Dark sides of deepfake technology

Sanela Z. Veljković[a], *Milica* T. Ćurčić[b], *Ilija* P.Gavrilović[c]

[a] University of Belgrade, 'Vinča' Institute of Nuclear Sciences - National Institute of the Republic of Serbia, Belgrade, Republic of Serbia,
e-mail: sanela.veljkovic@vin.bg.ac.rs, **corresponding author**,
ORCID iD: https://orcid.org/0009-0003-3650-290X

[b] University of Belgrade, 'Vinča' Institute of Nuclear Sciences - National Institute of the Republic of Serbia, Belgrade, Republic of Serbia,
e-mail: milica.curcic@vin.bg.ac.rs,
ORCID iD: https://orcid.org/0000-0002-4326-4036

[c] University of Belgrade, Faculty of Political Sciences,
Belgrade, Republic of Serbia,
e-mail: gavril502323@student.fpn.bg.ac.rs,
ORCID iD: https://orcid.org/0009-0005-1348-1792

https://doi.org/10.5937/vojtehg72-49630

FIELD: computer science, IT, security, artificial intelligence
ARTICLE TYPE: review paper

Abstract:

Introduction/purpose: Artificial intelligence can be used for both positive and negative purposes. In recent years, the use of deepfake technology has attracted significant attention. Deepfake technology replaces a person's face and creates events that never happened. While the use of deepfake was more noticeable in the past, the technology has advanced so rapidly that today it is impossible to determine if the content is fake or not. As a result, there is erosion of trust in the media and political institutions, manipulation of public discourse, as well as the spread of disinformation and fake news. The aim of this work is to examine the methods of creating deepfake content and explore the possibilities for detecting such content. A special focus is placed on investigating the dark side of deepfake technology, i.e., the negative purposes for which deepfake technology can be used.

ACKNOWLEDGMENT: This work was carried out within the scientific research activities of the 'Vinča' Institute of Nuclear Sciences - National Institute of the Republic of Serbia, funded by the Ministry of Science, Technology, and Innovation, grant number 451-03-66/2024-03/ 200017.

Methods: Through the use of literature review methods and content analysis, this work has provided a systematization of knowledge about deepfake technology, as well as an analysis of relevant data in this field regarding the potential misuse of deepfake technology. Deepfake technology and its use are viewed from a security perspective, i.e., how the use of these technologies can pose a social hazard. Future research should be designed to be multidisciplinary, integrating knowledge from social sciences (security, sociology, psychology) and technical sciences (information technology).

Results: The results of this research show that in a positive context, the use of deepfake is associated with medicine, the film industry, entertainment, and creative endeavors. However, deepfake is often used to create pornographic content, revenge porn, fake news, and various types of fraud.

Conclusion: Deepfake technology is neutral in the sense that the purpose of its use depends on the individual creating the content. The use of both artificial intelligence and deepfake technology raises complex legal and ethical questions. Although there is noticeable potential for societal improvement offered by these technologies, deepfake technology simultaneously poses a serious risk to human rights, democracy, and national security. Therefore, the misuse of deepfake technologies represents a social hazard for the entire population of any country. Women are particularly vulnerable due to the possibility of creating pornographic content and revenge porn using deepfake technology, although victims of this act can also be men.

Key words: deepfake, pornographic content, cyber violence, scams, fake news.

Introduction

Results of technical and technological progress often provoke certain resistance and apprehension among people, and in that regard, the development and potential application of artificial intelligence in everyday life is no exception. In society, demands for the prohibition of using artificial intelligence are increasingly heard, mostly due to the fear of losing jobs that could be replaced by this technology. In the 1990s, with the emergence of the Internet, people were concerned about the possibilities it offered. However, over time, they adapted, and many began to use the Internet not only for business but also for other purposes. The use of the Internet has facilitated people's lives by providing a lot of information that was not previously available. Today, that fear has been replaced by the fear of artificial intelligence. One of many examples is the emergence of ChatGPT, or language processing artificial intelligence models. This program is capable of understanding written and spoken human language,

as well as established rules of communication that enable it to understand posed questions and provide answers. However, the way this program will be used depends on the individual using it. The same goes for deepfake technology.

The development of deepfakes began with harmless filters on social media that allowed people to replace their friend's or family member's face with theirs own. These filters, although of low quality, caused immense enthusiasm among users of various social platforms. In the following years, other applications offering different possibilities emerged. The result of the development of this deepfake technology is that today it is impossible to determine whether the content is fake or not (Botha & Pieterse, 2020, p.62). This situation has led to numerous abuses of this technology, and its use is estimated to have negative consequences not only for individuals but also for the national security of states. Moreover, "products and services based on artificial intelligence technology have the potential to undermine human rights, democracy, and the rule of law" (Prlja et al, 2022, p.125). States are concerned about the spread of fake news and the creation of events that never happened, raising questions about new models of abuse of this future technology. On the other hand, the faces of many famous actresses and celebrities have been used to create deepfake pornographic video content.

The societal danger of this phenomenon stems from the fact that not only individuals in the public sphere are at risk, but anyone can become a victim of the use of this technology. Deepfake technology is easily accessible, and the costs of its use are minimal, almost conditioned solely by the motivation and knowledge of the individual abusing these technologies, while the consequences can be significant and serious. Therefore, in the first part of the paper, the development and functioning of deepfake technology will be presented, as well as the possibilities of detecting such content. In the second part of the paper, the dark sides of the use of deepfake technology will be presented in the form of creating pornographic content, revenge porn, fake news, as well as the possibilities of using this technology to carry out various types of scams.

