow to spot a deepfake[14]:

1) Edges of the face swap. Inconsistencies can be found at the edges of a face swap. These are mostly visible when those edges are close to clothing, hair or jewelry.
2) Earrings. Look closer, if you see two different earrings, it is suspicious. Two different ones and they are somewhat amorphous. For a Generative Adversarial Network (GAN) it is hard to detail intricate jewelry. Often, it puts different earrings in the ears, even though most people do not wear them this way. Also, have a close look at the point where the earring is connected to the ear, and you can say whether the image is a deepfake or not.
3) Glasses. You may think you are looking at someone who is wearing glasses, but are you? If you look closely, you can see that the glasses are different on the left and the right side. One glass has an oval shape, the other is more angular. Try to notice this.
4) Background. Backgrounds that look more like a Martian landscape than a blurry background due to depth of focus are not to be trusted.
5) Teeth. Teeth are hard to 'imagine' by GANs. Sometimes it is clearly visible that teeth are amorphous and do not have clearly defined edges.
6) Clothing. Clothing is individual, and there is a diverse array of different styles. But can it really be that this person’s neck hair is taking over his shirt?
7) Hair. This might well be the most difficult deepfake to spot. No one would expect to see hair growing in the middle of forehead except of face hair. But GAN makes mistakes and generate hair in random places.

In order to minimize the risks of targeted phishing attacks using deepfakes in a corporate environment, it is necessary to inform users about new types of malicious activity and be alert in situations when the behavior of the interlocutor in a telephone conversation or voice message seems unusual. In addition, it is recommended to:

— use multi-factor authentication of employees, electronic signature to protect e-mail messages,

[13] Cote J. Deepfakes and fake news pose a growing threat to democracy, experts warn. News@Northeastern, 2022. URL: https://news.northeastern.edu/2022/04/01/deepfakes-fake-news-threat-democracy/

[14] Fact check: The deepfakes in the disinformation war between Russia and Ukraine. Deutsche Welle, 2022. URL: https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433

— monitor the existence of programs for creating deepfakes on users' computers and attempts to search for such applications on the Internet, pay special attention to such employees and conduct internal checks against them,
— minimize the number of communication channels of the company,
— ensure coordinated dissemination of information,
— limit photos- and video content with the participation of company executives,
— develop a plan to respond to disinformation (by analogy with security incidents),
— organize centralized monitoring of channels and reporting,
— within the company and to communicate with counterparties, apply the practice of introducing oral passwords, code words or control questions, the answer to which is known only to two parties,
— monitor new detection methods deepfakes and methods of dealing with them.

There are several methods for detecting deepfakes:

1. Manual detection
One of the most promising manual methods for detecting deepfake videos is the analysis of human blinking in deepfake videos. Healthy adults blink every two to ten seconds, and one blink takes from one tenth to four tenths of a second. Therefore, you can expect to see similar blinking in a video where a person is talking, but this is not the case in many deepfake videos. The reason for the lack of blinking in deepfake videos is directly related to the algorithm used to create such videos. Deepfake algorithm training is based on images of faces, and very few such images show faces with their eyes closed. Consequently, the training data includes only persons with open eyes, which leads to the displacement of the training data. The last trained model will not be able to understand the action of a blinking eye and will not be able to produce a closed eye very well. The absence of blinking eyes in a deepfake video thus provides a simple but intuitive way to detect a deepfake.
It is possible to develop a method for determining when a person in the video blinks his eyes and whether he blinks in principle. The method will scan every frame of the video, automatically detecting the eyes. Using another deep learning, the neural network will allow you to determine whether the detected eyes are open or closed, depending on the eye appearance, geometric features and eye movement.

Another method of manual detection of deepfake material focuses on the discrepancy between the angle of inclination of the head and face. Methods are available to estimate the position of the head in three-dimensional (3D) space on two-dimensional (2D) video. When creating a deepfake the face is inserted into the video, where the head should point in the other direction at the camera. Therefore, the creators of the deepfake video must perform a 2D transformation to rotate the face so that it corresponds to the orientation of the head. However, the two-dimensional transformation introduces many disadvantages when the subject looks away from the camera or changes the viewing angle of the face position. Unfortunately, this particular detection method does not allow effective detection of deepfake videos when the person in the video is always looking directly into the camera without changing the angle of the face.

