DOI: https://doi.org/10.32782/PPSS.2023.1.8

DEEPFAKES AND SYNTHETICALLY REPRODUCED MEDIA CONTENT AS A FORM OF DISINFORMATION IN THE CONTEXT OF THE RUSSIAN AGGRESSION AGAINST UKRAINE

Hanna CHEMERYS,

PhD (Education) Associate Professor,
Head of Department of Design, Faculty of Social Pedagogy and Psychology,
Zaporizhzhia National University, Ukraine
https://orcid.org/0000-0003-3417-9910

AI technology manipulates videos by blending real and fake content, enhancing their realism and persuasiveness compared to videos generated entirely by AI. For instance, deepfake applications can interchange faces in a genuine video or modify lip movements, making it seem as though someone is saying something different from the original recording. Our research deals to analysis of the impact of deepfakes during the war, specifically delving into their usage in the initial stages of the Russian invasion of Ukraine. Although we couldn't document every instance of deepfakes, our focus was on identifying the most significant examples that had a substantial impact [1-5].

As an illustration, before the invasion of Ukraine by Russia in 2022, the United States exposed a Russian scheme to employ deepfake videos as a pretext for invading Ukraine [6]. Among the various online instances of deepfake usage during the Russo-Ukrainian war. Following the invasion, Ukrainian government officials cautioned about the potential dissemination of deepfakes by Russia depicting Ukrainian President Volodymyr Zelenskyy surrendering [7]. This apprehension seems to have materialized when hackers manipulated a Ukrainian news website to display a video purportedly showing President Zelenskyy. On video President Zelenskyy asking Ukrainian soldiers to lay down their weapons [8] (Figure 1).

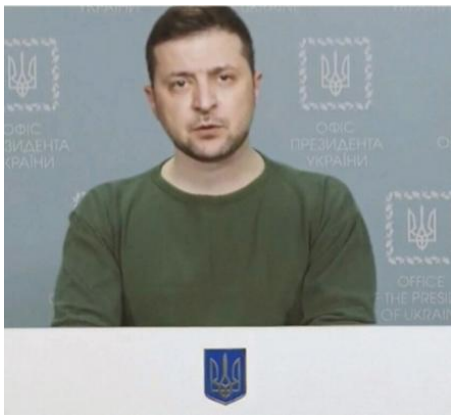

Figure 1: Deepfake Video. President V. Zelenskyy. Source: Twitter. Unknown creator.

This case highlighted the potential use of deepfakes, combined with compromised media services, to disseminate misleading messages. The consequence of this incident was the circulation of false information originating from a seemingly reliable source. This video give away a lot of signs about using deepfake technology: blurry outlines, flickering of the face (one of the obvious things, since some of these videos still look unnatural - this applies to the transitions between the face, neck and hair, which are not always organically combined with each other), unnatural facial expressions, especially when blinking and low video quality, which is often used to hide incorrect neural network operation. This case highlighted the potential use of deepfakes, combined with compromised media services, to disseminate misleading messages. The consequence of this incident was the circulation of false information originating from a seemingly reliable source.

Another sample of deepfake technology is video with V. Zaluzhnyi in November 2023 (Figure 2).

Figure 2: Deepfake Video. V. Zaluzhnyi.
Source: Unknown

This video is also fake: desynchronization of facial expressions and words, change of voice (not Zaluzhnyi's voice), pixelation in the forehead area due to the superimposition of another face, unrealism of ears that change position at different intervals of the video, and, of course, low quality of the video itself to hide the editing.

Comparable deepfake videos depicting russian president vladimir putin surrendering also surfaced amid the conflict.

We are now living in an era when "I saw with my own eyes" - this is not an indicator of credibility. All information must be analyzed. To approach everything with a cool head, and not lose vigilance due to emotions. Of course, it is necessary to analyze different sources to compare information, fact-checking should be done everywhere and always. And just in the case of the authorities, it is easier to do this, because the presence of such persons in the media space is greater, so it is easier to do fact-checking by comparing speeches and researching rhetoric.