Definition and characteristics of deepfake technology

The use of deepfake technology is conditioned by the technological advancement of information technologies, hence it does not have a long lifespan. These are new technologies, so the definitions of this type of technology date back only a few years. "Deepfake is a medium, usually video, but can also be just audio, or a combination of both - which is altered

using artificial intelligence and machine learning techniques" (Grothaus, 2021, p.1). Deepfake technology involves placing the face of one person onto the body of another (Mania, 2024, p.1). The goal of using this technology is to create events that never actually happened. Deepfake refers to "an image, sound, or other material that is entirely or partially created, or an existing image, sound, or other material that is manipulated using advanced technical means, and is difficult or impossible to distinguish from authentic material" (Van der Sloot & Wagensveld, 2022, p.4). There are two methods of creating deepfake content - autoencoders and generative adversarial networks (GANs). The development of deepfake technology has been almost fantastic over just a few years since its emergence. As Grothaus explains, when GANs were first created, it took hundreds or thousands of images of a person's face to achieve a convincing deepfake. By 2015, when the first deepfake application for smartphones, Face Swap Live, was released, face swapping could be done in real-time via the phone's camera, although the results were far from perfect. However, by 2020, deepfake technology on smartphones had advanced to the point where an application like Reface could create a realistic deepfake based on just one image of a person's face and then place that person into scenes from Hollywood movies with barely noticeable differences (Grothaus, 2021, p.26). Further development of deepfake technology raises concerns about distinguishing between fake and real content.

The compound word "deepfake" is formed by combining the English words "deeplearning" and "fake" (Vatreš, 2021, p.562). While "fake" translates as false, "deeplearning" is used in the IT world to denote deep learning, which is actually "a type of machine learning - one of the basic techniques that drives artificial intelligence. Deep learning is a key tool in the field of computer vision, where computer systems are tasked with identifying subjects in videos and photographs, and is used in everything from self-driving car systems and facial recognition systems to applications on smartphones that automatically tag owner's friends in pictures" (Grothaus, 2021, p.3). The first method of creating deepfake content relies on autoencoders. They represent a type of artificial neural network used for learning efficient data encoding, with two learning functions: an encoding function that transforms input data and a decoding function that recreates the input data from the encoded representation. In the case of generative adversarial networks, two neural networks compete against each other in a zero-sum game, where one network's gain is the other's loss. In the first technique, there is an encoder that extracts latent facial features from the image and a decoder that is used to reconstruct the facial

image, both using the same network that finds and learns similarities between two sets of facial images such as the position of eyes, nose, and mouth (Nguyen et al, 2022, p.3). In the GAN technique, there are two neural networks, a generator capable of producing images, and a discriminator whose role is to distinguish fake images from real ones (Nguyen et al, 2022, p.5). The techniques for creating deepfake content mentioned are illustrated in the following figures, where Figure 1 refers to the technique using autoencoders, while Figure 2 depicts the functioning of generative adversarial networks (GANs).

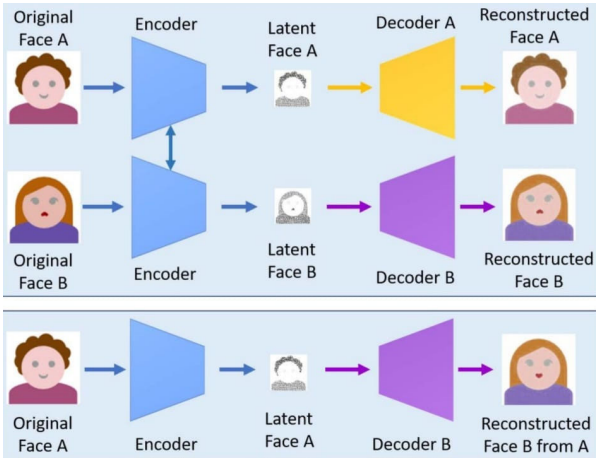

Figure 1 – Method of creating deepfake content through autoencoders (Nguyen et al, 2022, p.3)

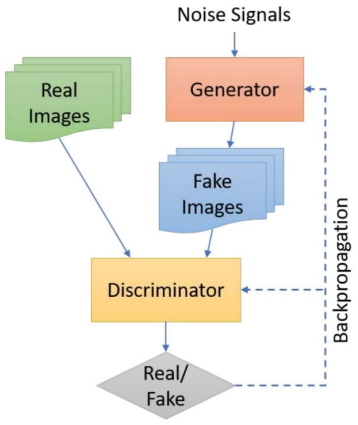

Figure 2 – Method of creating deepfake content using GANs (Nguyen at al, 2022, p.5)

Meskys et al. explain that the key idea of GANs is for the generator to try to generate images that would deceive the discriminator into believing they are real, while the central idea for autoencoders is for the encoder to try to encode the input image as a small vector, and the decoder to try to decode this small vector into the original image (Meskys et al, 2020, p.4). The GAN technique is a better version for creating deepfakes than autoencoders because its results are much more realistic. In this technique, two neural networks play the roles of a forger and an inspector, where "each time the AI inspector labels a fake photo as fake, the AI forger inspects aspects of that forgery, which naturally were not good enough to fool the AI inspector, analyzes a set of authentic photos, identifies what is different about them, and then tries again to fool the AI inspector with another, improved forgery" (Grothaus, 2021, p.4). Despite the possibilities offered by using GAN technology to create deepfake content, what is particularly concerning from a security perspective is that with the development of the technology, the number of photos needed to create fake content has decreased over time. This further indicates that not only politicians and celebrities are at risk but also every individual, given that almost everyone readily shares their photos on social media today. Moreover, the accessibility of deepfake technology is alarming, as extensive knowledge of information technology is not necessary to create fake content. This raises the question of who can create deepfake content, and the answer is that anyone with motivation and a smartphone can do so. In the literature, four groups of actors are usually distinguished: communities composed of individuals for whom this is a hobby; political players such as foreign governments and various activists; malicious actors aiming to carry out various types of scams; and legitimate actors such as television companies (Westerlund, 2019, p.41).