The third method of manual detection of deepfake material is based on the shortcomings that arise in the process of creating fake videos. Such disadvantages include:
— double chin or ghostly edges of the face;
— excessive blurring compared to other non-facial areas;

— change of skin tone at the edge of the face;
— double eyebrows or double edges on the face;
— the face is partially covered with hands or other objects;
— flicker or blur on video.

These disadvantages are caused by the fact that the creators of deepfake videos cut corners to reduce the time needed to create videos that can reduce the quality[15]. As a result, the number of pixels on the subject's face in the original video may vary depending on the distance from the camera and the size of the original image. Frames containing faces used to replace the original usually have a fixed size of 64x64 pixels or 128x128 pixels. To adapt to variations, fixed-size faces need to be transformed by enlarging, compressing, or rotating the image to match the original video. Such variations, if combined, will leave some disadvantages such as an excessively smooth face or loss of detail. These shortcomings can be detected by training a deep learning neural network to distinguish changes in the facial area in detail.

The last method of manual identification and detection of deepfakes is the use of detection of the blur. Blur detection is possible because the original face will have more color levels than the deepfake image when zoomed in. Blur detection includes the following steps:

— video stream analysis;
— finding the face area using the cascade Haar classifier (machine learning object detection algorithm);
— convolution using the Laplace operator;
— calculation of variance for the face area;
— division into two cases: the first is a known fake image and the second is an unknown video.

For the first case, the variances of the two facial dependencies are compared. An image with a larger variance is the original, while an image with a smaller variance is a deepfake image. In the second case, an unknown video will be used to find another reference area of the face. For this new area, the variance is also calculated. The next step is to calculate the ratio between these two selected areas of the face and compare the ratio with the threshold, which is set using a large collection of original images. If the ratio exceeds the threshold value, the video is original, otherwise it is a deepfake video.

2. Software-based detection

The first solution for detecting deepfakes with the power of software, called Shallow[16], is a web application that uses Keras convolutional neural networks specializing in image recognition. The solution focuses on distinguishing between real and fake videos in order to protect the reputation and integrity of anyone who may be influenced by deepfake videos. Instead of relying on pre-trained networks used to classify images, Shallow uses randomized networks to increase the accuracy of deepfake video classification.

Users can access the web interface and upload videos for processing. At the initial stage of processing, Shallow will detect and extract clippings of faces available in the uploaded video. The user can then select 20 facial clippings for testing and run the images through the model. The model will analyze the video and report the authenticity of the video.

---

[15] Ageev A. 5 techniques that allow (for now) to distinguish reality from deepfake. — URL: https://www.techcult.ru/technology/7549-5-priemov-pozvolyayushih-otlichit-realnost-ot-deepfake

[16] Shallow - Deepfake detection with deep learning. Github, 2022. URL: https://github.com/mvaleriani/Shallow

The dataset used to build and test the model for Shallow consists of four different categories: real training data, fake training data, real validation data, and fake validation data. To ensure the diversity of both training and verification, 30% of the images contain non-adults. The model was trained on two separate datasets and validated using an additional two datasets. The training was conducted using 15613 different images with a 50/50 ratio between the defect and real images. The test was carried out using 4872 different images with the same 50/50 ratio between deepfake and real images. Based on the use of available training datasets and through validation, the model was able to achieve 99% accuracy. However, it should be noted that the model has been trained and tested using available datasets and may not be representative of other datasets.

The second deepfake video detection solution, called MesoNet, is designed to detect fake faces in videos. The goal of MesoNet is to provide a method for automatic and effective detection of face insertion in a video, as well as to pay special attention to two recent methods used to create hyperrealistic fake videos: Deepfake and Face2Face (transferring the facial expression of an image from the source to the target person)[17].
Pre-trained networks are available for both built models. The dataset used to build and test the model for MesoNet consists of a training and validation set. The training was conducted on 5,111 fake images and 7,250 real images. The verification was carried out using 2998 fake images and 4259 real images. Both models demonstrated a very successful detection rate: over 98% for Deepfake and 95% for Face2Face detection.

