COUNTER EXTREMISM PROJECT

KONRAD ADENAUER STIFTUNG

Deep Fakes

On the Threat of Deep Fakes to Democracy and Society

Prof. Dr. Hany Farid
and Dr. Hans-Jakob Schindler

www.kas.de

Deep Fakes

On the Threat of Deep Fakes to Democracy and Society

Prof. Dr. Hany Farid and Dr. Hans-Jakob Schindler

Key messages

The present study is the result of a cooperation between the Konrad-Adenauer-Stiftung and the Counter Extremism Project. The authors, Prof Dr. Hany Farid and Dr. Hans-Jakob Schindler, deal with the destructive potential of so-called deep fakes – videos and images altered by artificial intelligence (AI) misused for political manipulation.

- Manipulated images and videos have already been posing major challenges to journalism, science, jurisdiction and politics in the past. New technology, however, has made the production of deep fakes widely available to the public – we are experiencing a democratization of deep fakes. Deep fakes used in disinformation campaigns can cause social cohesion and thus be a threat to democracy.

- Social media play an increasingly central role in informing the public and have become the main distribution platform of fake news. These platforms are operating unregulated and with different standards, presenting a particular challenge in the fight against disinformation campaigns.

- Germany is at an early stage of this new challenge. There is still enough time to develop an effective defense mechanism against the threat of deep fakes.

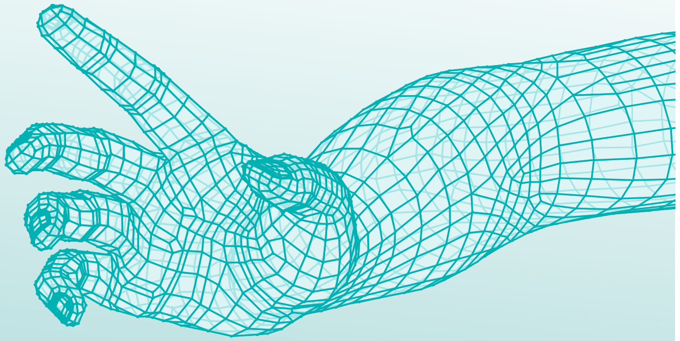
- The fight against this complex challenge requires a multipronged approach which combines technical solutions with legal and public education measures.

Authors

Prof. Dr. Hany Farid, Professor, University of California, Berkeley; Senior Advisor, Counter Extremism Project

Dr. Hans-Jakob Schindler, Senior Director, Counter Extremism Project

Table of contents



1. Introduction

Deep fakes, videos manipulated by artificial intelligence (AI) misused for political manipulation, are taking dis- and mis-information campaigns aimed at attacking open democracies and societies to a new level. This technology is only the latest development in a long history of image manipulation techniques that emerged when photography first gained widespread use. In the past, the manipulation of videos, which are generally perceived to be an accurate representation of reality, was the near exclusive domain of high-tech film studios and special-effects companies. Advances in technology, however, are lowering these technical hurdles and therefore opening up the possibility for the adoption of this technology by the general public, including nefarious state and non-state actors.

This technology, in combination with the global reach of social-media platforms, allows nefarious actors to launch far more sophisticated political influence campaigns that have the potential to reach large segments of the population. This is particularly the case as an increasing number of German citizens consume news via social media.[1] Fortunately, Germany – as compared to the United States – has not yet become a strategic target of such campaigns. However, there is no reason to assume that Germany will remain exempt from this development in the near future. Therefore, strategic discussion of this phenomenon should begin.[2]

While there may be still time for Germany to react, this study will outline that developing an effective defense mechanism against such threats will take time and should involve a multi-pronged approach. In order to contribute to this emerging debate, the Konrad-Adenauer-Stiftung partnered with the Counter Extremism Project to prepare the present report.

This report consists of five sections. Following the introduction, Chapter 2 provides a brief history of photo manipulation, demonstrating that the modification of images has a long history and has today become a fairly common practice that presents a challenge for media, legal procedures, political debates, as well as national security assessments.

Chapter 3 will then focus on synthesized content using AI. Such content constitutes what is now commonly referred to as “deep fakes”. This section will outline both the current state of play regarding the creation as well as the detection of deep fakes, highlighting that high-quality detection remains a considerable challenge. This section will also include a brief outlook on future creation and detection mechanisms, warning that the increasing democratization of this technology will present a growing risk.

Chapter 4 will outline the various incidents in which deep fakes have been used for criminal purposes and with the intent of exercising political manipulation. This section warns that in times of political uncertainty, the mere existence of this technology is enough to potentially cause political disruption, even if it is not taken advantage of.

Finally, in Chapter 5, the report describes the various elements of a risk mitigation strategy. Such a strategy should involve legal measures. These could include restrictions on the use of such technology as well as ensuring the confidentiality of methods used to identify deep fakes as well as new legal definitions that declare illegal the misuse of such technology when serving political manipulation. Because deep fakes misused for the purpose of political manipulation only reach their corrosive potential if they are circulated widely, the distribution mechanisms – social media platforms – must be involved in this endeavor. Technology-based elements of a defense system designed to prevent the misuse of deep fakes for political manipulation should encompass certification of original content and support for the ongoing development of detection technologies. Finally, an effective system will also engage public education and seek, for instance, to increase cyberliteracy at the secondary school level.

[1] According to the most recent online study conducted by ARD and ZDF, more than 90% of all Germans over the age of 14 use internet services, in particular social media platforms. ARD/ZDF Online Studie 2019, http://www.ard-zdf-onlinestudie.de/files/2019/ARD-ZDF-Onlinestudie-Grafik-2019.pdf [06.05.2020]. Already in 2016, around 31% of all German respondents said that they used social media applications for their news consumption. This number very likely has increased in the last 4 years, see: S. Hölig, U. Hasebrink, Nachrichtennutzung über soziale Medien im internationalen Vergleich. In: Media Perspektiven 11/2016, pp. 534–548, https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2016/11-2016_Hoelig_Hasebrink.pdf [06.05.2020].

[2] N. Lossau, Deep Fake: Gefahren, Herausforderungen und Lösungswege, Konrad-Adenauer-Stiftung, Analysen & Argumente Nr. 382/2020, https://www.kas.de/documents/252038/7995358/AA382+Deep+Fake.pdf/de479a86-ee42-2a9a-e038-e18c208b93ac?version=1.0&t=1581576967612 [06.05.2020].

2. A brief history of photo-manipulation

History is haunted by the specters of photographic fakery. Stalin, Mao, Hitler, Mussolini, Castro, and Brezhnev all had photographs manipulated in attempts to alter history. For this, they relied on intricate and time-consuming darkroom techniques. Starting in the early 1990s, however, powerful and low-cost digital technology made it far easier for nearly anyone to alter digital images. And the resulting fakes are often very difficult to identify. Over the past two decades, this photographic fakery has affected many different areas.

Media: Adnan Hajj was famous for producing striking war photographs from the ongoing struggle in the Middle East. On 7 August, 2006, the *Reuters* news agency published one of Hajj's photographs showing the aftermath of an Israeli bombing of a Lebanese town (shown in the lower panel of illustration 1).[3] In the week that followed, hundreds of bloggers and nearly every major media outlet reported that the photograph had been doctored; additional smoke had been added. The reactions were marked by outrage and anger – Hajj was accused of manipulating the image in order to exaggerate the impact of the Israeli attack. An embarrassed *Reuters* quickly withdrew the photograph and removed from its archives nearly 1,000 photographs contributed by Hajj. The case of Hajj is, of course, by no means unique. In 2003, Brian Walski, a veteran war photographer, manipulated a photograph that ran on the front page of the *Los Angeles Times*. After learning of the fake, the outraged editors of the *LA Times* fired Walski. The news magazines *Time* and *Newsweek* have each been rocked by similar scandals after it was revealed that photographs appearing on their covers had been doctored. Over the past few years, countless news organizations around the world have been shaken by similar experiences.

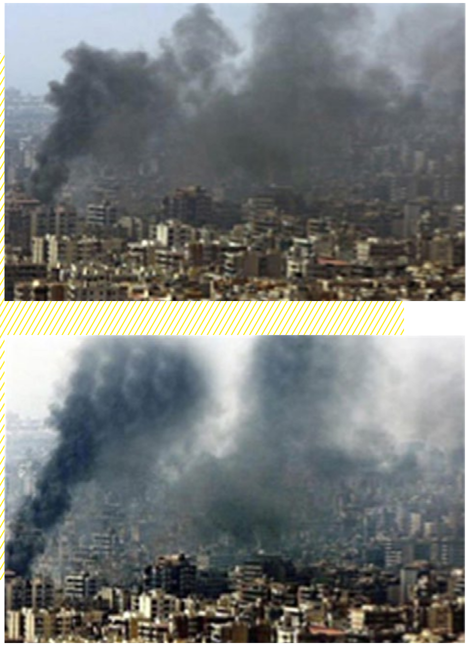


Illustration 1: The original (top) and manipulated (below) picture of an Israeli bombing of a Lebanese town. Picture: Adnan Hajj/Reuters

Science: Journalists are not the only ones tempted to manipulate photographs. In 2004, Professor Hwang Woo-Suk and colleagues published what appeared to be ground-breaking advances in stem cell research. The paper was published in one of the most prestigious scientific journals, *Science*. Evidence slowly emerged that the presented results were manipulated and/or fabricated. After months of controversy, Hwang withdrew the *Science* paper and resigned from his university position. An independent panel investigating the accusations found, among other things, that at least nine of the eleven customized stem cell colonies that Hwang had claimed to have cultivated were fakes. Much of the evidence for those nine colonies, the panel said, involved doctored photographs of two other, authentic, colonies. While this case attracted international coverage and provoked widespread outrage, it is by no means unique. In an increasingly competitive field, scientists are increasingly tempted to exaggerate or fabricate their results. Mike Rossner, editor in chief of *Cell Biology,* estimates that as many as 20% of the manuscripts submitted to his journal contain at least one figure in need of revision due to inappropriate image manipulation.[4]

Law: The child pornography charges against its police chief shocked the small town of Wapakoneta, Ohio. At court, the defendant's lawyer argued that if the state could not prove the authenticity of the seized images, the defendant's possession of the images was not illegal. In 1996, the Child Pornography Prevention Act (CPPA) amended existing federal criminal legislation against child pornography to include certain types of "virtual porn". In 2002, the United States Supreme Court found that sections of the CPPA, being overly broad and restrictive, violated First Amendment rights. The Court ruled that possession of "virtual" or "computer-generated" images depicting a fictitious minor does not violate the constitution. According to this logic, the burden to prove that the images are real and not computer-generated lies with the state. Given the sophistication of computer-generated images, several state and federal rulings have further found that juries should not be asked to distinguish real from virtual images. And at least one federal judge questioned the ability of even expert witnesses to make this determination.