Despite the presence of various legitimate actors and many positive purposes, the use of deepfakes is most commonly perceived as a threat to state security and the stability of the political system. The reasons are numerous, among which stand out: pressure on journalists struggling to distinguish real news from fake; national security is compromised by the spread of propaganda and interference in elections; citizens' trust in information coming from authorities is undermined; and questions about internet security for individuals and organizations arise (Westerlund, 2019, p.42). However, it should be borne in mind that the purpose of using any technology, including deepfake technology, is determined by the individual using that technology. The fact that tools, as well as technology in general, do not inherently carry essence or value is especially evident when Grothaus emphasizes that moral judgment can only be made about the

person who used the technology in a certain way. As an example, he cites the use of an airplane, which can be used as a tool for evacuating civilians from areas affected by natural disasters or as a tool for demolishing buildings. This confirms the hypothesis that the airplane as a form of technology is neutral in both cases, and it is the human who determines its purpose according to the given situation of use (Grothaus, 2021, p.25). Such perception can also be applied in the case of using deepfake technology. However, since there will always be individuals who will use this technology for negative purposes, the next section of the paper presents certain possibilities for detecting deepfake content.

Possibility of detecting deepfake content

For a short period, deepfake technology has evolved and been able to correct earlier flaws, which makes detecting fake content more difficult. There are numerous artificial intelligence tools used for detecting deepfake content. However, the problem with this approach lies in establishing a feedback loop between two groups of artificial intelligence. Namely, when a detection deepfake finds a new indicator revealing that content is fake, the deepfake creating that content learns how to correct the flaws and deceive the detection deepfake (Grothaus, 2021, p.223). In the literature, examples of different methods for detecting deepfake content can be found. According to one group of authors, detection methods include facial recognition such as tracking eye blinks, eye color, and dental flaws; multimedia forensics that interpret parameters such as pixel correlation, image continuity, and lighting; applying watermarks that allow identification of alterations; as well as convolutional neural networks working on machine learning principles enabling the detection of fake content through powerful image analysis functions (Albahar & Almalki, J, 2019, pp.3247-3248). On the other hand, some authors have classified methods for detecting fake content into general network-based methods; methods based on temporal consistency; methods based on visual artifacts; methods based on camera fingerprints; and methods based on biological signals. These methods are illustrated in Table 1.

Considering the variety of methods for detecting deepfake content proposed by different authors, it seems that there is no problem with their detection. However, given the short period in which deepfake technology has managed to overcome previous shortcomings, there remains a fear that in the future it will not be possible to distinguish deepfake content from genuine content. Today, "although you may be able to prove that a video is a deepfake, you can never prove with absolute certainty that the video

is not a deepfake" (Grothaus, 2021, p.2017). In this regard, new scientific research on deepfake technology mainly focuses on two levels of harm, specifically on: "individual harm to the dignity and emotional well-being of subjects of deepfake recordings, and broader societal harm that includes threats to national security and democratic institutions" (Kugler & Pace, 2021, p.623). Deepfake technology does not possess essence; rather, its manner of use is determined by the individual employing it. In a positive context, the use of deepfake technology is associated with medicine, the film industry, entertainment, and creative expression. However, it is essential to emphasize the impact that deepfake technology can have on each individual as well as on the state if used for negative purposes.

Table 1 – Methods of detecting deepfake content (Marković, 2022, pp.33-34).

Methods	Description
General network-based methods	In this method, detection is considered a frame-level classification task ending with a Convolutional Neural Networks.
Methods based on temporal consistency	It has been determined that there are inconsistencies between adjacent frames in deepfake videos due to flaws in the falsification algorithm. Therefore, Recurrent neural network is applied to detect such inconsistencies.
Methods based on visual artifacts	Mixing operations in the generation process would cause substantial deviations in the image within the blending boundaries. Methods based on on Convolutional Neural Networks are used to identify these artifacts.
Methods based on camera fingerprints	Due to the specific generation process, devices leave different traces on captured images. At the same time, it is acknowledged that faces and background images come from different devices. Therefore, the detection task can be completed using these traces.
Methods based on biological signals	GANs find it difficult to understand the hidden biological signals of faces, which complicates the synthesis of human faces with reasonable behavior. Based on this observation, biological signals are extracted to detect fake video clips.

Possibilities of abusing deepfake technology

The availability and ease of use of deepfake technology results in every individual possibly discovering that their face has been exploited to create deepfake pornographic content. Moreover, every individual may find themselves in a situation where they receive a deepfake video via email, featuring their face, showing actions that never actually occurred, potentially intended for blackmailing the person depicted. This raises questions about human perception, questioning whether one can trust their eyes and ears when watching a video of a politician discussing important international, regional, and national issues. Lastly, each of us could fall victim to identity theft and various other scams that are now feasible due to the advancement of deepfake technology. In the following section, the dark side of deepfake technology is presented, namely the potential for its misuse by individuals who employ it. Primarily, the use of deepfakes to create pornographic content is highlighted, as well as how the development of this technology has resulted in any woman becoming a potential victim of deepfake technology. Additionally, as a result of the development of this technology and the inability to distinguish between fake and authentic content, various forms of cyberbullying and the spread of fake news are possible. Thanks to deepfakes and their widespread availability, identity theft is now possible not only in the form of stealing someone's face but also their voice.

Pornographic content

The use of deepfakes to create pornographic content featuring famous and well-known personalities is often cited as one of the primary negative uses of this technology. The integrity of numerous actresses, singers, and other famous women became compromised when the first pornographic deepfake video content was created. In the early years, it was not possible to prove the misuse of the technology and thus demonstrate the falsehood of such posts, and the videos were further distributed over the Internet, creating the illusion that they depicted real events. The creation of deepfake pornography using the face of a famous individual poses several problems. Firstly, neither the famous personality whose face is used in the video nor the person whose body is used has given consent for the creation of this type of content (Grothaus, 2021, p.70). On the other hand, even though these videos feature famous personalities and are widespread on the Internet, the creators of such content fail to consider the consequences it may have on the individual. It seems that despite crossing many ethical boundaries, the creators of such

content have certain rules when making fake pornographic videos. These rules are as follows: there are no fake images or videos of any famous personality under the age of 18, and there are no fake images or videos of any famous personality over the age of 18 using their face or body from images or videos taken when they or the donor body were under 18 years old (Grothaus, 2021, p.86). Of course, this does not justify the creation of fake pornographic content using the face of a famous personality over 18 years old. Research indicates that pornographic deepfake content mainly elicits negative emotions among online users. Anger, disgust, and contempt are typically reported as responses to this type of content (Wang & Kim, 2022, p.7). However, in the following years, the development of deepfake technology has resulted in not only famous and prominent individuals but also any person worldwide being potentially subjected to victimization.

