SWP Comment

NO. 52 OCTOBER 2023

Deepfakes – When We Can No Longer Believe Our Eyes and Ears

Media Manipulation in Conflict: Challenges and Responses

Aldo Kleemann

Deception and media manipulation have always been an integral part of wartime propaganda. But never before has it been so easy to create high-quality fabrications of images as well as sound and video recordings. The tendency to react emotionally to these media opens up a whole new possibility for abuse by their creators. A call to surrender by President Volodymyr Zelensky, which was immediately exposed as a deepfake, is the first attempt to use the new technology in an armed conflict. The quality of such fabrications is improving, detecting them is becoming increasingly complex and there is no end in sight to these developments. Banning deepfakes would be futile. It is therefore time to look at current and potential applications and possible counter-strategies.

It is well known that truth is one of the first casualties of war, and that propaganda is used in conflicts. Conventional means of spreading disinformation include (social) media, political organisations, cultural associations, foundations and think tanks. Although the attribution of disinformation is sometimes difficult, Russia and China are undoubtedly among the most important actors in this field. Another instrument has been added to their toolbox recently by means of generative artificial intelligence (AI): deepfakes — artificially created or altered photos as well as video or voice recordings that look and sound deceptively real. Many first-generation deepfakes were easy to spot because of image flaws or tinny voices. High-quality fabrications were rare, which is why in 2020 the German government rightly deemed deepfakes to be of "low practical relevance" and did not feel compelled to develop a dedicated response strategy.

Today, easy-to-use AI tools capable of producing high-quality fabrications are freely available. Deepfakes have become a common commodity rather than a rarity. Three developments in particular have contributed to this in recent years: the continuous improvements to AI, the steady increase in available computing power and access to increasing amounts of data with which AI can be trained. An end to this process, and the full extent of its consequences, is difficult to foresee.

Stiftung Wissenschaft und Politik
German Institute for International and Security Affairs

How are deepfakes created?

In contrast to cheapfakes — whereby existing recordings are manually or digitally spliced together, slowed down or speeded up — the integration of AI allows for the automated creation or modification of media products. For example, faces or speech can be changed and still be lip-synched; gestures and facial expressions can be altered; or entire speeches can be invented and seemingly spoken by a person. To do this, two neural networks are combined in a GAN (Generative Adversarial Network) and trained using existing images as well as video and speech recordings. Especially in the case of public figures, the required data is often available in large quantities. The subsequent deep learning of the neural networks is so profound, and the results so realistic, that the now colloquial term “deepfake” can be traced back to this intensive process. In a GAN, there is an interplay between two components. While the design element (generator) creates fictitious images or voices, the other element (discriminator) evaluates them for authenticity against the given training dataset. The goal is for the generator to produce media that are as indistinguishable as possible from the training dataset. This process can be continuously improved by, for example, adjusting the dataset and the weighting of the selection criteria, or by bringing in real people to assist with the discriminator component. It can be expected that the ability to detect deepfakes will be drastically compromised with the continuous improvement of the above parameters.

Recent examples of deepfakes

Synthetic videos of Barack Obama and Angela Merkel show what is possible with the appropriate training data. Whereas the creators of the former video show President Obama insulting President Trump, the creators of the latter video have the German Chancellor giving a speech in verse about the behaviour of German citizens during the Corona pandemic. Both videos were produced to highlight the dangers of deepfakes and are labelled as such. Currently, realistic-looking deepfake images of well-known personalities are appearing, especially on Twitter, without always being identified as such. These include Pope Francis posing like a rapper in a down jacket and Donald Trump being arrested, kissing President Putin, or hugging and kissing the Chinese flag. New examples are added almost daily — some to entertain, some to warn and some to deceive. On 22 May 2023, a report of an explosion at the Pentagon was circulated on Twitter. The tweet gave the impression that it was an official report from the news agency Bloomberg. It was accompanied by an image showing black smoke over the Pentagon — a deepfake that was quickly spotted, but it was enough to cause the S&P 500 index to briefly drop by around 30 points.

All that is needed to create such images is an AI application such as Midjourney or Stable Diffusion and a precise description of the image you want to create.

Evolution of disinformation campaigns

The utilisation of generative AI to produce deepfakes changes the usage of such fabrications in disinformation campaigns in three fundamental ways:

- **Quantity** — Commercially available applications allow deepfakes to be produced en masse, quickly and cheaply. This allows not only states, but also resource-poor groups and individuals to run their own disinformation campaigns on a large scale.
- **Quality** — Deepfakes are improving in quality and appear more natural, making them harder to detect and increasing their credibility and persuasiveness.
- **Qualification** — Although the creation of deepfakes requires almost no skill, the expertise required to detect them is becoming more extensive.

These developments have the potential to significantly increase the reach and effectiveness of disinformation in the 21st century.