As deepfakes get better, researchers and companies are trying to keep up with tools to spot them. Abd-Almageed and Lyu use algorithms to detect deepfakes. Lyu's solution, the jauntily named DeepFake-o-meter, allows anyone to upload a video to check its authenticity, though he notes that it can take a couple hours to get results. And some companies, such as cybersecurity software provider Zemana, are working on their own software as well.
There are issues with automated detection, however, such as that it gets trickier as deepfakes improve. In 2018, for instance, Lyu developed a way to spot deepfake videos by tracking inconsistencies in the way the person in the video blinked; less than a month later, someone generated a deepfake with realistic blinking. Lyu believes that people will ultimately be better at stopping such videos than software. He would eventually like to see (and is interested in helping with) a sort of deepfake bounty hunter program emerge, where people get paid for rooting them out online. (In the United States, there has also been some legislation to address the issue, such as a California law passed in 2019 prohibiting the distribution of deceptive video or audio of political candidates within 60 days of an election.) "We're going to see this a lot more, and relying on platform companies like Google, Facebook, Twitter is probably not sufficient," he said. "Nothing actually beats human eyes."

The Congressional Research Service (CRS) prepared a document about deepfakes and national security with policy considerations and questions for Congress[18]. The Identifying Outputs of Generative Adversarial Networks Act (P.L. 116-258) directed NSF and NIST to support research on GANs. Specifically, NSF is directed to support

[17] Afchar, D. MesoNet: a Compact Facial Video Forgery Detection Network/D. Afchar., V. Nozick. Github, 2021. URL: https://github.com/DariusAf/MesoNet
[18] Sayler,Kelly M. Harris,Laurie A. Deep Fakes and National Security. URL: https://crsreports.congress.gov/product/pdf/IF/IF11333/5

research on manipulated or synthesized content and information authenticity, and NIST is directed to support research for the development of measurements and standards necessary to develop tools to examine the function and outputs of GANs or other technologies that synthesize or manipulate content.
In addition, DARPA has had two programs devoted to the detection of deep fakes: Media Forensics (MediFor) and Semantic Forensics (SemaFor). MediFor, which concluded in FY2021, was to develop algorithms to automatically assess the integrity of photos and videos and to provide analysts with information about how counterfeit content was generated. The program reportedly explored techniques for identifying the audio-visual inconsistencies present in deep fakes, including inconsistencies in pixels (digital integrity), inconsistencies with the laws of physics (physical integrity), and inconsistencies with other information sources (semantic integrity). MediFor technologies are expected to transition to operational commands and the intelligence community.
SemaFor seeks to build upon MediFor technologies and to develop algorithms that will automatically detect, attribute, and characterize (i.e., identify as either benign or malicious) various types of deep fakes. This program is to catalog semantic inconsistencies—such as the mismatched earrings seen in the GAN-generated image in Figure 1, or unusual facial features or backgrounds—and prioritize suspected deep fakes for human review. DARPA requested $28.9 million for SemaFor in FY2023, $7.9 million above the FY2022 appropriation. Technologies developed by both SemaFor and MediFor are intended to improve defenses against adversary information operations.

With the task of detecting deepfakes can handle the ICP systems. The limitation is that such systems should be well-organized and have developed algorithms of monitoring which should has a human administration. It is not easy to plan and organize the mechanism due to the constant development of deepfake tools. Here we can meet the dilemma in which limiting and administrating the monitoring is a very complicated issue because of the challenging control.

Conclusions

In conclusion of this article, it is worth noting that deepfake technology remains a dangerous threat to information security of the 21st century. It is necessary to constantly study the potential and risks associated with deepfakes. At the same time, the most promising areas of using deepfakes are political wars and fraud. In addition, taking into account the constant improvement of technologies, deepfakes can also harm judicial practice — in terms of trust in audio and video materials of the evidence base (dictaphone recordings, DVR files, etc.). Although this article defines and describes the sets of available methods for their detection, these methods have still limited usefulness of the pre-delivered capabilities and challenges in controlling such systems. To solve this problem, a framework should be provided to support the development of new and improved methods for detecting deepfakes.