DeepNude was the first software specifically aimed at any female individual, not just famous and well-known women, by enabling the removal of clothing from images and creating fake content that could then be shared via the internet and social media (Grothaus, 2021, p.97). This software provided the opportunity for one individual to easily harm another person, facilitated by the fact that people willingly share their pictures on various social media platforms today. Nowadays, the availability of photographs is not a problem because we all readily share our pictures on numerous social media platforms. However, DeepNude was not the only weapon against women. In 2020, the company Sensity released a shocking report revealing that "all it took for anyone who wanted to create 'deepfake' women was to send a photo of the victim to a chatbot in the popular Telegram messaging application. The bot would then undress the woman and send a fake nude photo to the person who requested it" (Grothaus, 2021, p.98). A vast number of women fell victim to this situation, with their faces being used on naked bodies generated by artificial intelligence. Many of them were unaware that their faces were being used in this way by colleagues, friends, or acquaintances. Moreover, there was the possibility of paying a fee to remove the watermarks from the fake nude images created by the bot (Grothaus, 2021, p.99). This is particularly concerning because if a victim's photo were to surface, they would not be able to prove that the image resulted from the use of deepfake technology. Although the integrity of every individual can be compromised through the use of artificial intelligence, practice has shown that women are a more vulnerable category, given the societal danger posed by this threat. Despite the stated use of DeepNude, today the software is used from images of men and women to images of animals and even objects.

The importance of a watermark lies in the fact that it indicates that the content is fake. However, the problem arises with the possibility of removing the watermark before sharing the image with other users via the Internet (Grothaus, 2021, p.101). The possibility of removing watermarks puts the victim of this technology's use in a particularly difficult position, making it challenging to prove that the content is false. Additionally, removing watermarks from images or videos produced by artificial intelligence is problematic in the context of revenge porn. An individual wishing to harm another can easily do so using deepfake technology. The ease of abusing deepfake technologies for pornographic purposes, as well as the significant, often irreparable consequences of publishing such content on the Internet, make these technologies particularly dangerous. Furthermore, most people are unaware of the existence of such technologies, so they interpret such content as authentic, further stigmatizing the victims. Such a situation can "erode trust, leading individuals to question the legitimacy of visual and auditory content, which further results in increased skepticism and confusion" (Wazid et al, 2024, p.2). This not only compromises the victim's privacy but also their mental health, deepening this negative social phenomenon.

Cyber violence

With the development and further refinement of deepfake technology, the possibility has emerged for fake content to be used as a means of extortion. "The most obvious example of this is the scenario in which someone creates 'deepfake' targets but does not publish the 'deepfake' publicly. Instead, the 'deepfaker' sends the target themselves in 'deepfake.' They might show the target physically or sexually assaulting someone or engaging in taboo sexual acts. The goal here is not to deceive the target into thinking it is real—because they know it is not. Instead, the aim is to induce anxiety, fear, and concern in the target that others might believe it is real if the 'deepfake' were publicly released" (Grothaus, 2021, p.115). These activities can leave serious negative consequences for the victim, and such behavior represents a form of cyber violence. Deepfake videos may not only contain sexual scenes that never occurred; today, there is a "threat that people will be implicated in 'deepfakes' showing them committing crimes they did not commit" (Grothaus, 2021, p.113). The impact this would have on the potential victim depends on numerous factors, but the consequences are immense in any case. Given the increasingly common cases of malicious software (ransomware) being used to block access to data on a computer until a ransom is paid, there is now the possibility of using deepfake ransomware. This software

operates as follows: instead of encrypting an individual's data to restrict access, it first analyzes the data on the computer to learn the user's identity. It then searches for publicly available images or videos of the person on the Internet or the computer itself and uses those media files to create a "deepfake" version of that person in an embarrassing situation. The ransomware would then display this "deepfake" version of the individual on the screen with a visible countdown timer. If the timer expires before the ransom is paid, the "deepfake" version of that person will be publicly released (Grothaus, 2021, p.117). The current development of technology has also enabled discussions about deepfake resurrections, referring to content featuring a deceased individual who appears to be alive and speaks or acts as the content creator desires (Lu & Chu, 2023).

Adults engage in creating deepfake content for various reasons, with women being a particularly vulnerable group. The question arises as to what the situation would look like if children and young people used this technology for cyberbullying. Children quickly adopt new technologies, having access to the Internet and various applications on their mobile phones. The use of deepfake technology does not seem complicated; on the contrary, it is widely available and easy to use because the software does most of the work for the user. It can be concluded that while in the 1990s, peer violence usually ended at the end of the school day, today, with the Internet, smartphones, and social media, victims of peer violence often have no respite from abuse. Peer violence has become digital in the 21st century and can be relentless, following people at all times, and attacking them anytime. Now, thanks to deepfake technology, it can be even more drastic than ever before (Grothaus, 2021, p.19). Although such events have not yet occurred, it is necessary to avoid such scenarios in the future because it cannot be assumed how detrimental they could be to the development of children and young people. Moreover, considering the short time in which deepfake technology has significantly advanced in eliminating previous shortcomings, we can speculate about what the future may hold.