Politics: “Fonda Speaks to Vietnam Veterans At Anti-War Rally”, read the headline, accompanied by a photograph purportedly showing Senator John Kerry sharing a stage with the controversial actress and anti-war activist Jane Fonda (Illustration 2).[5] Both the article and the photo were fakes – the latter was a composite of two separate and unrelated photographs. And in 2008, just days after being selected as a running mate to U. S. presidential hopeful John McCain, doctored images of a bikini-clad and gun-toting Sarah Palin circulated widely on the internet. The idea to pair one’s political enemies with controversial figures is certainly not new. It is believed that a doctored photograph contributed to Senator Millard Tydings’ electoral defeat in 1950. The photo of Tydings conversing with Earl Browder, a leader of the American Communist party, was meant to suggest that Tydings had communist sympathies. During the recent primary elections, political ads have involved a startling number of doctored photographs showing candidates in a flattering or damaging light.

Illustration 2: A photo composite of Senator Kerry and anti-war activist Jane Fonda. Picture: Ken Light, Owen Franken/The Guardian

Illustration 3: A photo composite in which the third missile from the left which failed to fire was digitally inserted. Picture: Sepahnews/The Guardian

National Security: As tensions mounted between the United States and Iran in 2008, the Iranian Government announced the successful testing of ballistic missiles. As evidence, the government released a photograph showing the simultaneous launch of four missiles. Shortly after its world-wide release, it emerged that the image had been doctored. In truth, only three missiles had launched, while the fourth missile, which had failed to launch, had been added digitally (Illustration 3).[6] This example made it clear how crucial access to authentic news images is, and highlighted the geo-political shockwaves potentially caused by fake photographs.

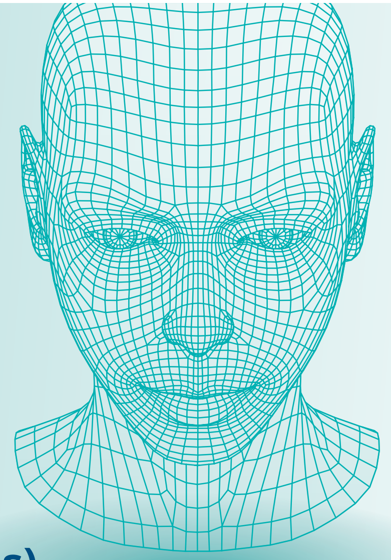
While historically they may have been the exception, over the past two decades doctored photographs have increasingly impacted nearly every aspect of society. Over the past few years, advances in AI and Machine Learning (ML), along with access to massive datasets and computing power, have brought us a great deal closer to the next revolution in digital manipulation.

[3] Photo created by Adnan Hajj (original and altered). Altered Photo originally distributed by Reuters.

[4] H. Pearson, Image manipulation: CSI: Cell biology. In: Nature 434/2005, pp. 952–953, https://www.nature.com/articles/434952a.pdf?proof=true&draft=collection%3F-proof%3Dtrue [06.05.2020].

[5] The picture is a combination of two original photos. The original photo of Senator Kerry was taken by Ken Light in 1971. The original photo of Jane Fonda was taken by Owen Franken in 1972.

[6] The altered picture was originally published by the Iranian news organisation Sepahnews.com. The picture is currently no longer accessible on the website. At the time, Associated Press also published the picture of Illustration 3.



3. AI-synthesized content (a. k. a. deep fakes)

3.1 Creating deep fakes

Recent advances in computer graphics, computer vision, and ML have made it easier to automatically synthesize compelling fake audio, image, and video. In the audio domain, highly realistic audio synthesis is now possible. A neural network, fed with enough sample recordings, can learn to synthesize speech based on a user's voice.[7] In the static image domain, highly realistic images of people can now be synthesized.[8] And in the video domain, highly realistic videos can be created of anybody saying and doing just about anything that their creator wants.[9]

Our focus will be on deep fake videos. Such manipulated videos fall into one of three categories: (1) face swap: the face in a video is automatically replaced with another person's face. Illustration 4 is an example of this technique, where the face of an impersonator gets replaced with that of Hillary Clinton.[10] This type of technique has been used to insert famous actors into a variety of movie clips in which they never appeared, or to create non-consensual pornography in which one person's likeness in an original video is replaced with another person's likeness; (2) lip sync: a source video is modified to make the mouth region consistent with an arbitrary audio recording. The actor and director Jordan Peele has produced a particularly compelling example of such media by altering a video of former U. S. President Barack Obama to depict him saying "President Trump is a total and complete dip shit."; and (3) puppet master: a target person is animated (head movements, eye movements, facial expressions) by a performer sitting in front of a camera acting out what they want their puppet to say and do.

Illustration 4: A Hillary Clinton impersonator (left) and a face-swap deep fake (right). Picture: Saturday Night Life/National Broadcasting Company (NBC)

There is an abundance of techniques for creating these types of deep fake videos, including *DeepFakes FaceSwap* and *FSGAN, Neural Textures, Face2Face,* and *FaceSwaps*, most of which are easily accessible open source projects.[11] The most common – but by no means exclusive – approach to creating deep fake videos (or images) leverages the power of generative adversarial networks (GAN). A GAN is composed of two main components, a generator and a discriminator. The generator's goal is to synthesize each video frame to be consistent with the distribution of a training dataset. The discriminator's goal is to determine if the synthesized video frame can be detected as belonging to the training dataset or not. The generator and discriminator work iteratively, eventually leading the generator to learn to synthesize a video – frame by frame – that fools the discriminator.

The popular *FaceSwap* software, for example, uses a GAN to create so-called face swap deep fakes in which one person's likeness in a video is replaced with another person's likeness. This approach has been popularized by, for example, adding the actor Nicholas Cage to movies in which he never appeared, including his highly entertaining appearance in *The Sound of Music*. While this technique can generate highly convincing fakes, it often requires a significant amount of training data. *FaceSwaps* uses a similar approach to generate deep fakes, similarly requiring large amounts of training data. The more recent method, *FSGAN*, on the other hand, creates high-quality fakes with less training data.

Neural Textures is a generic image synthesis framework that combines traditional graphics rendering with more modern learnable components. This framework can be used for novel-view synthesis, scene editing, and the creation of so-called lip-sync deep fakes which modify a person's mouth to be consistent with diverging audio input. This work generalizes earlier work that was designed to create lip-sync deep fakes on an individual basis.

Unlike these GAN-based methods, some methods rely on more traditional computer-graphics approaches to create deep fakes. *Face2Face*, for example, allows for the creation of so-called puppet master deep fakes in which one person's (the master's) facial expressions and head movements are mapped onto another person (the puppet). Similarly, *FaceSwap* builds a three-dimensional facial model of one person which is then mapped onto the recording of another person. These techniques allow users to generate facial expressions in real time using standard consumer cameras. A related technique, puppet mastering, is able to build a photo-realistic avatar GAN that synthesizes facial expressions and orientations in real time on mobile devices.

3.2 Detecting deep fakes

The field of digital forensics has produced ample research on detecting traditionally manipulated images and videos.[12] Here we focus only on techniques for detecting the types of deep fake videos described above. There are three broad categories for detecting deep fake videos: (1) manual review, (2) low-level computational techniques, and (3) high-level computational techniques.

Early deep fake videos were marked by obvious artifacts, including blur and glitches. Faked subjects lacked typical functions such as blinking.[13] While an experienced eye can also identify elaborate deep fakes, it is becoming increasingly difficult to visually distinguish real recordings from fakes. As the quality of deep fakes has continued to improve, therefore, it is becoming necessary to build on automatic and computational techniques to detect deep fake videos.

Low-level approaches focus on detecting pixel-level artifacts introduced by the AI synthesis process. One such approach uses a convolutional neural network (CNN) to detect pixel-level artifacts produced during the process of transposing one face onto another. Another learning-based approach trains a twin neural network to find inconsistencies between the image and the camera metadata (e.g., focal length, ISO, aperture, exposure time, etc.). An image is then authenticated using this network to determine if each image patch is consistent with the same imaging pipeline. Although not necessarily focused on deep fakes, *ManTra-Net* (Manipulation Tracing Network)[14] uses end-to-end training of a CNN to detect and localize different types of image manipulation, including splicing, removal, and copy-move. Another approach focuses on detecting and localizing facial manipulations by using a network to holistically classify a face as manipulated or not. A second network exploits low-level features in small patches to determine if a face region is consistent with the rest of the image; a final prediction is generated by combining these two predictions. Other approaches have shown that GAN-generated content contains distinct digital fingerprints which can be identified and used to classify images as GAN-generated or not.

The advantage of these and similar low-level approaches is that they can automatically extract artifacts and differences between synthetic and real content. The drawback is that they can be highly sensitive to intentional or unintentional laundering, including resizing or transcoding, as well as adversarial attacks and when extrapolated to novel datasets.

By contrast, the high-level approaches described next tend to be more resilient to these types of manipulation and are likely to be more robust when generalizing to novel datasets.

High-level approaches focus on more semantically meaningful features. For example, previous work recognized that the creation of face swap deep fakes introduces inconsistencies in the head pose, as the head pose is extrapolated from the central, transposed portion of the face and the surrounding, original head. These inconsistencies leverage 3-D geometry, and are currently difficult for synthesis techniques to correct. Because training data sets often do not include depictions of people with their eyes closed, it has also been observed that early face swap deep fakes were marked by an unusually low frequency of eye blinks. More recent deep fakes, however, seem to have fixed this problem. A related technique exploits spatial and temporal physiological traits that are not consistently recorded in real video footage and disrupt face swap deep fakes. Other research analyzed hours of video recordings of specific individuals (in this case, various world leaders and U. S. presidential candidates) in order to extract distinct and predictable patterns of facial expressions and head movements. Related research found that lip sync deep fakes failed to accurately model the mouth when transposing certain sounds (phonemes).

The advantage of these high-level approaches is that unlike low-level approaches, they are more robust against laundering attacks and can easier contribute to identifying a large variety of deep fakes, from face-swapping to lip-syncing to puppet-mastering. The drawback of these approaches is that they can require more effort to develop, test, and deploy.

Despite efforts by digital forensic researchers to develop both low- and high-level forensic techniques, no technology exists that can contend with the vast array of different types of deep fakes at a speed and accuracy that can be deployed at internet-scale.

There are several challenges that the digital forensic community is facing. Deep fakes are a relatively new phenomenon. In terms of their sophistication, they have developed much faster than expected. There are significantly more researchers working on synthesizing ever more realistic audio, image, and video data than there are those trying to identify such content. This means that the nature and quality of deep fakes is developing at an unprecedented rate that is difficult to keep pace with. In addition, the scale and speed of the internet makes deploying effective technology incredibly challenging: Facebook, for example, registers around one billion uploads daily,[15] and some 500 hours of video are uploaded to YouTube every minute.[16] The sheer amount of information uploaded every day makes employing effective filtering technology incredibly difficult.

