Malicious Use of Deepfakes and Political Stability

Evgeny Pashentsev
Institute of Contemporary International Studies, Diplomatic Academy, Ministry of Foreign Affairs of the Russian Federation, Moscow, Russia; School of International Relations, Saint Petersburg State University, Saint Petersburg, Russia
icspsc@mail.ru

Abstract

In many cases, deepfakes have become a source of joy, fun and innocent tricks, as well as being a perfect example of the way impressive advances in technology have helped people, with respect to education, health care and other areas. However, the malicious use of deepfakes (MUD) is a current social problem which threatens to do much more damage to individuals, democratic institutions, social stability and international security in the future. The author aims to analyze the current practice of MUD, its role in the destabilization of national psychological security and political stability and possible ways of detecting and neutralizing MUD. The systematic approach adopted in this study seeks to avoid one-sided assessments, rather looking to give an objective analysis of the interrelated aspects of the development and use of deepfakes. The unity of opposites is the central category of dialectics. In dialectical contradictions, the content of the opposites forming them is not strictly deterministic or unambiguous because they are in a state of change and self-development. The application of the dialectical method suggests that any phenomenon should be considered in the duality of its properties and characteristics to find contradictions and mutual disposition (e.g. objective connection, unity, dependence). This is particularly important for our case, since we are dealing with dual-use technology which, because of the nature of social relations in modern society, can be deliberately used to either benefit or harm people. The paper proves that due to the widespread availability of internet technologies, the cheapness, variety of forms, and the growing quality of deepfakes, they pose a growing threat to democracy and political stability. However, there is a wide range of socially useful applications of deepfakes which, in our opinion, makes it undesirable to apply a complete ban on the production and distribution of deepfakes. Through deepfakes (in the context of other measures), false targets can be identified for criminals and the safety of potential victims of criminal actions is thereby increased.

Keywords: Deepfakes, artificial intelligence, politics, psychological warfare, political stability, ethics for AI.

1. Introduction

Political stability is very important for the successful development of any country. The topic of political stability has attracted the attention of researchers for many years (Ake, 1975; Dowding and Kimber, 1983; Bulut and Yildirim, 2020, etc.). New factors that can disrupt political stability, including the malicious use of advanced technologies, are particularly dangerous, since it is initially difficult to foresee their consequences. Especially if such an aspect of political destabilization is a quantitative and qualitative increase in the ability to manipulate public consciousness.

The ability to bewitch people and make them see things that do not really exist is reflected in many myths and legends. In a fantastic story by Lino Aldani (1964), the invention of ‘onirofilm’ meant a death sentence for the regular movie. In the story, having turned on the projection device, the viewer perceived smells and touch to such an extent that they merged with what was happening and found themselves ‘inside’ the film. They were no longer a passive spectator, but the main character. Wonderful dreams of love, power, and glory became the only needs as everything else in life lost its meaning. The 17th-century philosopher René Descartes imagined a scenario wherein he might be deceived by a demon. This evil demon is imagined to present a complete illusion of an external world (see more on this in Descartes, 2014). Nearly 400 years later in *Simone* (a 2002 US science fiction film) this idea was implemented in an unexpected way. According to the plot, the shooting of the film is put in jeopardy when the actress, who was supposed to play the main role, refuses to shoot. The Director decides to replace her with a digital actress, Simone, created with the help of a computer program. However, everyone takes the artificial girl to be the real one and admires her work. 15 years later, a user of a social news platform (Reddit), known as "deepfakes," started posting pornographic celebrity videos using face-swapping technology, including Gal Gadot, Scarlett Johansson, Aubrey Plaza, and Taylor Swift (Merrefield, 2019). Doctored imagery is neither new nor rare but deepfakes creates fakes that are high-quality, cheap, and quickly-produced and, as a result, very effective. The distribution of deepfakes can be supported by other AI technologies.

In September 2019, researchers at the cyber-security company Deeptrace found 14,698 deepfake videos online, compared with 7,964 in December 2018. They said 96% were pornographic in nature, often with a computer-generated face of a celebrity replacing that of the original adult actor in a scene of sexual activity. At the same time other forms of non-pornographic deepfakes have gained popularity (Ajder et al., 2019, p. 1). Advances in software have made it much easier to create deepfakes than before. Researchers at the Samsung AI Center in Moscow developed a way to create "living portraits" from a very small dataset (as few as one photograph, in some of their models (Zakharov et al., 2019, p. 1).

The rising turbulence in international relations and political instability in many countries makes any new tool for distorting information about the world around us, such as malicious use of deepfakes (MUD), very dangerous for society. Therefore, it requires careful study and an adequate means of countering, including those based on AI technologies.

An additional risk of MUD arises in connection with the coronavirus pandemic because it is accompanied by a major downturn in the world economy, which could reach the scale of Great Depression and increase socio-political tensions and panic. In such a difficult environment, the skillful use of higher-quality, faster and easier to create deepfakes can become a serious threat to political stability.

The coronavirus pandemic has encouraged the transition to distant forms of working, based on modern information technologies. Working from home and the growth of AI implementation, in addition to the undoubted advantages, gives new opportunities for malicious use of artificial intelligence (MUAI), including the sphere of deepfakes.