Fake news

Creating fake content on the Internet is not a novelty. The faces of politicians have been previously used to create entertaining content. Today, with the development of deepfake technology, we find ourselves in a situation where we are not sure if the content is fake or not. In this regard, it is necessary to differentiate between shallowfake and deepfake content. Although they have the same goal, unlike deepfakes, which rely on deep machine learning techniques, shallowfakes rely on traditional editing tools

and manual work to alter existing media. They are often called 'cheapfakes' because, due to the required manual editing and human editing skills, some of them may look cheaply made if the creator lacks the necessary expertise in graphics and video editing (Grothaus, 2021, pp.39-40). Therefore, due to the numerous shortcomings of shallowfake content, it is much easier to conclude that it is fake. As Marković and Dimovski emphasize, given the advancement of deepfake technology, "governments worldwide are concerned that material obtained using these applications could be used to undermine national security, spread political propaganda, and disrupt electoral campaigns" (Marković & Dimovski, 2023, p.299). In earlier deepfake videos, there were flaws based on which the observer could conclude that it was fake, which is not the case today. Today, "deepfake video technology - AI techniques for creating real video footage of fake events, often with real people saying things they never actually said - is being used for political purposes" (Schneier, 2021, p.16). This can result in enormous consequences for society and state institutions. "Deepfake videos can have deep negative consequences for democracies: fake news created by deepfake videos may aim to tarnish the reputation of certain individuals, portray false events (e.g., fake terrorist attacks), or influence democratic processes such as election campaigns or other socially significant events" (Meskys et al, 2020, p.8). Faces of many politicians worldwide have been used to create deepfake content, causing concern for many governments. Precisely because of the quality of today's deepfake videos featuring certain influential politicians saying things they never actually said, some countries fear the potential consequences of fake content.

Some political leaders around the world have started using deepfake technology for propaganda purposes, both among the population of their own country and among the populations of other countries. The dangers of propaganda are numerous, as the consequences of using deepfake technology as a geopolitical tool are immeasurable (Grothaus, 2021, p.206). Deepfake technology can be a powerful weapon for causing instability in certain states and regions. Moreover, this technology can be used during international conflicts to: legitimize war; fabricate orders; create confusion; divide armies; undermine support; polarize society; divide allies; and discredit leaders (Byman et al, 2023, pp.6-8). It seems that it has never been easier to disseminate false information and content than it is today. All of this leads us to question whether we can trust our eyes and ears. "We will soon live in a world where we have to wonder 'is it real?' about everything because we will no longer be able to believe that the photos we see, the videos we watch, and the sound we hear are

authentic representations of facts. Even now, with every passing month - just months - deepfake algorithms are becoming increasingly polished and powerful" (Grothaus, 2021, pp.214-215). In the coming years, there is a possibility that deepfake technology will advance to the point where it will be impossible to determine whether the content is fake by any detection method. Research shows that people are overly confident in their ability to detect fake and deepfake content, yet in practice, this is often not the case (Köbis et al, 2021). The availability of deepfake technology, as well as the simple process of creating fake content, already has enormous negative consequences for human rights, democracy, society, state institutions, and national security. From this perspective, what awaits in the future seems daunting considering the speed of deepfake technology development.

Frauds

In the world, there have already been cases of using deepfake technology to carry out certain types of fraud. Due to the possibilities offered by this technology, identity theft is no longer an unexpected phenomenon. Today, there is a significant shift toward biometric authentication for security verifications, and the goal of deepfake content creators is to appear as somebody else to deceive biometric systems (Grothaus, 2021, p.117). With the development of deepfake technology, the creator of deepfake content does not necessarily have to use only someone's face but also their voice. This is not an unexpected occurrence in today's world. The creator of deepfake content can perpetrate numerous scams by using someone else's face or voice. Many applications today require verification in the form of your picture or a recording of your voice. Images are usually available on numerous social media platforms and can easily be abused for these purposes. However, although faking someone else's voice may seem like a slightly bigger problem, that is not the case. Every recording of your voice or the potential availability of your video recordings on social media is what the deepfake content creator needs. After successfully cloning your voice, based on a minute or less of recording, the deepfake creator will be able to create a fake recording of you saying whatever they want you to say (Grothaus, 2021, p.120). There are two main methods for creating deepfake speech: text-to-speech synthesis (which uses written text to produce synthesized speech that sounds like a specific person) and voice conversion (which uses the original voice uttering the desired phrase and a target voice, and the output is the original phrase spoken by the target voice) (Firc et al, 2023, p.14). Identity theft in the form of stealing your face and voice is not the only type of fraud that deepfake content creators can carry out in today's world.

Frauds can be directed toward both individuals and various companies, financial markets, central banks, and financial regulators (Bateman, 2020). Certain attacks that have used deepfake technology and were carried out in the past few years with the aim of gaining huge financial gains are presented in Table 2.

Table 2 – Deepfake attacks launched for financial gain (Kshetri, 2023, p.91).

Year	Victim and location	Type of attack	Amount scammed
2019	Californian women	Deepfake videos	About US $300,000
2019	United Kingdom-based energy company	Vishing	US $234,000
2020	A United Arab Emirates bank	Deepfake voice technology	>US $35 million
2018-2021	Japanese Manga artist	Deepfake videos	>US $500,000