Deepfakes are the perfect tool for disinformation campaigns because they produce believable fake news that takes time to debunk. Rather small percentage of ordinary people can digest the information and make a right decision that this video is a fake and you can not believe it. Meanwhile, the damages caused by deepfakes — especially those that affect people's reputations — are often long-lasting and irreversible. Combating disinformation, however, has always been a challenge for democracies as they try to

uphold freedom of speech. Human-AI partnerships can help deal with the rising risk of deepfakes by having people verify information. Introducing new legislation or applying existing laws to penalize producers of deepfakes for falsifying information and impersonating people could also be considered. Multidisciplinary approaches by international and national governments, private companies and other organizations are all vital to protect democratic societies from false information. It is better to start form the early childhood with the help of education, when adolescents will learn how to analyze the information, know the basics of critical thinking and learn to trust or distrust in more safe environment which will prevent from mistakes in future.

Attachments

Attachment 1. Deepfake timeline (table)

2017	2018	2019	2020	2021	2022
November - /r/deepfakes Subreddit Created. The term deepfakes originated around the end of 2017 from a Reddit user named "deepfakes".	January - Deepfake Creation Services. The first of many websites offering a deepfake creation service is launched, funded by user donations.	June - DeepNude. a downloadable Windows and Linux application called DeepNude was released, which used neural networks, specifically generative adversarial networks, to remove clothing from women's images.	February - Indian Election Campaign. A day ahead of the Legislative Assembly elections in Delhi, two deepfake videos of the Bharatiya Janata Party (BJP) President Manoj Tiwari criticising the incumbent Delhi government of Arvind Kejriwal went viral on WhatsApp.	January - President of Algeria's video. The President of Algeria's video created great confusion in the public as he was too sick to speak to his citizens. Many people did not believe this video is real, so they checked it on our scanner.	February - Deepfake democracy: South Korean candidate goes virtual for votes
	February - Platforms Banning Deepfakes. Several websites, including Discord, Gfycat, Pornhub and Twitter, ban deepfakes with varying degrees of success.	June – FaceApp. AI photo editor FaceApp goes viral after adding AI-based age filter.	August - AI-generated Elon Musk joined a Zoom call has gone viral. A new deepfake-powered filter allows video conferencing participants to make themselves look like practically any famous person, living or dead, ranging from the Mona Lisa to Steve Jobs and Elon Musk.	March - Tom Cruise Deepfake. The convincing Tom Cruise deepfakes that went viral took lots of skill to create.	March - Unknown published a deepfake of Zelensky with an appeal to "lay down arms"
	April - Rana Ayyub is Targeted by Deepfake Pornography. Indian journalist Rana Ayyub is targeted by deepfake pornography using her image and fake impersonation accounts on social media, with the intent to humiliate and defame.	June - House holds hearing on "deepfakes" and artificial intelligence amid national security concerns. The House Intelligence Committee heard from experts on the threats that so-called "deep fake" videos and other types of artificial intelligence-generated synthetic data	September - Putin Deepfake Video. A new advertising campaign from anti-corruption non-profit RepresentUs uses deepfake technology to deliver a message about democracy.	March - The video of one minister's corruption confession divided people. On social media, many denounced it as "poorly edited," a "deepfake" and "forged."	March – Fake: Putin did not announce the peace in Ukrainian conflict

		pose to the U.S. election system and national security at large.			
	April - Obama Deepfake Video. Deepfake video of former US president Barack Obama raises mainstream awareness.	July – FaceApp. AI photo editor FaceApp goes viral after adding AI-based age filter.	September - Microsoft Deepfake Detection Tool. Microsoft made public that they are developing a Deepfake detection software tool.	April - European MPs targeted by deepfake video calls imitating Russian opposition.	April - Reface and partners have created a deepfake video in which Putin tells the truth about the war
	May - US Senator Voices His Concern About Deepfakes. US Senator Marco Rubio voices his concerns about deepfakes at the Senate Intelligence Committee nomination hearing.	August - DARPA Is Taking On the Deepfake Problem. The Defense Department is looking to build tools that can quickly detect deepfakes and other manipulated media amid the growing threat.	October - Twitter Introducing New Policy Towards Deepfakes. Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media	May - A 'beautiful' female biker was actually a 50-year-old man using FaceApp. He said he used photo editing apps such as FaceApp to make himself appear like a youthful woman.	
		August - DARPA Is Taking On the Deepfake Problem. The Defense Department is looking to build tools that can quickly detect deepfakes and other manipulated media amid the growing threat of "large-scale, automated disinformation attacks."	October - Deepfake Telegram Bot. Deepfake bot on Telegram is violating women by forging nudes from regular pics.	May - New York's Right to Publicity and Deepfakes Law Breaks New Ground. WilmerHale Counsel Matthew Ferraro and Partner Louis Tompros have contributed an article published in The Computer & Internet Lawyer (April 2021).	
		September - CEO Deepfake Scam. Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 in what cybercrime experts described as an unusual case of artificial		June - Facebook doubles down on detecting deepfakes. Facebook has collaborated with researchers at the Michigan State University (MSU) to develop a method of detecting and attributing deepfakes.	