There is a family of technologies that could be widely deployed. **Control-capture technologies** can confirm content integrity by extracting, at the time of recording, a unique digital signature from any recorded digital content, cryptographically signing this signature, and then placing it on a secure central server or a distributed immutable ledger, such as a blockchain. This signature can then be compared to any version of the same content found online to determine if the content has been modified. Although this approach tackles disinformation differently than forensic techniques – by telling us what is real instead of what is fake – these technologies are available today and can operate at internet scale. Both these control-capture and classic forensic techniques merit further exploration.

3.3 Future of deep fake creation and detection

Today, a fair amount of computing power is required to create a long and visually compelling deep fake video. While this type of computing power is readily available in the cloud, it does have its barriers. However, the underlying software for creating deep fake videos is readily and freely available online. The emerging trend in the creation of deep fakes is that the quality continues to increase, while the amount of required data and computing power decreases. In addition, commercial products and websites have begun to appear, only accelerating the technology's democratization.

At the same time, and as we will discuss below, the channels for distributing deep fakes through social media remain readily accessible. And as the general public, inundated every day with misinformation, conspiracies, and lies, struggles to make sense of the world around, we see the emergence of the "liar's dividend":[17] as fake information overwhelms the online ecosystem, anyone can claim that news that they don't like, or that doesn't corroborate their world view, is simply fake. It is perhaps these distribution and consumption patterns that pose larger threats to democracy and society than the fake content itself.

[7] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior,K. Kavukcuoglu, Wavenet: A generative model for raw audio, 2016, https://arxiv.org/abs/1609.03499 [06.05.2020].

[8] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410; T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen,T. Aila, Analyzing and improving the image quality of stylegan, 2019, https://arxiv.org/abs/1912.04958 [06.05.2020].

[9] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deep-fakes and beyond: A survey of face manipulation and fake detection, 2020, https://arxiv.org/abs/2001.00179 [06.05.2020].

[10] Original video on the left is a screen grab for an episode of the popular comedy show Saturday Night Life of the Nationa Broadcasting Company (NBC) in 2016. The deep fake picture on the right was generated for illustrative purposes.

[11] See for example: Britt Paris, Joan Donovan, Deepfakes and Cheap Fakes, p. 15; https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf [27.05.2020]

[12] H. Farid, Photo Forensics. Cambridge, MA/London: MIT Press. 2016.

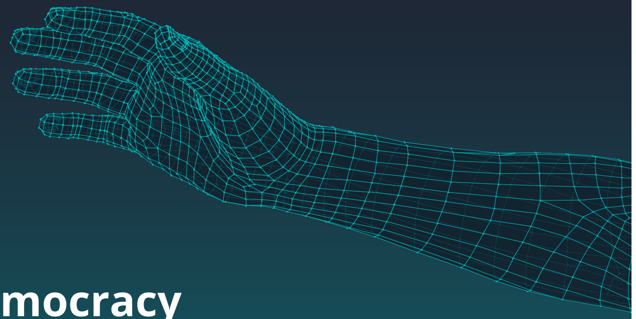
[13] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: IEEE International Workshop on Information Forensics and Security, 2018, pp. 1–7, https://arxiv.org/pdf/1806.02877.pdf [06.05.2020].

[14] Y. Wu, W. Abdalmageed, P. Natarajan, ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, https://openaccess.thecvf.com/content_CVPR_2019/papers/Wu_ManTra-Net_Manipulation_Tracing_Network_for_Detection_and_Localization_of_Image_CVPR_2019_paper.pdf [06.05.2020].

[15] D. Noyes, The Top 20 Valuable Facebook Statistics – Updated January 2020, Zephoria, https://zephoria.com/top-15-valuable-facebook-statistics/.

[16] J. Hale, More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute, Tubefilter, 7 May 2019, https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. [05.06.2020]

[17] R. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security. In: California Law Review 107/2019, pp. 1753–1819, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [06.05.2020].



4. Threats to democracy and society: From state to non-state actors

4.1 Deep fakes misused for criminal purposes

The malicious use of deep fakes presents a threat when used by state as well as non-state actors. The first and currently most prevalent use of deep fake videos is their use for criminal purposes, especially in the creation of non-consensual pornography.[18] This continues to pose a threat to all women, particularly celebrities, and journalists, but also those who simply attract unwanted attention. In response, several U. S. states have recently passed legislation to minimize the harm posed by such content, and similar legislation is being considered at the U. S. federal and international levels.

In addition, the democratization of sophisticated technology to synthesize highly realistic fake audio, image, and video promises to complicate the struggle against dis- and mis-information campaigns designed to commit fraud. A striking example of this occurred in March 2019, when fraudsters used synthetic voice impersonation to mimic a company's chief executive officer in a phone call and demand a fraudulent transfer of EUR 200,000.[19] Less sophisticated deep fakes involve the creation of synthesized photos of non-existing individuals[20] to create artificial identities in order to conduct fraud and espionage.[21] Finally, experts have also highlighted that deep fakes can be misused to exacerbate local grievances and conflicts.[22]

4.2 Threat of deep fakes misused for political manipulation

The main emphasis of this report concerns the misuse of deep fakes to disrupt democratic elections and sow civil unrest. Cases of attempted manipulation have risen continually over the last two years. In January 2019, the worldwide threat assessment of the United States intelligence community identified deep fakes as one of the major global threats:

> *Adversaries and strategic competitors probably will attempt to use deep fakes or similar machine-learning technologies to create convincing – but false – image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners.*[23]

Prominent cases of attempted political manipulation in 2019 involved fairly low-quality deep fakes. For example, during the recent British parliamentary elections at the end of 2019 misinformation, including deep fakes, seems to have been strategically used to damage the prospects of the Leader of the Liberal Democrats and a prominent Labor politician.[24]

In the United States, two manipulated videos of the current speaker of the U. S. House of Representatives, Nancy Pelosi, circulated on social media. Without using sophisticated technology, the videos were slowed down slightly to make the speaker seem drunk.[25] Although the manipulation was easy to identify, the video was shared among the speaker's political opponents even after being flagged as manipulated.[26] Interestingly, Facebook, one of the major platforms on which the video was shared, refused to take the video down.[27]

The challenges arising from weak online monitoring, even among major social media platforms, are showcased by Twitter's certification of a fake U. S. Congress candidate in February 2020. In December 2019, Twitter announced that it would certify candidates for the 2020 U. S. elections[28] and that it would cooperate with Ballotpedia, a not-for-profit organization that maintains a database of political candidates.[29] Apparently, though, Twitter failed to realize that a 17-year-old high school student had created a fake account for a fictitious U. S. congress candidate by using a photo from a website containing a collection of synthesized pictures of fake individuals,[30] and verified the account.[31]

A rather strange attempt of open political manipulation involves a low-quality deep fake video produced for the Flemish Socialist Party in Belgium. In 2018, the party posted on its website a deep fake video of U. S. president Donald Trump which showed him allegedly demanding that Belgium leave the Paris climate accord. The figure posing as Donald Trump is speaking English; Dutch subtitles were added. The only part of the fake video not included in the subtitles is the last sentence, when the stand-in says that the footage has been faked.[32] Some of the viewers of the video failed to realize it was a fake, as documented by comments posted on the party's Facebook page.[33] Apparently, the party's aim was to redirect voters to an online petition calling on the Belgian government to take stronger action on climate change. However, because some believed the video was real, the party saw itself forced to publicly emphasize that the video was a joke and a fake. The case demonstrates the power of deep fakes, and that producers are not always able to control their particular impact.[34]

However, deep fakes used for political manipulation also entail a deeper risk, eroding public trust in political institutions. Therefore, even when deep fakes are not employed, political damage can be caused, as was the case in Gabon in 2018. The president of Gabon, Ali Bongo, apparently fell ill in 2018 and did not appear in public for several months. When the government released a video of him giving a New Year's address, allegations were raised that the video was a deep fake, although there was no evidence to substantiate the claims.[35] The allegations were apparently part of a campaign launched by fractions of the country's military who were planning a coup later that year.[36] In this case, therefore, the potential availability of deep fake technology was enough to trigger political disruptions.[37]

While currently there are no known cases of terrorist groups using deep fakes to cause political disruption, technological availability could lead to legal difficulties in the prosecution of returning foreign terrorist fighters. Some current cases in Europe are exploring serious crimes committed by foreign terrorist fighters that are members of the Islamic State in Iraq and the Levant (ISIL).[38] In these cases, images and video footage are used as evidence. Therefore, an ISIL defendant could now credibly claim that images or videos have been tampered with, which could present the prosecution with new technical challenges, which would have to prove the documents' authenticity.

Finally, as deep fakes become increasingly easy to produce, and sophisticated computational power is no longer needed, they allow malicious state actors to add an additional layer of distance between themselves and the deep fakes they may choose to circulate as part of their political manipulation campaigns. This will complicate the already difficult challenge of attribution,[39] as "outsourcing" the creation and circulation of deep fakes as part of their misinformation campaigns will help them plausibly deny their responsibility – this is the so-called "liar's dividend".

In recent years, coordinated misinformation campaigns by malicious state actors have gained traction both in the United States[40] as well as Europe.[41] Boosted by deep fake technology, these misinformation campaigns could significantly increase their corrosive impact. As the examples above demonstrate, so far deep fakes have not played a major role in such attempts. Therefore, there is still sufficient time to develop a multilayered defense system to counter the impact this technology could have if used in political manipulation campaigns. The next chapter will address legal, technology-based and educational measures as potential components of such a defense system. As developing such a system will take time, first steps towards its design should be taken now.

[18] A study by Deeptrace, a company specialized in finding deep fakes online, found that 94% (13,254 of 14,678) of deep fake videos identified by the company in 2019 were non-consensual pornography: H. Ajder, G. Patrini, F. Cavalli, L. Cullen, The State of Deep Fakes. Landscape, Threats and Impact, Deeptrace, September 2019, p. 6, https://share.hsforms.com/1cg_h2aPnRrufZeN8HDjWPw3hq83.

[19] C. Stupp, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Scams using artificial intelligence are a new challenge for companies, Wall Street Journal, 30 August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [06.05.2020].

[20] Synthesized from pictures of existing individuals.

[21] Ajder, Patrini, Cavalli, Cullen, The State of Deep Fakes, p. 13.

[22] T. Simonite, Forget Politics. For Now, Deepfakes Are for Bullies Wired, 4 September 2019, https://www.wired.com/story/forget-politics-deepfakes-bullies/ [06.05.2020].

[23] D. R. Coats, Director of National Intelligence, Worldwide Threat Assessment of the Intelligence Community, 29 January 2019, p. 7, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf [06.05.2020].

[24] The Soufan Center, IntelBrief: The Use of Disinformation in the British Election, 13 December 2019, https://thesoufancenter.org/intelbrief-the-use-of-disinformation-in-the-british-election/ [06.05.2020].

[25] D. Harwell, Faked Pelosi videos, slowed to make her appear drunk, spread across social media, Washington Post, 24 March 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [06.05.2020].