In a comparative perspective, there is a noticeable disproportion between different types of attacks and the targets that have been subjected to these attacks. Vishing, the technique of using fake phone calls to deceive victims and obtain sensitive information, was used during 2019, with the target being an energy company based in the United Kingdom. On that occasion, the perpetrators gained a profit of $234,000. On the other hand, the target of the fraud, with a total profit exceeding $35 million, was a bank in the United Arab Emirates, and the means used to achieve the goal was deepfake voice technology. The use of deepfake technology may not necessarily be aimed at gaining financial profit but can result in reputational damage and loss of trust among stakeholders (Mustak et al, 2023, p.12). Often, malicious software (ransomware) is used to block access to data on a computer until a ransom is paid. Numerous hospitals and healthcare facilities worldwide have been targeted by hackers using this software. This prevents healthcare workers from accessing medical records and other vital patient health-related data. In addition to the potential use of malicious software for various types of extortion and forms of cyber violence, there is also the possibility of using deepfake technology to manipulate CT (computerized tomography) scans. 'Researchers have found that deepfake technology can be used to generate CT-quality images of internal organs that a hacker can then insert into a patient's medical records - replacing authentic CT scans with

deepfaked CT scans' (Grothaus, 2021, p.125). There are several reasons for such types of fraud: the hacker wants to inflict psychological or emotional stress on a specific individual; the hacker does this to gain money and a competitive advantage; and finally, the hacker may compel the doctors of a particular patient to perform a risky operation to try to alleviate the problem, thereby endangering the patient's life (Grothaus, 2021, p.126). Although these cases may seem extreme, they are not impossible today precisely because of the development and availability of deepfake technology. The technology, while neutral in itself, can be abused in various ways depending solely on the individual using it. However, assuming that the development of this technology will continue in the future, with further refinement, there remains a fear of the various other ways malicious individuals worldwide may exploit it.

Conclusion

As an integral part of artificial intelligence, deepfake technology is causing concern among individuals and states worldwide. Despite its positive applications, deepfake technology is often used for negative purposes. The creation of deepfake content impacts the human rights of individuals, national security, and the political systems of countries. Fake videos can cause instability in certain countries and regions. The faces of politicians are used to create fake content and events that never actually happened. Likewise, the faces of famous actresses and celebrities have been used to create pornographic video content that circulates widely on the Internet. With the development of deepfake technology, every person is at risk because this technology allows faces to be used on artificially generated nude bodies. Although every individual is at risk, based on previous practice, it can be concluded that women are a particularly vulnerable and endangered category. Privacy is seriously compromised, and the consequences for victims are immeasurable from various perspectives. Everyone's photos are publicly available on social media and other platforms, resulting in anyone potentially becoming a victim. While it once required a vast number of photos to create fake content, today only one is needed. People's safety is compromised as the development of deepfake technology has provided opportunities for numerous scams. Identity theft is now more possible than ever in history.

A deepfake creator can not only steal someone's face but also someone's voice. It is not just individuals at risk on an individual level, but also countries, various companies, and financial institutions. In recent years, scams have occurred worldwide due to the development of

deepfake technology, targeting both individuals and companies. Furthermore, the use of malicious software and deepfake technology can result in various forms of extortion representing a form of cyberattack. Although there are different methods of detecting deepfake content, it is difficult to prove that a video was created using this technology. It is challenging to trust our eyes and ears in a world where manipulation of public discourse is widespread. Previously, it was possible to detect certain irregularities in a video and realize that the content was fake. Since 2014 and the development of GAN techniques, deepfakes have become more realistic, making it difficult to determine whether the content is true or not. The technique works by improving itself and eliminating flaws, resulting in fantastic results. What should always be kept in mind is that deepfake technology is neither inherently good nor bad. Its usage depends on the individual employing it. However, given that deepfakes are now being used for various negative purposes, the question arises: what can be expected in the future? Therefore, for a comprehensive approach to this researched problem, it is recommended that future research be of a multidisciplinary nature, incorporating not only security but also an information technology approach.

References

Albahar, M. & Almalki, J. 2019. Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22), pp.3242-3250 [online]. Available at: https://www.jatit.org/volumes/Vol97No22/7Vol97No22.pdf [Accessed: 02 March 2024].

Bateman, J. 2020. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios* [e-book]. Washington: Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/files/Bateman_FinCyber_Deepfakes_final.pdf [Accessed: 21 January 2024].

Botha, J. & Pieterse, H. 2020. Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security. In: *Proceedings of the 15th International Conference on Cyber Warfare and Security,* Norfolk, Virginia, USA, pp.57-67, March 12-13.

Byman, D.L., Gao, C., Meserole, C. & Subrahmanian, V.S. 2023. *Deepfakes and international conflict* [e-book]. Washington: Brookings Institution. Available at: https://www.brookings.edu/wp-content/uploads/2023/01/FP_20230105_deepfakes_international_conflict.pdf [Accessed: 21 January 2024].

Firc, A., Malinka, K. & Hanáček, P. 2023. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*, 9(4), e15090. Available at: https://doi.org/10.1016/j.heliyon.2023.e15090.

Grothaus, M. 2021. *Trust No One: Inside the World of Deepfakes*. London, UK: Hodder Studio. ISBN: 9781529347982.

Köbis, N.C., Doležalová, B. & Soraperra, I. 2021. Fooled twice: People cannot detect deepfakes but think they can. *iScience*, 24(11), art.number:103364. Available at: https://doi.org/10.1016/j.isci.2021.103364.

Kshetri, N. 2023. The Economics of Deepfakes. *Computer*, 56(8), pp.89-94. Available at: https://doi.org/10.1109/MC.2023.3276068.

Kugler, M.B. & Pace, C. 2021. Deepfake Privacy: Attitudes and Regulation. *Northwestern University Law Review 611, Northwestern Public Law Research Paper No. 21-04*. Available at: https://doi.org/10.2139/ssrn.3781968.

Lu, H. & Chu, H. 2023. Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts. *Computers in Human Behavior,* 145, art.number:107761. Available at: https://doi.org/10.1016/j.chb.2023.107761.

Mania, K. 2024. Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study. *Trauma, Violence, & Abuse,* 25(1), pp.117-129. Available at: https://doi.org/10.1177/15248380221143772.

Marković, D. 2022. *Detektovanje manipulacije u video snimcima stvorenih „deepfake" tehnikom sistemom učenja prostorno vremenskih karakteristika*. PhD thesis. Belgrade: Singidunum University [online]. Available at: https://nardus.mpn.gov.rs/handle/123456789/20763 (in Serbian) [Accessed: 02 March 2024].