It is more difficult to identify fakes from non-media persons, because fact-checking in this case is more difficult. Therefore, it is worth being aware of the basic markers for identifying deepfakes, and here the rule "Trust no one" should work, even if the markers are not found. It is better to question the information

than to fall for a possible deepfake. After all, not only generative video can be fake. There is a high probability of staging. We can recall the case with the "crucified boy" (The story was first published by Aleksandr Dugin on 9 July 2014) [9]. That video was not a deepfake. It was a real person who said that text on camera. Here it is necessary to add common sense to the analysis of information.

However, examples of the use of synthetically recreated media content during the war were not only for the purpose of disinformation. For example, on January 15, the official channel of the Verkhovna Rada posted a picture generated by artificial intelligence on the topic of the terrorist attack in Dnipro on January 14 (Figure 3). Previously, images from Midjorney with a neutral message appeared on the channel: Ukrainian symbols, silhouettes of Ukrainian defenders, etc. However, it was this incident that touched the audience for life. The picture shows a small crying boy with scratches on his face against the background of a destroyed high-rise building. The image is very realistic: at first, it can be confused with a photo with too much retouching. It was this believability that blew the audience away. After a dozen comments, the account deleted the post. When Ukrainian media and diplomats make maximum efforts to convey to the Western audience the consequences of Russian aggression, the realistic image of war generated by artificial intelligence can become a space for manipulation, so it is safe to assume that the active development of AI art can cause a new kind of fake images, and therefore, and a new level of information warfare [4].

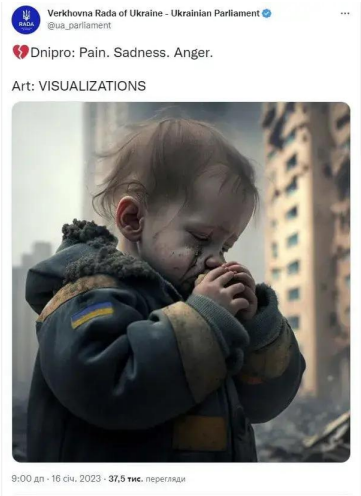

Figure 3: Generative Image on the Topic of the War in Ukraine. Source: official channel of the Verkhovna Rada

The above-mentioned examples of artificial intelligence, although not all of them were aimed at disinformation, had a clear negative impact on society. This incident highlights the utilization of (deep)fakes for disinformation objectives. Instances like the aforementioned underscore the hazardous nature of this type of disinformation, with the objective of heightening existing conflicts, undermining trust in state institutions, and inciting emotions and anger in general. The erosion of trust is likely to complicate law enforcement efforts.

Here are examples of the positive aspects of using generative images during the war. With the onset of a full-scale war in Ukraine, the Sirens Gallery project appeared, created on the basis of an open-source model by the Ukrainian startup ZibraAI (Figure 4) [10]. This project is designed to draw attention to the war and help raise funds for the reconstruction of cities destroyed by russia, and to remind the world that the war continues. At the beginning of work on it, we conducted a little research and chose as a basis the approach implemented in Disco Diffusion. Disco Diffusion is based on the class-conditional diffusion model from OpenAI together with CLIP, which connects text queries with images.

The project developers researched the parameters, available models, and artistic styles, chose the option that best suited the theme, and began generating paintings. For better images, super-resolution was added to the pipeline and a convenient interface was made for internal needs [11]. However, in reality, more time and effort of the development team was spent not on working on the technological part, but on the timeline of the war. To select the most important events, and then write stories about them, many events and photos were processed. In total, there are about 2,000 pictures generated by artificial intelligence based on textual descriptions of the most important events of the war. 1,991 paintings were put up for sale as NFTs on the Opensea.io and Paras.id platforms (already sold for over 250,000 UAH).

a) b)

Figure 4: Sirens Gallery Project:
a) Ghost of Kyiv; b) Heroes of Azovstal hold defense of Mariupol for 85 days.