Possible application of deepfakes in conflicts

In March 2022, the first notable attempt to use a deepfake in an armed conflict was made. After the website of the television channel Ukraine 24 was hacked, a video of President Zelensky appeared, in which he declared: “There is no tomorrow. At least not for me. Now I have to make another difficult decision: To say goodbye to you. I advise you to lay down your arms and return to your families. It is not worth dying in this war.” Ukraine was prepared for such a deepfake attack. Within minutes, a real video of the president’s response was recorded and circulated on social media. The poor quality of the deepfake, the speed with which it was detected and rebutted, and the ability to distribute the real video via a largely stable internet connection all contributed to the failure of the fake appeal to surrender. However, these conditions will not always be present in future conflicts. Moreover, as deepfake technology develops, limitations to the production and effective use of deepfakes will no longer be technical in nature, but exclusively linked to the question of creativity.

Paralysis — Deepfakes could be used in the form of fabricated evidence to paralyse or divide allies. Such an approach was proposed in the run-up to the invasion of Ukraine. US security experts, for example, suggested that Russia was planning to produce fake video evidence of Ukrainian war crimes against Russian communities to justify the attack on Ukraine. Such video evidence would have been suitable to start a discussion in European states about the legitimacy of Russia crossing the border to protect Russian minorities, which could have prevented an immediate reaction in favour of Ukraine. In this specific case, it was not the use of deepfakes that was suspected, but rather a traditional fake using props and actors. Deepfakes might facilitate the fabrication of such scenes in the future. The creation or alteration of eyewitness testimony and allegedly authentic recordings of orders issued in violation of international law can already be generated today.

Mobilisation — Deepfakes could also be used to mobilise populations against security forces. Existing ethnic, cultural, social or religious fault lines within and between societies could be exploited. For example, the maltreatment of religious symbols could be faked by creating photos or videos of desecrations or by falsifying eyewitness accounts. The potential for the real or perceived maltreatment of religious symbols to mobilise people was illustrated by the riots following the Muhammad cartoons controversy in 2005 and the burning of Korans by US forces in Afghanistan in 2012.

Subversion — Deepfakes could be used to create fear and uncertainty. Fake videos of political or military leaders making calls to surrender or disdainful remarks about their own forces killed in action, or questioning the sense and purpose of the military operation would likely demoralise the armed forces. Similarly, the fabrication of atrocities committed by their own forces against the civilian population could be used to undermine popular support for the armed forces. In addition, massive amounts of very graphic images and audio recordings illustrating the horrors of war could be produced to prevent mobilisation of the population and encourage desertion.

Recommendations for action

Deepfakes are here to stay. The incentive to create convincing media content that fits one’s own narrative quickly and cheaply is simply too great. This is already evident today — outside armed conflicts — in democratic discourse. In both Germany and the

United States, parties and their supporters in domestic political debates are already resorting to deepfakes to reinforce their messages. Moreover, the underlying AI technology offers a whole range of positive applications in addition to the negative ones listed above.

A silver bullet, that is, a simple and universally applicable weapon against deepfakes, will never exist. Current assessments of the potential and limitations of generative AI are, of course, based on a snapshot in time. The pace of development in this area of technology has repeatedly surprised even the experts. Moreover, it is unclear what the capabilities of the AI models currently being developed by private and public actors are, and what restrictions they are subject to.

As a result, many of the approaches tackling this issue are either very specific and tailored to individual cases or — in order to keep up with the rapidly changing dynamics — have to be designed more broadly and are likely to require continual adjustments. What is needed is a mix of preventive and reactive measures to limit the impact of deepfakes.

Preventive approaches and their limitations

Preventive approaches aim to raise the barriers to using of deepfakes and limit their potential impact from the outset.

Reducing the number of actors and controlling them — The creation of deepfakes requires specialised software and hardware. Access to these resources is a possible starting point for regulatory measures to reduce the number of actors capable of creating deepfakes and control them. Some approaches currently under discussion include export restrictions on hardware components and restrictions on access to computing power, training data as well as ready-to-use AI models.

An example here is the restriction on the export to China of semiconductors and other materials required to build supercomputers, introduced by the United States in October 2022. Such a restriction on the hardware side may help slow the growth of computing power serving as the basis for generative AI models. However, there is a significant need for regulation, as not only direct exports but also indirect supplies via third countries need to be considered in order to effectively implement such a restriction.

If a state does not have its own computers, recourse to cloud computing is an easy way to circumvent export restrictions on hardware supplies. It is therefore sometimes suggested that access to cloud computing power should be restricted. In practice, such a regulation is difficult to implement. On the one hand, almost all global cloud providers would have to be covered by the restrictions; on the other hand, it is difficult to determine whether leased computing power is being used for a climate simulation or for training AI.

Similarly, it is difficult to enforce a restriction on access to training data such as images and videos. It is true that the volume of training data has a significant impact on the performance of AI, and limiting it is therefore in principle a good way of reducing the number of actors who are capable of training a powerful generative AI application. However, this would also require the agreement of all stakeholders. Moreover, it is questionable whether such a regulation could be effectively implemented for data that is freely available on the internet.

Once an AI model has been trained, developers decide how the model can be used and who has access to it. This results in some access control options that can actually be implemented effectively. However, they are only effective as long as a large number of providers participate, and as long as there are no open source alternatives to these AI models.