Marković, E. & Dimovski, D. 2023. Deepfake as a new form of crime. In: *XII International conference on social and technological development,* Trebinje, Republic of Srpska, Bosnia & Herzegovina, pp.296-308, June 15-18 [online]. Available at: https://stedconference.com/wp-content/uploads/2024/02/Book-of-Proceedings_STEDC_2023-sa-DOI-brojevima_compressed.pdf (in Serbian) [Accessed: 02 March 2024].

Meskys, E., Liaudanskas, A., Kalpokiene, J. & Jurcys, P. & 2020. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice*, 15(1), pp.24-31. Available at: https://doi.org/10.1093/jiplp/jpz167.

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. & Dwivedi, Y.K. 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154, art.number:113368. Available at: https://doi.org/10.1016/j.jbusres.2022.113368.

Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.Ti., Nguyen, D.Th., Nuynh-The, T., Nahavandi, S., Nygen, T.T., Pham, Q.- V. & Nygen, C.M. 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223, art.number:103525. Available at: https://doi.org/10.1016/j.cviu.2022.103525.

Prlja, D., Gasmi, G. & Korać, V. 2022. *Human rights and artificial intelligence* [e-book]. Belgrade: Institute of Comparative Law (in Serbian). Available at: http://ricl.iup.rs/1295/1/Ljudska%20prava%20i%20ve%C5%A1ta%C4%8Dka%20inteligencija-Prlja-Gasmi-Korac.pdf (in Serbian). [Accessed: 21 January 2024]. ISBN: 978-86-80186-82-5.

Schneier, B. 2021. *The Coming AI Hackers* [e-book]. United States: Harvard Kennedy School, Belfer Center for Science and International Affairs. Available at: https://www.belfercenter.org/sites/default/files/2021-04/HackingAI.pdf [Accessed: 21 January 2024].

Vatreš, A. 2021. Deepfake Phenomenon: An Advanced Form of Fake News and Its Implications on Reliable Journalism. *Društvene i humanističke studije DHS (Social Sciences and Humanities Studies),* 6(3), pp.561-576 (in Serbian). Available at: https://doi.org/10.51558/2490-3647.2021.6.3.561.

Van der Sloot, B. & Wagensveld, Y. 2022. Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review*, 46, art.number:105716. Available at: https://doi.org/10.1016/j.clsr.2022.105716.

Wang, S. & Kim, S. 2022. Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, 134, art.number:107305. Available at: https://doi.org/10.1016/j.chb.2022.107305.

Westerlund, M. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11), pp.39-52. Available at: https://doi.org/10.22215/timreview/1282.

Wazid, M., Mishra, A.K., Mohd, N. & Das, A.K. 2024. A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact. *Cyber Security and Applications*, 2, art.number:100040. Available at: https://doi.org/10.1016/j.csa.2024.100040.

El lado oscuro de la tecnología deepfake

Sanela Z. Veljković[a], **autor de correspondencia**,
Milica T. Ćurčić[a], *Ilija* P.Gavrilović[b]

[a] Universidad de Belgrado, Instituto de Ciencias Nucleares 'Vinča' - Instituto Nacional de la República de Serbia, Belgrado, República de Serbia

[b] Universidad de Belgrado, Facultad de Ciencias Políticas, Belgrado, República de Serbia

CAMPO: ciencias de la computación, TI, seguridad, inteligencia artificial
TIPO DE ARTÍCULO: artículo de revisión

Resumen:

Introducción/objetivo: La inteligencia artificial puede utilizarse tanto con fines positivos como negativos. En los últimos años, el uso de la tecnología deepfake ha atraído una atención significativa. La tecnología deepfake reemplaza el rostro de una persona y crea eventos que nunca sucedieron. Si bien el uso de deepfake era más notorio en el pasado, la tecnología ha

avanzado tan rápidamente que hoy es imposible determinar si el contenido es falso o no. Como resultado, hay erosión de la confianza en los medios y las instituciones políticas, manipulación del discurso público, así como la difusión de desinformación y noticias falsas. El objetivo de este trabajo es examinar los métodos de creación de contenido deepfake y explorar las posibilidades de detectar dicho contenido. Se pone atención especial en investigar el lado oscuro de la tecnología deepfake, es decir, los fines negativos para los que se puede utilizar la tecnología deepfake.

Métodos: Mediante el uso de métodos de revisión de bibliografía y análisis de contenido, este trabajo ha proporcionado una sistematización del conocimiento sobre la tecnología deepfake, así como un análisis de datos relevantes en este campo con respecto al posible mal uso de la tecnología deepfake. La tecnología deepfake y su uso se analizan desde una perspectiva de seguridad, es decir, cómo el uso de estas tecnologías puede representar un peligro social. Las investigaciones futuras deben diseñarse para ser multidisciplinarias, integrando el conocimiento de las ciencias sociales (seguridad, sociología, psicología) y las ciencias técnicas (tecnología de la información).

Resultados: Los resultados de esta investigación muestran que, en un contexto positivo, el uso de deepfake se asocia con la medicina, la industria cinematográfica, el entretenimiento y los esfuerzos creativos. Sin embargo, deepfake se utiliza a menudo para crear contenido pornográfico, pornografía de venganza, noticias falsas y varios tipos de fraude.

Conclusión: La tecnología deepfake es neutral en el sentido de que el propósito de su uso depende del individuo que crea el contenido. El uso tanto de inteligencia artificial como de tecnología deepfake plantea cuestiones legales y éticas complejas. Aunque estas tecnologías ofrecen un potencial notable de mejora social, la tecnología deepfake supone al mismo tiempo un grave riesgo para los derechos humanos, la democracia y la seguridad nacional. Por tanto, el uso indebido de las tecnologías deepfake representa un peligro social para toda la población de cualquier país. Las mujeres son especialmente vulnerables debido a la posibilidad de crear contenido pornográfico y pornografía vengativa mediante la tecnología deepfake, aunque las víctimas de este acto también pueden ser hombres.