Our research on the impact of deepfakes during the early stages of the Russian invasion of Ukraine reveals significant challenges. While not capturing every instance, we focused on notable examples, such as Russia's alleged plan to use deepfakes as a pretext for invasion and the subsequent manipulation of a video showing President Zelenskyy. This incident exemplifies the disinformation threat, highlighting the dangers of (deep)fakes in intensifying conflicts and eroding trust.

Observers and analyze emphasize that deepfake technology facilitates the rapid and effortless creation of fake videos compared to previous methods. Our research uncovered numerous instances where the presence of deepfakes led to skepticism or confusion. Notably, our data revealed cases where authentic videos were mistakenly accused of being fake. Additionally, we observed instances of people losing trust in all videos related to the conflict, entertaining theories that suggested world leaders were replaced by deepfakes.

Beyond disinformation, our study underscores the broader impact of deepfakes. The technology's ability to blend real and fake content creates convincing videos, causing skepticism, confusion, and even the mistrust of authentic footage. Notably, the Zelensky video falsely claiming the war's end exemplifies how deepfakes can spread misleading messages from seemingly reliable sources. However, our research also recognizes the dual nature of AI-generated content. While deepfakes serve disinformation goals, projects like the Sirens Gallery use generative images to draw attention to the war's consequences positively. As society grapples with the challenges posed by synthetic media, cultivating critical thinking and a discerning approach to information becomes crucial in navigating the complexities of the digital age.

The work is performed within the research CEFRES, UAR 3138 CNRS–MEAE. Research topic "Development of Methodology of Critical Thinking and Pedagogical Support to Counteract Disinformation and Manipulation of Artificially Reproduced Media Content" (2023).

References:

1. Borges L., Martins B., Calado P. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*. 2019. Vol. 11. No. 3. Pp. 1-26.

2. Chawla R. Deepfakes: How a pervert shook the world. *International Journal of Advance Research and Development*. 2019. Vol. 4. No. 6. Pp. 4-8.

3. Chemerys H. Truth & Trust in the Age of Deepfakes: Recognize & Overcome. *Українські студії в європейському контексті: зб. наук. пр.* Київ : ГО "Інноваційні обрії України", 2023. № 7. Pp. 403-407 DOI: https://doi.org/10.31110/2710-3730/2023-7

4. Chemerys H., Briantseva H. V., Briantsev O. A. The Urgency of the Problem Synthetically Reproduced Media Content. *International scientific conference "Interaction of culture, science and art in terms of moral development of modern European society" : conference proceedings*, December 28–29, 2021. Riga, Latvia : "Baltija Publishing", 2021a. Pp. 85-88. DOI: 10.30525/978-9934-26-178-7-20

5. Day C. The future of misinformation. Computing in Science & Engineering. 2019. Vol. 21. No. 1. Pp. 108 -108.

6. CBS News, 'U.S. reveals Russian plot to use fake video as pretense for Ukraine invasion', 2022, URL: https://www.cbsnews.com/news/russia-disinformation-video-ukraine-invasion-united-states/. (accessed on 10 March 2022)

7. Metro, 'Ukraine warns Russia may deploy deepfakes of Volodmyr Zelensky surrendering', 2021, URL: https://metro.co.uk/2022/03/04/ukraine-warns-russia-may-deploy-deepfakes-of-zelensky-surrendering-16217350. (accessed on 15 May 2022)

8. National Public Radio, 'Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn', 2021, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia (accessed on 17 February 2022)

9. Snyder T. (12 April 2022). "The sadness of Sloviansk". Thinking about... (accessed on 12 April 2022)

10. Zibra.AI, Sirens, 2022. URL: https://sirens.gallery/ (accessed on 12 December 2022)

11. Ragot M., Martin N., Cojean S. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 2020. Pp. 1-10