[26] A. Fichera, Manipulated Video Targeting Pelosi Goes Viral, FactCheck.Org, 24 May 2019, https://www.factcheck.org/2019/05/manipulated-video-targeting-pelosi-goes-viral/ [06.05.2020].

[27] J. Waterson, Facebook refuses to delete fake Pelosi video spread by Trump supporters, Guardian, 24 May 2019, https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site [06.05.2020].

[28] B. Coyne, Helping identify 2020 U. S. election candidates on Twitter, Twitter, 12 December 2019, https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html [06.05.2020].

[29] Ballotpedia: About, https://ballotpedia.org/Ballotpedia:About [06.05.2020].

[30] This Person Does Not Exist, https://thispersondoesnotexist.com/ [06.05.2020]; R. Metz, These people do not exist. Why websites are churning out fake images of people (and cats), CNN, 28 February 2020, https://edition.cnn.com/2019/02/28/tech/ai-fake-faces/index.html [06.05.2020].

[31] D. O'Sullivan, A high school student created a fake 2020 candidate. Twitter verified it, CNN, 28 February 2020, https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html [06.05.2020]; Twitter has since suspended the account.

[32] J. Lytvynenko, A Belgian Political Party is Circulating a Trump Deepfake Video, BuzzFeed, 20 May 2018, https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video [06.05.2020].

[33] H. v. d. Burchard, Belgian socialist party circulates ‘deep fake’ Donald Trump video, Politico, 21 May 2018, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/ [06.05.2020].

[34] O. Schwartz, You thought fake news was bad? Deep fakes is where news goes to die, Guardian, 12 November 2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth [06.05.2020].

[35] A. Breland, The Bizarre and Terrifying Case of the ‘Deepfake’ Video that Helped Bring an African Nation to the Brink, Mother Jones, 15 March 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [06.05.2020].

[36] Ajder, Patrini, Cavalli, Cullen, The State of Deep Fakes, p. 10.

[37] S. Cahlan, How misinformation helped spark an attempted coup in Gabon, Washington Post, 13 February 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/ [06.05.2020].

[38] P. Bąkowski, L. Puccio, Foreign fighters – Member State responses and EU action, European Parliamentary Research Service, March 2016, p. 8, https://www.europarl.europa.eu/EPRS/EPRS-Briefing-579080-Foreign-fighters-rev-FINAL.pdf [06.05.2020].

[39] S. Bradshaw, P. N. Howard, The Global Disinformation Disorder: 2019 Global Inventory of Organised Social Media Manipulation, Working Paper 2019.2, Oxford, UK: Project on Computational Propaganda, p. 9, https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [06.05.2020].

[40] K. Roose, S. Frenkel, N. Perlroth, Facebook, Google and Twitter Struggle to Handle November’s Election, New York Times, 29 March 2020, https://www.nytimes.com/2020/03/29/technology/facebook-google-twitter-november-election.html [06.05.2020].

[41] See for example: L. Benková, The Rise of Russian Disinformation in Europe, Austria Institut für Europa- und Sicherheitspolitik, Fokus 3/2018, https://www.aies.at/download/2018/AIES-Fokus_2018-03.pdf [06.05.2020].

5. Managing the threat of deep fakes

Deep fake technology, like all human inventions, is neither inherently good nor inherently bad. It can be used for positive purposes, such as in the arts,[42] or for disseminating public messages in several languages.[43] However, as deep fake technology can also be used for widespread criminal purposes as well as political manipulation, both have the potential to cause widespread social harm.

As this technology will continue to spread, reaching large numbers of users, and hard- and software requirements will cease to be an issue, a range of measures should be discussed. These measures should be designed to minimize the social harm potentially caused by misuse of this technology, but they must guarantee freedom of speech and ensure that such technology can be employed in positive contexts, such as the arts and culture.

Given the current pace of developments in technology and distribution, no single measure seems sufficiently powerful to tackle the problem effectively. For example, wholesale legal bans are unlikely to be effective as this technology is freely available globally through the internet and deep fake videos are distributed on a multitude of platforms.

Therefore, a suite of measures should be employed, targeting both the production and, more importantly, the distribution of deep fake videos. These should comprise legal, technology-based, and educational measures.

5.1 Legal measures

5.1.1 Potential legal restrictions on technology

Taking into view the production side of deep fake videos, one central concern is to maintain a degree of control over new technology. This requires, on the one hand, efforts to ensure that technology-related barriers remain in place that have the potential to prevent non-state actors, including criminals, from employing this technology, but mainly to create a legal framework to prosecute nefarious actors if attribution is possible. This would entail declaring the use of particular types of deep fake software illegal within a jurisdiction.

The other, potentially more important legal restriction concerns widespread access to detection technology. Making this technology fully available to the market and to users would enable nefarious actors to quickly adapt their production methods. This resulting technological “arms race” could be slowed down by keeping advanced detection technology off the market. Therefore, legal safeguards preventing the release of advanced detection technologies could be an effective measure.

In order to tackle malicious use of deep fakes, it is important to distinguish deep fakes used to commit criminal acts – such as fraud or the production of non-consensual pornography – and deep fakes used to drive political manipulation. In both cases, targeting misuse as well as distribution is important, albeit on different levels. To target criminal activity, the misuse of deep fake technology should be the primary concern. In this case, damage is done to an individual or a group of individuals.

Regarding the misuse of deep fakes to drive political manipulation – which is the main emphasis of this report – it is important first to make misuse a prosecutable crime, but legislation must also define appropriate and necessary exceptions, including for satire, comedy, or political critique. However, since politically manipulative deep fakes only reach their full corrosive potential when distributed widely, targeting the distribution mechanisms of such material is a second, equally important approach. In the following sections, we will outline actions aimed at tackling the misuse of deep fakes produced to drive political manipulation.

5.1.2 Legislation targeting misuse for political manipulation purposes

Current regulatory action targets the misuse of such technology for political manipulation purposes by creating new legal categories. For example, the state of California passed AB 730 in 2019, the first law to ban deep fakes from being used with malice in political campaigns:

> *a person, firm, association, corporation, campaign committee, or organization shall not, with actual malice, produce, distribute, publish, or broadcast campaign material that contains (1) a picture or photograph of a person or persons into which the image of a candidate for public office is superimposed or (2) a picture or photograph of a candidate for public office into which the image of another person or persons is superimposed.*
> *(…)*
> *within 60 days of an election at which a candidate for elective office will appear on the ballot, distribute, with actual malice, materially deceptive audio or visual media.*[44]

AB 730 has an important provision, safeguarding the right to distribute deep fakes as part of a news story or political satire:

> *This section does not apply to an internet website, or a regularly published newspaper, magazine, or other periodical of general circulation, including an internet or electronic publication, that routinely carries news and commentary of general interest, and that publishes materially deceptive audio or visual media prohibited by this section, if the publication clearly states that the materially deceptive audio or visual media does not accurately represent the speech or conduct of the candidate.*[45]

This law is a first step towards defending political discourse against the malicious use of deep fake technology. However, in an online environment its application will have to take into account the difficult issue of attribution. Only time will tell whether the continuing challenge of attributing a deep fake video to a potential perpetrator in a legally sound manner will in many cases prevent the effective application of this new legal provision in California in the defense against deep fakes used to drive political manipulation.

The law immediately drew criticism from those fearing it will place undue restrictions on free speech, in particular in political settings.[46] Others argued that the law is "too feeble" to have the desired effects, especially since it requires that the material is produced with malign intent.[47] Finally, the law only applies to the jurisdiction of California and therefore does not cover actors outside the state, leading some to call for action at the federal level.[48]

A similar bill was indeed introduced to the U. S. Congress in December 2019 but has not progressed since.[49] However, in December 2019, the U. S. Congress passed the National Defense Authorization Act for Fiscal Year 2020 (NDAA).[50] Section 5709 establishes a new annual report to be submitted to Congress by the Director of National Intelligence concerning:

> *(A) the potential national security impacts of machine-manipulated media (commonly known as "deepfakes"); and*
>
> *(B) the actual or potential use of machine-manipulated media by foreign governments to spread disinformation or engage in other malign activities.*[51]

This report, according to Section 5709, should also include:

> *An updated identification of the counter-technologies that have been or could be developed and deployed by the United States Government, or by the private sector with Government support, to deter, detect, and attribute the use of machine-manipulated media and machine-generated text by foreign governments, foreign-government affiliates, or foreign individuals, along with an analysis of the benefits, limitations and drawbacks of such identified counter-technologies, including any emerging concerns related to privacy.*[52]

Section 5724 of the NDAA also establishes a "Deep Fake Prize Competition" with an award of $5,000,000 "to stimulate the research, development, or commercialization of technologies to automatically detect machine-manipulated media."[53]

Therefore, while there is no political consensus at the federal level on how to reign in the malicious use of deep fakes, Congress has identified their potential use for political manipulation purposes as a threat to national security.

These legislative efforts reflect the emergence of a growing specialized regulatory system focused on tackling deep fakes. Of course, in addition to drafting new legislation, policy makers could also build on existing laws, such as copyright law or publicity rights (the right to one's own image). However, these existing laws have their own limits when applied to the issue.[54]

Therefore, legislative efforts concentrating on the misuse of deep fakes can only be one element of the response. Politically motivated deep fakes only reach their full disruptive potential if they achieve widespread distribution. Therefore, in addition to legislative efforts targeting their production and misuse, tackling the distribution mechanisms of such material is a second, equally important element in preventing the misuse of this technology.

5.1.3 Legislation targeting distribution mechanisms for political manipulation

Deep fakes produced for political manipulation purposes are circulated via internet services. Here, it is important to distinguish between individual sharing of deep fakes, i.e. the distribution of deep fakes via personal communication, such as email and private messaging, and the sharing of deep fakes on public platforms, in particular via social media. Regulatory action focusing on personal communication would require a significant infringement of legally protected personal communication, and therefore should not be considered a primary focus. Individual distribution, despite the availability of automated mass email software, is also likely to be slower, giving the target of such a deep fake attack time to respond.

Public distribution via social media, in particular via global platforms, ensures instant impact of such material. Its corrosive impact on political discourse is therefore much greater. Therefore, platform providers, notably those with a significant number of users, must contribute to the defense system.

In the United States, the Communication Decency Act (CDA) of 1996 significantly limits attempts to legally require platform providers to set up defense systems against the misuse of their services by individuals or entities posting harmful content. Section 230 of the CDA provides that:

> *No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.*
> *(...)*
> *No provider or user of an interactive computer service shall be held liable on account of-*
>
> *(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or*
>
> *(B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).*[55]

This means that platform providers are free to restrict access to, block, or remove content that they deem undesirable, but they are not liable to third parties harmed by their failure to do so, with the exception of content related to sex-trafficking. Consequently, neither the removal nor the non-removal of content can be litigated as a general matter.