Palabras claves: deepfake, contenido pornográfico, ciberviolencia, estafas, noticias falsas.

Темные стороны дипфейк технологии

Санела З. Велькович[а], **корреспондент**,
Милица Т. Чурчич[а], *Илья* П. Гаврилович[б]

[а] Белградский университет, Институт ядерных исследований «Винча» − Институт государственного значения для Республики Сербия, г. Белград, Республика Сербия

[б] Белградский университет, факультет политических наук, г. Белград, Республика Сербия

РУБРИКА ГРНТИ: 28.23.00 Искусственный интеллект
ВИД СТАТЬИ: обзорная статья

Резюме:

Введение/цель: Искусственный интеллект можно использовать как в положительных, так и в отрицательных целях. В последние время использование технологии дипфейк привлекает большое внимание. С помощью технологии дипфейк можно заменить лицо определенного человека и создавать события, которых никогда не было. Раньше использование дипфейков было гораздо заметнее, но технологии за короткий промежуток времени настолько прогрессировали, что на сегодняшний день практически невозможно отличить фейковый контент от настоящего. В результате образовались: эрозия доверия к средствам массовой информации и политическим институтам, возможность манипулирования общественным дискурсом, а также распространение дезинформации и фейковых новостей. Целью данной статьи является изучение способов создания дипфейкового контента, а также исследование возможностей обнаружения такого контента. Особое внимание уделено исследованию темной стороны технологии дипфейк и негативным целям, в которых ее можно использовать.

Методы: На основании тщательного обзора литературы и контент-анализа в данной статье представлена систематизация знаний о технологии дипфейк, а также анализ соответствующих данных в этой области относительно возможности злоупотребления технологии дипфейк. Технологии дипфейк и их использование рассматриваются с точки зрения безопасности, то есть исследуется, насколько они опасны и какую социальную угрозу может представлять использование этих технологий. Рекомендуется, чтобы будущие исследования имели мультидисциплинарный характер, интегрируя знания социальных наук (безопасность, социология, психология) и технических наук (информационные технологии).

Результаты: Результаты данного исследования показывают, что в положительном контексте использование дипфейка

связано с медициной, киноиндустрией, развлечениями и творчеством. Однако дипфейки часто используются для создания порнографического контента, порномести, фейковых новостей, а также для осуществления разных видов мошенничества.

Выводы: Технология дипфейк сама по себе нейтральна в том смысле, что цель ее использования зависит от человека, создающего конкретный контент. Однако использование как искусственного интеллекта, так и технологии дипфейк поднимает сложные юридические и этические вопросы. Хотя потенциал, который эти технологии обеспечивают для улучшения общества, заметен, технология дипфейк одновременно представляет собой серьезный риск для прав человека, демократии и национальной безопасности стран. Ввиду вышеизложенного, недобросовестное использование дипфейковых технологий представляет социальную опасность населения каждой страны в целом. Женщины являются особо уязвимой категорией населения из-за возможности создания порнографического контента и порномести с использованием технологии дипфейк, однако жертвами могут стать и мужчины.

Ключевые слова: дипфейк, порнографический контент, кибернасилие, мошенничество, ложные новости.

Тамне стране дипфејк технологије

Санела З. Вељковић[а], **аутор за преписку,**
Милица Т. Ћурчић[а], *Илија* П. Гавриловић[б]

[а] Универзитет у Београду, Институт за нуклеарне науке „Винча” – Национални институт Републике Србије, Београд, Република Србија

[б] Универзитет у Београду, Факултет политичких наука, Београд, Република Србија

ОБЛАСТ: рачунарске науке, ИТ, безбедност, вештачка интелигенција
КАТЕГОРИЈА (ТИП) ЧЛАНКА: прегледни рад

Сажетак:

Увод/циљ: Последњих година употреба дипфејк технологије привлачи огромну пажњу. Њоме се могу креирати лажни ликови као и догађаји који се никада нису десили. Ранијих година била је препознатљива, али је у последње време толико напредовала да је данас немогуће проценити да ли је неки садржај лажан или не. Резултат тога јесте недостатак поверења у медије и политичке институције, могућност манипулисања јавним дискурсом, као и ширење дезинформација и лажних вести. Циљ овог рада јесте

испитивање начина креирања дипфејк садржаја, као и истраживање могућности за његово откривање. Посебно су истраживане могућности злоупотребе дипфејк технологије.

Методе: Посредством коришћења литературе и анализе садржаја, у раду су систематизована знања о дипфејк технологији и анализирани релевантни подаци у овој области у вези са могућностима злоупотребе. Дипфејк технологија и њена употреба сагледане су и са безбедносног аспекта, односно показано је на који начин она може представљати друштвену опасност. Препорука је да будућа истраживања буду мултидисциплинарна – да интегришу знања друштвених наука (безбедност, социологија, психологија) и техничких наука (информационе технологије).

Резултати: Резултати овог истраживања показују да се дипфејк технологија примењује медицини, филмској индустрији, забави и креативном стваралаштву. Међутим, неретко се користи како би се креирали порнографски садржаји, лажне вести и извеле одређене врсте превара.

Закључак: Дипфејк технологија јесте неутрална, што значи да сврха њене употребе зависи од појединца који креира одређени садржај. Употреба како вештачке интелигенције, тако и дипфејк технологијенамеће сложена правна и етичка питања. Иако је уочљив потенцијал који ове технологије пружају унапређењу друштва, дипфејк технологија истовремено представља озбиљан ризик по људска права, демократију и националну безбедност држава. Посебно угрожену категорију становништва представљају жене због могућности креирања порнографских садржаја и осветничке порнографије употребом дипфејк технологије, иако жртве могу бити и мушкарци.

Кључне речи: дипфејк, порнографски садржај, сајбер насиље, преваре, лажне вести.

Paper received on: 04.03.2024.
Manuscript corrections submitted on: 25.09.2024.
Paper accepted for publishing on: 26.09.2024.