		intelligence being used in hacking.			
		October - California and Texas ban political deepfake videos. Two states have passed a law meant to prevent altered "deepfake" videos from influencing elections in a plan that has raised free speech concerns.			
		October - Deepware Scanner Released. Deepware Scanner released as a first deepfake detection tool against deepfakes where people can easily scan and detect deepfake videos.			
		December - Deepfake Detection Challange (DFDC).			
		December - David Beckham Deepfake Video. The recent global campaign showing Malaria survivors speaking through David Beckham to help raise awareness around the Malaria Must Die initiative spooked a lot of people.			

References

1) Deepfake timeline by Deepaware. URL: https://deepware.ai/deepfakes-timeline/
2) Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), p. 1146–1151. DOI:10.1126/science.aap9559
3) Giles M. (2018). The GANfather: The man who's given machines the gift of imagination. MIT Technology Review. URL: https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/
4) European MPs targeted by deepfake video calls imitating Russian opposition. The Guardian, 2021. URL: https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition
5) 'Deepfake' that supposedly fooled European politicians was just a look-alike, say pranksters. The Verge, 2021. URL: https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus
6) A deepfake double was created for the presidential candidate of South Korea. Bird in Flight, 2022. URL: https://birdinflight.com/ru/novosti/20220217-deepfake-candidate.html (In Russ.)
7) Fake: Putin did not announce peace between Russia and Ukraine. RE:Baltica/RE:Check, 2022. URL: https://ru.rebaltica.lv/archives/3747 (In Russ.)
8) l deepfake di Putin che dichiara la pace con l'Ucraina. La Stampa, 2022. URL: https://www.lastampa.it/esteri/2022/03/18/video/il_deepfake_di_putin_che_dichiara_la_pace_con_lucraina-2876409/?fbclid=IwAR1cXgr-3O8duuj9rg5OW9jNl08vattXK3lSvSjvtNv3yepJzfVfE9W2d3A
9) Deepfakes are now trying to change the course of war. CNN Business, 2022. URL: https://edition.cnn.com/2022/03/25/tech/deepfakes-disinformation-war/index.html
10) Cote J. Deepfakes and fake news pose a growing threat to democracy, experts warn. News@Northeastern, 2022. URL: https://news.northeastern.edu/2022/04/01/deepfakes-fake-news-threat-democracy/
11) Fact check: The deepfakes in the disinformation war between Russia and Ukraine. Deutsche Welle, 2022. URL: https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433
12) WebFeatCompleate. Comparison of Visual and Plain Text Content, 2021. URL: https://www.webfeatcomplete.com/comparison-of-visual-and-plain-text-content/
13) Ageev A. 5 techniques that allow (for now) to distinguish reality from deepfake. — URL: https://www.techcult.ru/technology/7549-5-priemov-pozvolyayushih-otlichit-realnost-ot-deepfake (In Russ.)
14) Shallow - Deepfake detection with deep learning. Github, 2022. URL: https://github.com/mvaleriani/Shallow
15) Afchar, D. MesoNet: a Compact Facial Video Forgery Detection Network/D. Afchar., V. Nozick. Github, 2021. URL: https://github.com/DariusAf/MesoNet
16) Sayler,Kelly M. Harris,Laurie A. Deep Fakes and National Security. URL: https://crsreports.congress.gov/product/pdf/IF/IF11333/5