Despite these limitations, political pressure on the industry in the United States has led several major platforms to announce new policies to tackle the distribution of deep fakes. For example, since the end of 2019, Facebook,[56] Twitter,[57] TikTok,[58] Reddit,[59] and Google/YouTube[60] announced policy changes to limit the distribution of deep fakes. However, these new policies vary significantly in terms of their reach.

For example, TikTok updated its policies by introducing a very broad definition that encompasses various forms of deep-fake material: [61]

> *Content that is intended to deceive or mislead any of our community members endangers our trust-based community. We do not allow such content on our platform. This includes activities such as spamming, impersonation, and disinformation campaigns.*
> *(…)*
> ***Impersonation***
> *We do not allow users to impersonate other individuals or organizations in order to deceive the public. When we confirm reports of impersonation, we remove the violating accounts. We do allow exceptions for parody, commentary, or fan accounts, as long as the account does not mislead others with regard to its identity or purpose on TikTok.*
>
> ***Do not post:***
> *As another person or organization by using someone else's name, biographical details, or profile picture in a misleading manner*
>
> ***Misleading information***
> *We do not permit misinformation that could cause harm to our community or the larger public. (…) We also remove content distributed by disinformation campaigns.*
>
> ***Do not post:***
> *Misinformation meant to incite fear, hate, or prejudice*
> *(…)*
> *Content that misleads community members about elections or other civic processes.*[62]

The advantage of such a broad definition is that it potentially encompasses a wide range of malicious activities and forms of deep fake material. However, the company does not define in detail its understanding of what it means to "mislead", "deceive", or distribute "disinformation campaigns", allowing it to take a liberal approach in implementing its policies. Therefore, the quality of internal decision-making mechanisms will be the main factor determining whether such platforms can successfully prevent misuse of their services.

At the other end of the spectrum of potential definitions of deep fakes, Facebook, currently the largest social media platform, has adopted an extremely narrow definition covering only the most advanced deep fake video productions. Its updated Community Guidelines state:

> ***Policy Rationale***
> *Media, including image, audio, or video, can be edited in a variety of ways. In many cases, these changes are benign, like a filter effect on a photo. In other cases, the manipulation isn't apparent and could mislead, particularly in the case of video content. We aim to remove this category of manipulated media when the criteria laid out below have been met.*
> *(…)*
> ***Do not post:***
> *Video that has been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person, and would likely mislead an average person to believe that a subject of the video said words that they did not say*
>
> ***AND***
> *is the product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic.*
>
> *This policy does not extend to content that is parody or satire or is edited to omit words that were said or change the order of words that were said.*[63]

This definition makes clear that the guidelines only cover deep fake videos which have been altered using sophisticated technologies, such as AI, ML or deep learning. Lower quality deep fakes, even if potentially used for political manipulation purposes, are not included, allowing Facebook to ignore their distribution altogether.

Facebook's narrow approach led to immediate criticism highlighting that currently the vast majority of misleading videos are produced using less sophisticated technologies and that its new policy thus failed to effectively address the issue of political manipulation.[64] Facebook's current definition, for instance, does not cover the manipulated video of Nancy Pelosi mentioned earlier.[65]

These widely diverging standards across major global social media platforms are creating a fairly uneven defense landscape that is easily exploited by organized campaigns to use deep fakes for political manipulation. Therefore, while it is positive that platforms seem to have understood the need to take action, this uneven landscape will likely remain ineffective in defending against the misuse of deep fakes for political manipulation. This, in turn, only highlights the need for democratic governments to define a set of minimum standards for the defense mechanisms and systems that social media platforms should deploy to avoid the exploitation of gaps in the respective provisions.[66]

With its *Netzwerkdurchsetzungsgesetz* (Network Enforcement Act, "NetzDG"),[67] Germany has already created a legal instrument that could be used to tackle the distribution of deep fake material, including material distributed for the purpose of political manipulation, and set minimum standards. However, the law's underlying notice and take-down mechanism only has limited impact since it does not require platform providers to take proactive measures, as shown by a study conducted by CEP in early 2020.[68] Still, the current process to amend the act presents an opportunity to introduce stricter regulations to tackle the threat posed by deep fakes.[69] For example, deep fake videos used for criminal or politically manipulative purposes could be defined as violating the victim's right to their own image, and thus potentially constitute illegal content according to section 1(3) of the NetzDG. Thus, they would already be covered by the law.[70]

However, the current notice and take-down mechanism is based on the premise that users will identify malicious content and notify platforms, which are then required to remove such content if it is found to be in violation of the NetzDG.[71] This principle seems too weak to prevent the distribution of deep fake material, including material that is aimed at manipulating political discourse. This is particularly true for cases that require rapid response, such as manipulations launched prior to an election or a crucial vote. It is also important to note that most social media posts gain the majority of views in the first few hours after being published. Therefore, a system relying on users to flag manipulative content and notify the respective platform, which then reviews complaints and only subsequently removes such content, is very likely to be inefficient in preventing political manipulation via deep fakes.

Without specialized software, high-quality deep fake videos are difficult, if not impossible, to distinguish from authentic videos. Such tools are generally not available to the average user. Platforms, especially those with a global reach, thus have a responsibility to take proactive measures. These companies are particularly at risk since deep fakes posted on their platforms will likely expose their large user base to effective political manipulation. Therefore, these platforms should be required to proactively reinforce their defense systems.

Furthermore, additional technology-based measures should be undertaken to limit the overall impact of deep fake political manipulation attempts in Germany.

5.2 Technology

The corrosive nature of deep fake technology used to drive political manipulation lies in its potential to erode public perceptions of reality and, therefore, public trust.[72] To establish a defense against such attacks, it is necessary to take into view solutions that can help to create a repository of confirmed original recordings, along with technology that allows for the forensic verification of deep fake videos.

5.2.1 Certification of original content

In general, videography is perceived to be a representation of reality. “The moving image became the promise of a purer truth, one that could not be complicated by the threat to alteration made feasible by still images.”[73] Deep fakes systematically undermine this deeply held conviction. Consequently, potential technology-based solutions aimed at generating a repository of verified original recordings could be an important step forward. Such an endeavor would of course require a long-term industrial shift as regards the technology used to run recording devices.

As mentioned above, control-capture technology may be an effective way forward.[74] If a digital watermark in the form of a hash[75] can be inserted when the material is originally recorded, any manipulation of the material would automatically update the accompanying hash and indicate that the video had been altered. Hashing technology is a very established verification technology, used regularly to confirm the integrity of datasets after transmission.[76] Blockchain technology[77] may also play a supporting role as a repository of “original” hashes, in particular for recordings of particular public significance, such as statements by heads of states on important policy matters, etc.[78] Such an open repository of hashes linked to original recordings could then be used to quickly and reliably verify alterations made to the respective recordings.

5.2.2 Support for the development of deep fake detection technologies

The development of deep fake detection technologies is a second important technology-based element to help defend against deep fake videos created to drive political manipulation. Such technologies will be crucial to the forensic analysis of potential deep fake videos. Moreover, this development would create the necessary expertise that would enable regulators to effectively audit and assess the effectiveness of a social media platform’s defense systems in combating political manipulation via deep fakes.

In the United States, there are several ongoing initiatives to develop deep fake detection technologies. The NDAA 2020 established a national competition,[79] and several companies, including Facebook,[80] have announced individual programs. Microsoft, Facebook, Amazon Web Services (AWS), and the Partnership on AI[81] launched the Deep Fake Detection Challenge in 2019.[82]

These initiatives demonstrate that the development of such technology requires not only advanced technical expertise, but also significant financial resources. Germany is well-placed to contribute to its development, with several German research institutes already working on AI and machine learning.[83]

5.3 Public education

Both legal and technology-based solutions are needed to tackle deep fake-driven political manipulation. These, of course, do not address the challenge arising from the fact that such manipulations may already be taking effect and altering public perceptions. The risks are especially high prior to elections or votes in parliament. Unfortunately, forensic approaches designed to debunk deep fakes will ultimately not be effective by themselves.

In 2019, a study of the online communities in several European countries ahead of the European Union elections found that:

> *the reach of fact-checkers is limited, often to those digital communities which are not targets for or are propagating disinformation.[84]*

This may also be put down to the fact that false news spreads faster than truthful information on social media. A longitudinal research study conducted in 2018 found that:

> *Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information.*[85]

The researchers also concluded that human behavior was a greater determinant for the spread of false news than automated systems. In such a case, behavioral interventions, such as labeling and disincentivizing the spread of false news, would be more effective in slowing the dissemination of false news than other measures, such as subsequent fact-checker interventions or the implementation of technology-based solutions.[86]

These findings can also be applied to deep fakes, which, if used to drive political manipulation, form a specific subcategory of false news. One indispensable defense element against deep fake-driven political manipulation is civic education – increasing cyberliteracy among users.

Of course, some solutions might focus on promoting so-called counter-speech, a concept which assumes that truth will eventually prevail in an open marketplace of ideas. However, the assumption that an uninhibited dissemination of ideas will push back deep fakes is based on a lack of understanding of the technical distribution mechanisms for news in social media. The algorithms of social media platforms, which are designed to increase time spent on the platform and which feed into the propensity of individuals to share false information, including deep fakes, act as gatekeepers in this marketplace. Consequently,

> *conditions, such as the structural and economic changes that have affected the news media, increased fragmentation and personalization, and increasingly algorithmically-dictated content dissemination and consumption, affect the production and flow of news in ways that may make it more difficult than it has been in the past to assume that legitimate news will systematically win out over false news.*[87]

Although important, counter-speech should be accompanied by broader public education campaigns that create awareness for the fact that seeing is no longer believing.[88] This could be achieved by efforts to increase overall cyberliteracy, starting with specialized courses as part of the regular school curriculum. Some efforts in this regard are already ongoing in Germany, and they should be intensified.[89] These efforts could also be led by media associations. In 2019, for example, the Canadian non-governmental media development organization Journalists for Human Rights (JHR) launched its “Fighting Disinformation through Strengthened Media and Citizen Preparedness in Canada” project, supported by the government of Canada.[90]

However, one effect of growing public skepticism towards online content is that the “liar’s dividend” is used more frequently and more effectively.[91]

A second necessary public education strategy should include efforts to help the public distinguish between content distribution mechanisms, such as social media platforms, and news production mechanisms. Professional media organizations are typically guided by codes of conduct. For example, the German Press Council’s code of conduct outlines in section 1 that:

> *Respect for the truth, preservation of human dignity and accurate informing of the public are the overriding principles of the Press.*[92]

In addition to promoting awareness and cyberliteracy, existing functions, such as reverse image searches, could be made available to the public, and providing access to these should be made mandatory for social media platforms.

> *Reverse image search has empowered journalists, fact-checkers, and everyday netizens to unearth original photos from which forgeries are made. This type of tool allows users to upload an image, then use computer vision to discover similar photos online, which can reveal the photo as altered or presented outside its original context.*[93]

Fostering cyberliteracy and public awareness in combination with an increased availability of tools, such as reverse image searches, should become part of an overall educational policy that could be adopted to bolster society's defenses against deep fake-driven political manipulation.