Whilst working on this paper, the author relied on a wide range of academic literature which examined the history of deepfakes (Westerlund, 2019), their nature, options and risks associated with use (Tech, 2018; Patriau and Patriau, 2019; Delfino, 2019; Ajder et al., 2019; Pindrop, 2020). Special attention was paid to those studies that focused on the political risks of spreading deepfakes (Chesney and Citron, 2019; Young, 2019; Adams, 2019).

This paper hypothesises that, in the near future, MUD could become a dangerous tool for destabilizing political systems in different countries within the framework of integrated use. The systematic research approach seeks to avoid one-sided assessments, instead looking to give an objective analysis of the interrelated aspects of the development and double-use of deepfakes.

The application of the dialectical method suggests that any phenomenon should be considered in the duality of its properties and characteristics. This is particularly important for our case, since we are dealing with dual-use technology which, because of the nature of social relations in modern society, can be deliberately used to either benefit or harm people.

The study is structured as follows: the meaning of the term deepfake is analyzed in the narrow and broad sense of the word, then the role of deepfakes in MUAI is studied and, finally, the use of MUD in politics. The main conclusions on the topic are then formulated.

2. Understanding of deepfakes and their role in the MUAI

In its narrowest sense, deepfake is a way of adding a digital image or video over another image or video, so that it appears to be part of the original. A deepfake is an image or video that has been changed in this way (Collins, 2020). The term *deepfake* is typically used to refer to a video that has been edited using an algorithm to replace the person in the original video with someone else (usually a public figure) in a way that makes the video look authentic (Merriam-Webster, (2020). Deepfakes are not only images and sounds but texts as well. Mika Juuti, a doctoral student at Aalto University in Finland, and a team of researchers developed a new way to make algorithmically generated reviews more believable. The study, presented at the European Symposium on Research in Computer Security in September 2018, asked participants to read real reviews written by humans and fake machine-generated reviews. The researchers then asked the participants to identify the fake ones. Up to 60 percent of the fake reviews were mistakenly thought to be real (Cole, 2018).

The methods that now fall under the deepfake umbrella include face swaps, audio deepfakes (copying someone's voice), deepfake puppetry or facial re-enactment (mapping a targets face to an actor's and manipulating it), deepfake lip-synching (a video of someone speaking created from audio and footage of their face) (Vincent, 2018), fake machine-generated texts, targeted image transformation, and others. The methods used are increasing in number.

Without rejecting the differences in specific technologies, we believe it is possible to use the term deepfake in the broader sense of the word, i.e. when the term combines a set of current and future technologies for constructing pseudo-reality, which are based on the growing capabilities of artificial intelligence to create or modify images, sounds, and texts.

Human ingenuity will, no doubt, conceive many beneficial uses for deepfake technology. For now, the most obvious possibilities for beneficial uses fall under the headings of education, art, and healthcare. Perhaps most notably, deep-fake audio technology holds the promise to restore the ability of persons suffering from certain forms of paralysis, such as ALS, to speak with their own voice (Chesney and Citron, 2019). Good results have been achieved in education, e.g. museum activities (Braunstein, 2018) etc.

Unfortunately, there are not only positive uses of deepfakes but also increasing cases of MUD. Some misuse is isolated in nature, the other part integrating perception management without the use of other AI technologies. The potentially most dangerous cases include MUD as part of malicious use of artificial intelligence (MUAI) within perception management and high–tech psychological warfare (HTPW). Deepfakes do not exhaust the possibilities of MUAI to create a distorted view of reality. Here we should pay attention to the possibly malicious activities of *AI- based chatbots* (Balch, 2020) or the use of *prognostic weapons* (Pashentsev, 2016). For example *lie detectors* enhanced with the use of AI, could cause dissonance in public opinion if their data is poisoned and they can be used against democratic institutions. *Sentiment analysis* provides a very accurate analysis of the overall emotion of the text content incorporated from sources like blogs, articles, forums, and surveys. Sentiment analysis is an opinion-mining process from the perspective of computer linguistics (Azizan and Abdul Aziz, 2017).

Deepfakes are not good or bad in themselves. It is only a technology that expresses people's constructive or destructive (conscious or unconscious) intentions. They should not be demonized but nor should the real threat of their use for anti-social purposes be ignored.

History suggests that those working to manipulate the media are often one step ahead of those working to protect against such manipulation. As social media platforms change, so too will the ways in which computer-based propaganda is spread. The technology that is costly today is likely to be cheap and easy to use tomorrow. Moreover, disinformation campaigns are nearly impossible to put back in the box once they have been launched (Wolley, 2020).

Our understanding of deepfakes as a *set of technologies* for the construction of *pseudo-reality*, which are based on the growing capabilities of artificial intelligence, may be helpful for further research in this area. Furthermore, an understanding of deepfakes as being only *one of the sets of MUAI technologies* is rather important for researching deepfakes and MUAI as a whole, as well as for individuals and society, which have to defend themselves against MUAI *in all its forms,* including MUD.

See more: Pashentsev, E. Malicious Use of Deepfakes and Political Stability, Proceedings of the 2nd European Conference on the Impact of AI and Robotics, a virtual conference hosted by Instituto Universitario de Lisboa (ISCTE-IUL), Portugal, 22-23 October 2020 (ed. F. Matos), Academic Conferences and Publishing International Limited, Reading, UK, pp. 100–107. https://www.academic-conferences.org/conferences/eciair/

https://www.academic-bookshop.com/ourshop/cat_1643278-2020-Conferences.html