Mandatory labelling — Mandatory end-user labelling, as currently envisaged by the EU in Article 52 of the Artificial Intelligence Act, is not likely to reduce the number of deepfakes. A software-based labelling requirement would be better. This would ensure that the common, freely available AI applications in Europe would only produce recognisable deepfakes. Such labels could be removed, but the expertise required to do so would limit the number of people who could produce a deepfake without a label.

Awareness-raising — Knowledge of deepfakes and their possible application can be used to help develop a more critical approach to audiovisual media. With a view to conflicts and crises, this knowledge should be promoted especially among political leaders and authorities and organisations with security tasks. However, such an approach must not presuppose that it will continue to be possible to detect deepfakes with the naked eye and without technical aids.

Promoting trustworthy content — Some ideas are not directly aimed at detecting deepfakes, but at creating transparency, and thus facilitating the dissemination of authentic and trustful audio and visual recordings. One such approach is the Content Authenticity Initiative. The companies involved — including the BBC, Nikon, Reuters and Adobe — are attempting to create cross-platform industry standards that will enable the origin of digital content to be securely authenticated. The aim is to attach tamper-proof identity and history data to recordings so that the authorship and any modifications to the files can be tracked permanently. As a result, the standard creates transparency in terms of the distribution process, but the information value is limited to the origin and any subsequent modification of the file. There is no guarantee that the recording itself is an authentic representation of reality.

Examining potential applications — In order to effectively counter the use of deepfakes in times of crisis and conflict, security authorities must also examine the potential applications of the technology itself. In the United States, there are warnings about the dangers that deepfakes pose for democracy, but at the same time the Special Operations Command is intensively exploring how the technology can be used for military purposes. Such an ongoing debate could take place in Germany, the EU and NATO within existing structures:

- The interministerial working group on hybrid threats, which has been working under the leadership of the Federal Ministry of the Interior since 2019, and the associated task force on disinformation is a suitable format for pooling the experience of different departments.
- Under the auspices of the Federal Ministry of Defence, the Centre for Operational Communication monitors the information space and is already examining the impact of propaganda on the armed forces.
- Exchanges with EU and NATO partners could take place through the Helsinki-based European Centre of Excellence for Countering Hybrid Threats and the Riga-based NATO Strategic Communications Centre of Excellence.

Reactive approaches and their limitations

Reactive approaches aim to reduce the impact of a deepfake that has already been released. In an age where information spreads in minutes rather than days, the ability to quickly detect and respond to a deepfake is essential.

Technical detection — The sheer variety of ways to manipulate media makes it unlikely that automated detection — in the sense of a one-size-fits-all solution — will be available in the foreseeable future. Moreover, the economic incentive to create better and better deepfakes is currently much higher than the incentive to work on techniques to detect them. The state must counteract this by specifically promoting media forensic expertise. The bandwidth of detection methods is vast, ranging from an individual recording of the facial expressions and speech rhythms of high-ranking leaders, for example, to the collection of power grid fluctuations in order to verify the time and place of a recording or identify the equipment used. In order to make it difficult for a potential attacker to use a convincing deepfake effectively, it is crucial that the detection methods are varied and, in some cases, kept secret. Otherwise, the discriminator in a GAN will be continually adapted to the known detection methods in order to evade them.

Response strategy — An effective response strategy encompasses many of the points already mentioned: a general awareness of the issue, a constant level of engagement with the issue of deepfakes, combined with media monitoring and the ability to rapidly identify and assess potential fabrications technically. Then there is the need for tried and tested procedures: within the government, between departments, but also with partners in the EU and NATO.

The fact that Ukraine was able to expose the deepfake of the alleged Zelensky speech so quickly was partly due to the fact that the president was one of the most intensely monitored people in the media in March 2022, and also that the Ukrainian authorities had anticipated the use of deepfakes. In addition to Ukraine, there are also other states dealing with persistent disinformation. In Taiwan, which has repeatedly been the target of Chinese news manipulation, government ministries are required to respond to misinformation within 60 minutes of it being released — a timeframe that Germany should also adopt, given the speed at which disinformation spreads. This requires a hitherto unseen level of adaptability and reaction speed on the part of state institutions, whose stages of progress are usually assessed in years, if not decades, rather than minutes and hours.

Lieutenant Colonel (G.S.) Aldo Kleemann is a Visiting Fellow in the International Security Research Division at SWP.

This Comment reflects
the author's views.

The online version of
this publication contains
functioning links to other
SWP texts and other relevant
sources.

SWP Comments are subject
to internal peer review, fact-checking and copy-editing.
For further information on
our quality control procedures, please visit the SWP
website: https://www.swp-berlin.org/en/about-swp/quality-management-for-swp-publications/

SWP
Stiftung Wissenschaft und
Politik
German Institute for
International and
Security Affairs

Ludwigkirchplatz 3 — 4
10719 Berlin
Telephone +49 30 880 07-0
Fax +49 30 880 07-100
www.swp-berlin.org
swp@swp-berlin.org

ISSN (Print) 1861-1761
ISSN (Online) 2747-5107
DOI: 10.18449/2023C52

(English version of
SWP-Aktuell 43/2023)