42 In 1994, the film Forest Gump made particularly innovative use of deep fake video manipulation by inserting the main character, played by actor Tom Hanks, into several original historical recordings.

43 For example, an Indian political party used deep fake technology to spread its campaign message in several languages. See: J. Fergus, Deepfake video in multiple languages is the first of its kind in an Indian election. It's a 'positive campaign', INPUT, 19 February 2020, https://www.inputmag.com/culture/a-deepfake-video-is-the-first-of-its-kind-in-indian-election-campaign [06.05.2020]. Systematic approaches outlining the potential use and misuse of this technology can be found here: R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security; E. Meskys, A. Liaudanskas, J. Kalpokiene, P. Jurcys, Regulating deep fakes: legal and ethical considerations. In: Journal of Intellectual Property Law & Practice 15/2020, Issue 1, pp. 24–31, https://academic.oup.com/jiplp/article/15/1/24/5709090 [06.05.2020].

44 AB 730, Elections: deceptive audio or visual media, October 2019, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 [06.05.2020].

45 Ibid.

46 A. Metwally, JP Mohler, Manipulated Media: Examining California's Deepfake Bill, JOLT Digest, 12 November 2019, http://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-billhttp://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-bill [06.05.2020].

47 B. M. Nonnecke, Opinion: California's Anti-Deepfake Law Is Far Too Feeble. While well intentioned, the law has too many loopholes for malicious actors and puts too little responsibility on platforms, Wired, 5 November 2019, https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ [06.05.2020].

48 D. Castro, State Government Might Not Be Enough to Stop Deepfakes, Governing, 7 January 2020, https://www.governing.com/news/headlines/State-Government-Might-Not-Be-Enough-to-Stop-Deepfakes.html [06.05.2020].

49 H. R. 3230 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, https://www.congress.gov/bill/116th-congress/house-bill/3230/text.

50 National Defense Authorization Act for Fiscal Year 2020, S. 1790, 116th Congress, 1st Session (2019), https://www.govinfo.gov/content/pkg/BILLS-116s1790enr/pdf/BILLS-116s1790enr.pdf [06.05.2020].

51 Ibid.

52 Ibid.

53 Ibid.

54 For example, the "fair use" doctrine allows portions of copyrighted material to be used for public purposes like commentary, satire, parody, reporting, education, or research.

55 Section 230 (c) (1) and (2) Telecommunications Act of 1996, Pub. LA. No. 104–104, 110 Stat. 56 (1996), https://transition.fcc.gov/Reports/tcom1996.pdf [06.05.2020].

56 M. Bickert, Enforcing Against Manipulated Media, Facebook, 6 January 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [06.05.2020].

57 Y. Roth, A. Achuthan, Building rules in public: Our approach to synthetic & manipulated media, Twitter, 4 February 2020, https://blog.twitter.com/

en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [06.05.2020].

[58] L. Mahendran, N. Alsherif, Adding clarity to our Community Guidelines, TikTok, 8 January 2020, https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines [06.05.2020].

[59] Updates to Our Policy Around Impersonation, Reddit, 9 January 2020, https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/ [06.05.2020].

[60] How YouTube supports elections, YouTube, 3 February 2020, https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html?m=1 [06.05.2020].

[61] The company made this policy change while reportedly working on deepfake technology itself. See: J. Constine, ByteDance & TikTok have secretly built a deepfakes maker, TechCrunch, 3 January 2020, https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/ [06.05.2020].

[62] Community Guidelines, TikTok, January 2020, https://www.tiktok.com/community-guidelines?lang=en [06.05.2020].

[63] Community Standards, Manipulated Media, Facebook, https://www.facebook.com/communitystandards/manipulated_media [06.05.2020].

[64] See for example: G. Edelman, Facebook's Deepfake Ban Is a Solution to a Distant Problem. The platform has a plan to deal with tomorrow's disinformation. But what about today's? Wired, 7 January 2020, https://www.wired.com/story/facebook-deepfake-ban-disinformation/ [06.05.2020]; J. Sachs, Facebook's Ban On Deepfakes Not Likely To Help Stop Spread Of Misinformation, Grit Daily, 8 January 2020, https://gritdaily.com/ban-on-deepfakes-facebook/ [06.05.2020]; A. Khalid, Facebook's deepfake ban ignores most visual misinformation, Quartz, 9 January 2020, https://qz.com/1781809/facebooks-deepfake-ban-wont-remove-most-visual-misinformation/ [06.05.2020].

[65] See Chapter 4.2 above.

[66] In the United States, several Senators have already publicly demanded that the tech industry establish standards in this regard. See: B. Vincent, Sens. Marco Rubio and Mark Warner want Facebook, YouTube, TikTok and others to create industry standards for handling synthetic content, Nextgov, 2 October 2019, https://www.nextgov.com/emerging-tech/2019/10/lawmakers-press-social-media-giants-confront-deepfake-threats/160325/ [06.05.2020].

[67] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

[68] A. Ritzmann, M. Macori, H.-J. Schindler, NetzDG 2.0. Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG) und Untersuchung zu den tatsächlichen Sperr- und Löschprozessen von YouTube, Facebook und Instagram, Counter Extremism Project, 12 March 2020, https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper.pdf [06.05.2020].

[69] Bundesregierung, Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes, 31 March 2020, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Aenderung_NetzDG.pdf?__blob=publicationFile&v=2 [06.05.2020].

[70] H.-J. Schindler, N. Semaan, Democratising Deepfakes. How Technological Development Can Influence Our Social Consensus. In: International Reports of the Konrad-Adenauer-Stiftung 1/2020, pp. 60-68, https://www.kas.de/documents/259121/8620647/Democratising+Deepfakes.pdf/8b3a9ba0-b2ff-2e8d-32be-f7992894a5e5?version=1.0&t=1585317007608 [06.05.2020].

[71] Section 1(3), NetzDG, https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

[72] W. A. Galston, Is seeing still believing? The deepfake challenge to truth in politics, Brookings Institution, 8 January 2020, https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [06.05.2020].

[73] T. Le Wagner, A. Blewer, 'The Word Real Is No Longer Real': Deepfakes, Gender, and the Challenges of AI-Altered Video. In: Open Information Science 3/2019, p. 33, https://www.researchgate.net/publication/334730810_The_Word_Real_Is_No_Longer_Real_Deepfakes_Gender_and_the_Challenges_of_AI-Altered_Video [06.05.2020].

[74] See Chapter 3.2 above.

[75] Hashing is defined as the generation of a value or values from a string of text using a mathematical function. See: Definition Hashing, Techopedia, 21 November, 2017, https://www.techopedia.com/definition/14316/hashing [06.05.2020].

[76] Definition Hash, Tech Terms, 21 April, 2018, https://techterms.com/definition/hash [06.05.2020].

[77] A blockchain, or distributed ledger, is the open decentralized distribution of cryptographically secured hashes that is managed by a peer-to-peer network. This provides open access to all participants, while the employed encryption methodology used for the hashes ensures that these cannot be altered. See: Blockchain. What Is Blockchain Technology? How Does Blockchain Work?, BuiltIn, https://builtin.com/blockchain [06.05.2020].

[78] A. G. Martinez, The Blockchain Solution to Our Deepfake Problems. Technology to hack videos will only keep getting better. A decentralized ledger might help us know when we're seeing the truth, Wired, 26 March 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/ [06.05.2020]; P. Madsen, Combating deepfakes with distributed ledgers, Hedera Hashgraph, 10 June 2019, https://www.hedera.com/blog/using-distributed-ledgers-to-combat-deepfakes [06.05.2020].

[79] See above, section 5.2.1.

[80] M. Schroepfer, Creating a data set and a challenge for deepfakes, Facebook AI, 5 September 2019, https://ai.facebook.com/blog/deepfake-detection-challenge/ [06.05.2020].

[81] Partnership of AI is a multi-stakeholder organization of companies and experts on artificial intelligence. See: Frequently Asked Questions, Partnership on AI, https://www.partnershiponai.org/faq/ [06.05.2020].

[82] Deepfake Detection Challenge, https://deepfakedetectionchallenge.ai/ [06.05.2020].

[83] See for example: Nationales Forschungszentrum für angewandte Cybersicherheit ATHENE, https://www.athene-center.de/; Max Planck Institute for Intelligent Systems, https://www.is.mpg.de/; Fraunhofer Big Data and Artificial Intelligence Alliance BIG DATA AI, https://www.fraunhofer.de/en/institutes/institutes-and-research-establishments-in-germany/

fraunhofer-alliances/big-data-and-artificial-intelligence-alliance.html; Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, https://www.iais.fraunhofer.de/en/research/artificial-intelligence.html; Max Planck Institute for Machine Learning, https://www.cis.mpg.de/machine-learning/. A useful overview of ongoing research activities in Germany can be found here: Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP – Drucksache 19/15210 – Beschäftigung der Bundesregierung mit Deepfakes, 2 December 2019, http://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf [06.05.2020].

[84] How Effective Are Fact-Checkers? A preliminary analysis on how successful fact-checkers are at disseminating content across different digital communities of opinion, Alto Analytics, 12 July 2019, https://www.alto-analytics.com/en_US/fact-checkers/ [06.05.2020].

[85] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. In: Science 359/ 2018, p. 1146, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf [06.05.2020]. The researchers analyzed around 126,000 stories tweeted by around 3 million people more than 4.5 million times between 2006 and 2017.

[86] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, p. 1150.

[87] P. M. Napoli, What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble. In: Federal Communications Law Journal 70/2017–2018, Issue 1, p. 59, http://www.fclj.org/wp-content/uploads/2018/04/70.1-Napoli.pdf [06.05.2020].

[88] See for example: I. Beridze, J. Butcher, When seeing is no longer believing. In: Nature Machine Intelligence 1/2019, pp. 332–334, https://www.nature.com/articles/s42256-019-0085-5 [27.05.2020] H. K. Hall, Deepfake Videos: When Seeing Isn‘t Believing. In: Catholic University Journal of Law and Technology 27/2018, Issue 1, pp. 51–76 [06.05.2020].

[89] See for example: Digitally Educated, deutschland.de, 6 February 2018, https://www.deutschland.de/en/topic/knowledge/digital-literacy-for-school-pupils-three-good-examples [06.05.2020].

[90] Journalists for Human Rights (JHR), Launching JHR's program on 'Fighting Disinformation through Strengthened Media and Citizen Preparedness in Canada', CISION, 27 September 2019, https://www.newswire.ca/news-releases/launching-jhr-s-program-on-fighting-disinformation-through-strengthened-media-and-citizen-preparedness-in-canada--899686785.html [06.05.2020].

[91] P. Chadwick, The liar's dividend, and other challenges of deep-fake news, Guardian, 22 July 2018, https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin [06.05.2020].

[92] Presserat, German Press Code, p. 2, https://www.presserat.de/files/presserat/dokumente/download/Press%20Code.pdf [06.05.2020]. The German Press Council is also the body responsible for handling complaints against journalists and media organizations that violate the Press Code, see pp. 11 [26.05.2020].

[93] A. Engler, Fighting deepfakes when detection fails, Brookings Institution, 14 November 2019, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/ [06.05.2020].

6. Conclusion

This report outlined the current state of play concerning deep fake videos and highlighted the ways in which the growing spread of this technology can undermine societal cohesion. Rapid progress is opening up new frontiers in the production of deep fakes, giving nefarious actors, including malicious state actors, new opportunities for interventions. Currently, the high volume of training data necessary for the production of sophisticated deep fakes seems to restrict mass production. However, recently, new advances enabling the production of moving images synthesized from a single still image indicate that this barrier may dissolve in the near future.[94]

Currently, the use of deep fakes continues to be prevalent in a criminal context, particularly in non-consensual pornography. Known cases of deep fakes produced to drive political manipulation involved relatively easy to detect manipulated videos. Nevertheless, even this low-quality material caused political disruptions. Should this technology be deployed en masse as part of an organized and well-targeted misinformation campaign steered by malicious state actors, the disruptive effects of such a campaign could be significant.

The long-term challenge presented by deep fake technology rests in the gradual erosion of public trust as it undermines society's established understanding of truth. Ultimately, misuse of this technology poses a threat to free and open political debate. The discursive principle, which the German philosopher Jürgen Habermas defines as one of the core elements of a democracy,[95] is subverted if discourse itself is based on manipulated perceptions of reality.

In this sense, deep fakes are the technically most advanced aspect of the continuing rise of the threat of fake news. The growing prevalence of low-cost global distribution mechanisms – in particular global social media platforms and their increasingly central role in serving information to the public – exacerbates this risk. The unregulated nature of these platforms and their highly diverging operational policies represent a significant challenge in defending political discourse and maintaining social cohesion.

Consequently, we urgently need to begin a strategic discussion aimed at developing an effective defense mechanism against this emerging threat. Currently, Germany has not been affected by this technology to the same extent as other countries. This indicates that Germany is at an early stage in facing this new challenge and that there is still sufficient time to develop effective responses. Measures are unlikely to have an effect if adopted in isolation. Meeting the complex challenge posed by advanced deep fake-driven political manipulation should involve a multipronged approach that combines legal measures, technology-based solutions, and public education measures that mutually complement each other.

All of the various elements and measures outlined in this report will require political debate, which must begin now. With this report, the Konrad-Adenauer-Stiftung and CEP hope to contribute to launching this crucial debate.

[94] M. Weisberger, This Animated Mona Lisa Was Created by AI, and It Is Terrifying, Live Science, 27 May 2019, https://www.livescience.com/65573-mona-lisa-deepfakes.html [06.05.2020].

[95] J. Habermas, The Theory of Communicative Action. Vol. I: Reason and the Rationalization of Society, T. McCarthy (trans.). Boston: Beacon, 1984.

7. Literature

A **Ajder, Henry / Patrini, Giorgio / Cavalli, Francesco / Cullen, Laurence,** *The State of Deep Fakes. Landscape, Threats and Impact,* Deeptrace, September 2019, https://regmedia.co.uk/2019/10/08/deepfake_report.pdf [06.05.2020].

Alto Analytics, *How Effective Are Fact-Checkers? A preliminary analysis on how successful fact-checkers are at disseminating content across different digital communities of opinion,* 12 July 2019, https://www.alto-analytics.com/en_US/fact-checkers/ [06.05.2020].

ARD/ZDF *Online Studie 2019,* http://www.ard-zdf-onlinestudie.de/files/2019/ARD-ZDF-Onlinestudie-Grafik-2019.pdf [06.05.2020].

B **Bakowski, Piotr / Puccio, Laura,** *Foreign fighters – Member State responses and EU action,* European Parliamentary Research Service, March 2016, https://www.europarl.europa.eu/EPRS/EPRS-Briefing-579080-Foreign-fighters-rev-FINAL.pdf [06.05.2020].

Benková, Lívia, *The Rise of Russian Disinformation in Europe,* Austria Institut für Europa- und Sicherheitspolitik, Fokus 3/2018, https://www.aies.at/download/2018/AIES-Fokus_2018-03.pdf [06.05.2020].

Beridze, Irakli / Butcher, James, *When seeing is no longer believing,* Nature Machine Intelligence 1/2019, pp. 332–334, https://www.nature.com/articles/s42256-019-0085-5 [27.05.2020]

Bickert, Monika, *Enforcing Against Manipulated Media,* Facebook, 6 January 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [06.05.2020].

Bradshaw, Samantha / Howard, Philip N., *The Global Disinformation Disorder: 2019 Global Inventory of Organised Social Media Manipulation.* Working Paper 2019.2, Oxford, UK: Project on Computational Propaganda, https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [06.05.2020].

Breland, Ali, *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink,* MotherJones, 15 March 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [06.05.2020].

BuiltIn, *Blockchain. What Is Blockchain Technology? How Does Blockchain Work?,* https://builtin.com/blockchain [06.05.2020].

Bundesregierung, *Antwort der Bundesregierung auf die Kleine Anfrage der Abgeordneten Manuel Höferlin, Frank Sitta, Grigorios Aggelidis, weiterer Abgeordneter und der Fraktion der FDP* – Drucksache 19/15210 – Beschäftigung der Bundesregierung mit Deepfakes, 2 December 2019, http://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf [06.05.2020].

Burchard, Hans von der, *Belgian socialist party circulates 'deep fake' Donald Trump video,* Politico, 21 May 2018, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/ [06.05.2020].

C **Cahalan, Sarah,** *How misinformation helped spark an attempted coup in Gabon,* The Washington Post, 13 February 2020, https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/ [06.05.2020].

Castro, Daniel, *State Government Might Not Be Enough to Stop Deepfakes,* Governing, 7 January 2020, https://www.governing.com/news/headlines/State-Government-Might-Not-Be-Enough-to-Stop-Deepfakes.html [06.05.2020].

Chadwick, Paul, *The liar's dividend, and other challenges of deep-fake news,* Guardian, 22 July 2018, https://www.theguardian.com/commentisfree/2018/jul/22/deep-fake-news-donald-trump-vladimir-putin [06.05.2020].

Chesney, Robert / Citron, Danielle Keats, *Deep fakes: A looming challenge for privacy, democracy, and national security.* In: California Law Review 107/2019, pp. 1753–1819, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [06.05.2020].

Coats, Daniel R., *Statement for the Record: Worldwide Threat Assessment of the U. S. Intelligence Community,* 29 January 2019, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf [06.05.2020].

Constine, Josh, *ByteDance & TikTok have secretly built a deepfakes maker*, Techcrunch, 3 January 2020, https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/ [06.05.2020].

Coyne, Bridget, *Helping identify 2020 U. S. election candidates on Twitter,* Twitter, 12 December 2019, https://blog.twitter.com/en_us/topics/company/2019/helping-identify-2020-us-election-candidates-on-twitter.html [06.05.2020].

D **_Digitally Educated,_** deutschland.de, 6 February 2018, https://www.deutschland.de/en/topic/knowledge/digital-literacy-for-school-pupils-three-good-examples [06.05.2020].

E **Edelman, Gilad,** *Facebook's Deepfake Ban Is a Solution to a Distant Problem. The platform has a plan to deal with tomorrow's disinformation. But what about today's?,* Wired, 7 January 2020, https://www.wired.com/story/facebook-deepfake-ban-disinformation/ [06.05.2020].

Engler, Alex, *Fighting deepfakes when detection fails,* Brookings, 14 November 2019, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/ [06.05.2020].

F **Farid, Hany,** *Photo Forensics*. Cambridge, MA/London: MIT Press. 2016.

Fergus, J., *Deepfake video in multiple languages is the first of its kind in an Indian election*. It's a 'positive campaign', INPUT, 19 February 2020, https://www.inputmag.com/culture/a-deepfake-video-is-the-first-of-its-kind-in-indian-election-campaign [06.05.2020].

Fichera, Angelo, *Manipulated Video Targeting Pelosi Goes Viral*, FactCheck.Org, 24 May 2019, https://www.factcheck.org/2019/05/manipulated-video-targeting-pelosi-goes-viral/ [06.05.2020].

G **Galston, William A.,** *Is seeing still believing? The deepfake challenge to truth in politics*, Brookings, 8 January 2020, https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [06.05.2020].

H **Habermas, Jürgen**, *The Theory of Communicative Action. Vol. I: Reason and the Rationalization of Society*, T. McCarthy (trans.). Boston: Beacon, 1984.

Hale, James, *More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute*, Tubefilter, 7 May 2019, https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/ [06.05.2020].

Hall, Holly Kathleen, *Deepfake Videos: When Seeing Isn't Believing.* In: Catholic University Journal of Law and Technology 27/2018, Issue 1, pp. 51–76, https://scholarship.law.edu/cgi/viewcontent.cgi?article=1060&context=jlt [06.05.2020].

Harwell, Drew, *Faked Pelosi videos, slowed to make her appear drunk, spread across social media*, The Washington Post, 24 March 2019, https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [06.05.2020].

Hölig, Sascha / Hasebrink, Uwe, *Nachrichtennutzung über soziale Medien im internationalen Vergleich.* In: Media Perspektiven 11/2016, pp. 534–548, https://www.ard-werbung.de/fileadmin/user_upload/media-perspektiven/pdf/2016/11-2016_Hoelig_Hasebrink.pdf [06.05.2020].

J **Journalists for Human Rights (JHR),** *Launching JHR's program on „Fighting Disinformation through Strengthened Media and Citizen Preparedness in Canada",* CISION, 27 September 2020, https://www.newswire.ca/news-releases/launching-jhr-s-program-on-fighting-disinformation-through-strengthened-media-and-citizen-preparedness-in-canada--899686785.html [06.05.2020].

K **Karlsefni, Thorfinn,** *Nicholas Cage, Sound of Music (deepfake),* https://youtu.be/MHkZEpfUnAA [06.05.2020].

Karras, Tero / Laine, Samuli / Aila, Timo, *A style-based generator architecture for generative adversarial networks.* In: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

Karras, Tero / Laine, Samuli / Aittala, Miika / Hellsten, Janne / Lehtinen, Jaakko / Aila, Timo, *Analyzing and improving the image quality of stylegan,* 2019, https://arxiv.org/abs/1912.04958 [06.05.2020].

Khalid, Amrita, *Facebook's deepfake ban ignores most visual misinformation,* Quartz, 9 January 2020, https://qz.com/1781809/facebooks-deepfake-ban-wont-remove-most-visual-misinformation/ [06.05.2020].

L **Li, Yuezun / Chang, Ming-Ching / Lyu, Siwei,** *In ictu oculi: Exposing AI created fake videos by detecting eye blinking.* In: IEEE International Workshop on Information Forensics and Security, 2018, pp. 1–7, https://arxiv.org/pdf/1806.02877.pdf [06.05.2020].

Lossau, Norbert, *Deep Fake: Gefahren, Herausforderungen und Lösungswege,* Konrad-Adenauer-Stiftung, Analysen & Argumente Nr. 382/2020, https://www.kas.de/documents/252038/7995358/AA382+Deep+Fake.pdf/de479a86-ee42-2a9a-e038-e18c208b93ac?version=1.0&t=1581576967612 [06.05.2020].

Lytvynenko, Jane, *A Belgian Political Party is Circulating a Trump Deepfake Video,* BuzzFeed, 20 May 2018, https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video [06.05.2020].

M **Mahendran, Lavanya / Alsherif, Nasser,** *Adding clarity to our Community Guidelines,* TikTok, 8 January 2020, https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines [06.05.2020].

Madsen, Paul, *Combating deepfakes with distributed ledgers,* Hedera Hashgraph, 10 June 2019, https://www.hedera.com/blog/using-distributed-ledgers-to-combat-deepfakes [06.05.2020].

Martinez, Antonia Garcia, *The Blockchain Solution to Our Deepfake Problems. Technology to hack videos will only keep getting better. A decentralized ledger might help us know when we're seeing the truth,* Wired, 26 March 2018, https://www.wired.com/story/the-blockchain-solution-to-our-deepfake-problems/ [06.05.2020].

Meskys, Edvinas / Liaudanskas, Aidas / Kalpokiene, Julija / Jurcys, Paulius, *Regulating deep fakes: legal and ethical considerations.* In: Journal of Intellectual Property Law & Practice, 15/2020, Issue 1, pp. 24–31, https://academic.oup.com/jiplp/article/15/1/24/5709090 [06.05.2020].

Metwally, Amre / Mohler, JP, *Manipulated Media: Examining California's Deepfake Bill,* JOLT Digest, 12 November 2019, http://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-bill [06.05.2020].

Metz, Rachel, *These people do not exist. Why websites are churning out fake images of people (and cats),* CNN, 28 February 2020, https://edition.cnn.com/2019/02/28/tech/ai-fake-faces/index.html [06.05.2020].

N **Napoli, Philip M.,** *What If More Speech Is No Longer the Solution? First Amendment Theory Meets Fake News and the Filter Bubble.* In: Federal Communications Law Journal 70/2017–2018, Issue 1, pp. 55–104, http://www.fclj.org/wp-content/uploads/2018/04/70.1-Napoli.pdf [06.05.2020].

Nonnecke, Brandie M., *Opinion: California's Anti-Deepfake Law Is Far Too Feeble. While well intentioned, the law has too many loopholes for malicious actors and puts too little responsibility on platforms.* Wired, 5 November 2019, https://www.wired.com/story/opinion-californias-anti-deepfake-law-is-far-too-feeble/ [06.05.2020].

Noyes, Dan, *The Top 20 Valuable Facebook Statistics – Updated January 2020,* Zephoria, https://zephoria.com/top-15-valuable-facebook-statistics/ [06.05.2020].

O **Oord, Aaron van den / Dieleman, Sander / Zen, Heiga / Simonyan, Karen / Vinyals, Oriol / Graves, Alex / Kalchbrenner, Nal / Senior, Andrew / Kavukcuoglu, Koray,** *Wavenet: A generative model for raw audio,* 2016, https://arxiv.org/abs/1609.03499 [06.05.2020].

O'Sullivan, Donie, *A high school student created a fake 2020 candidate. Twitter verified it,* CNN, 28 February 2020, https://edition.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html [06.05.2020].

P **Paris, Britt / Donovan, Joan,** *Deepfakes and Cheap Fakes. The Manipulation of Audio and Visual Evidence,* 2019, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf.

Pearson, Helen, *Image manipulation: CSI: Cell biology.* In: Nature 434/2005, pp. 952–953 https://www.nature.com/articles/434952a.pdf?proof=true&draft=collection%3Fproof%3Dtrue [06.05.2020].

Presserat, *German Press Code,* https://www.presserat.de/files/presserat/dokumente/download/Press%20Code.pdf [06.05.2020].

R **Reddit,** *Updates to Our Policy Around Impersonation,* 9 January 2020, https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/ [06.05.2020].

Ritzmann, Alexander / Macori, Marco / Schindler, Hans-Jakob, *NetzDG 2.0. Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG) und Untersuchung zu den tatsächlichen Sperr- und Löschprozessen von YouTube, Facebook und Instagram,* CEP Policy Paper, 12 March 2020, https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper.pdf [06.05.2020].

Roose, Kevin / Frenkel, Sheera / Perlroth, Nicole, *Facebook, Google and Twitter Struggle to Handle November's Election,* The New York Times, 29 March 2020, https://www.bizjournals.com/philadelphia/news/2020/03/30/facebook-google-and-twitter-struggle-to-handle.html [06.05.2020].

Roth, Yoel / Achuthan, Ashita, *Building rules in public: Our approach to synthetic & manipulated media,* Twitter, 4 February 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [06.05.2020].

S **Sachs, Julia,** *Facebook's Ban On Deepfakes Not Likely To Help Stop Spread Of Misinformation,* Grit Daily, 8 January 2020, https://gritdaily.com/ban-on-deepfakes-facebook/ [06.05.2020].

Schindler, Hans-Jakob / Semaan, Nauel, *Democratising Deepfakes. How Technological Development Can Influence Our Social Consensus.* In: International Reports of the Konrad-Adenauer-Stiftung 1/2020, pp. 60–68, https://www.kas.de/documents/259121/8620647/Democratising+Deepfakes.pdf/8b3a9ba0-b2ff-2e8d-32be-f7992894a5e5?version=1.0&t=1585317007608 [06.05.2020].

Schroepfer, Mike, *Creating a data set and a challenge for deepfakes,* Facebook AI, 5 September 2019, https://ai.facebook.com/blog/deepfake-detection-challenge/ [06.05.2020].

Schwartz, Oscar, *You thought fake news was bad? Deep fakes is where news goes to die,* The Guardian, 12 November 2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth [06.05.2020].

Simonite, Tom, *Forget Politics. For Now, Deepfakes Are for Bullies.* Wired, 4 September 2019, https://www.wired.com/story/forget-politics-deepfakes-bullies/ [06.05.2020].

Soufan Center 2019, *IntelBrief: The Use of Disinformation in the British Election,* 13 December 2019, https://thesoufancenter.org/intelbrief-the-use-of-disinformation-in-the-british-election [06.05.2020].

Stupp, Catherine, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Scams using artificial intelligence are a new challenge for companies,* Wall Street Journal, 30 August 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [06.05.2020].

T **Technopedia,** *Definition hashing,* https://www.techopedia.com/definition/14316/hashing [06.05.2020].

Techterms, *Definition hash,* https://techterms.com/definition/hash [06.05.2020].

Tolosana, Ruben / Vera-Rodriguez, Ruben / Fierrez, Julian / Morales, Aythami / Ortega-Garcia, Javier, *Deep-fakes and beyond: A survey of face manipulation and fake detection,* 2020, https://arxiv.org/abs/2001.00179 [06.05.2020].

V **Vincent, Brandi,** *Lawmakers Press Social Media Giants to Confront Deepfake Threats. Sens. Marco Rubio and Mark Warner want Facebook, YouTube, TikTok and others to create industry standards for handling synthetic content,* 2 October 2019, https://www.nextgov.com/emerging-tech/2019/10/lawmakers-press-social-media-giants-confront-deepfake-threats/160325/ [06.05.2020].

Vosoughi, Soroush / Roy, Deb / Aral, Sinan, *The spread of true and false news online.* In: Science 359/2018, pp. 1146–1151, https://science.sciencemag.org/content/sci/359/6380/1146.full.pdf [06.05.2020].

W **Wagner, Travis Le / Blewer, Ashley,** *"The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video.* In: Open Information Science 3/2019, pp. 32–46, https://www.researchgate.net/publication/334730810_The_Word_Real_Is_No_Longer_Real_Deepfakes_Gender_and_the_Challenges_of_AI-Altered_Video [06.05.2020].

Waterson, Jim, *Facebook refuses to delete fake Pelosi video spread by Trump supporters,* The Guardian, 24 May 2019, https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site [06.05.2020].

Weisberger, Mindy, *This Animated Mona Lisa Was Created by AI, and It Is Terrifying*, Live Science, 27 May 2019, https://www.livescience.com/65573-mona-lisa-deepfakes.html [06.05.2020].

Wu, Yue / AbdAlmageed, Wael / Natarajan, Premkumar, *ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features.* In: IEEE Conference on Computer Vision and Pattern Recognition, 2019.

Y **YouTube,** *How YouTube supports elections,* 3 February 2020, https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html?m=1 [06.05.2020].

Legal texts:

B **Bundesregierung,** *Entwurf eines Gesetzes zur Änderung des Netzwerkdurchsetzungsgesetzes,* 31 March 2020, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/RegE_Aenderung_NetzDG.pdf?__blob=publicationFile&v=2 [06.05.2020].

C ***California Assembly Bill-730*** *Elections: deceptive audio or visual media,* https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 [06.05.2020].

Communications Decency Act of 1996, *Section 230, Pub. LA. No. 104–104, 110 Stat. 56,* 1996, https://transition.fcc.gov/Reports/tcom1996.pdf [06.05.2020].

G ***Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken*** *(Netzwerkdurchsetzungsgesetz – NetzDG),* https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html [06.05.2020].

N ***National Defense Authorization Act (NDAA) 2020,*** https://www.govinfo.gov/content/pkg/BILLS-116s1790enr/pdf/BILLS-116s1790enr.pdf [06.05.2020].

U **United States Congress,** *House Resolution 3230: Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019,* https://www.congress.gov/bill/116th-congress/house-bill/3230/text [06.05.2020].

Konrad-Adenauer-Stiftung and Counter Extremism Project

This report was produced in cooperation between the Konrad-Adenauer-Stiftung and the Counter Extremism Project. They have coordinated the creation, publication and placement of this report. The foundation contributes to and encourages the debate of the impact deep fakes have on our society among the German legislature and public.

The Counter Extremism Project (CEP) is a non-profit, non-partisan international organization that aims to counter the threat of extremist ideologies and to strengthen pluralistic-democratic forces. CEP deals with extremism in all forms – this includes Islamist extremism/terrorism as well as right-wing and left-wing extremism/terrorism. To this end, CEP exerts pressure on financial and material support networks of extremist and terrorist organizations through its own research and studies, works against extremist and terrorist narratives and their online recruitment tactics, and develops good practices for the reintegration of extremists and terrorists, and promotes effective regulations and laws.

In addition to offices in the United States, CEP has offices and a separate legal entity as Counter Extremism Project, Germany gGmbH in Berlin, as well as maintains representations in Brussels. CEP‘s activities are led by an international group of former politicians, senior government officials and diplomats. CEP supports policymakers to develop laws and regulations to effectively prevent and combat extremism and terrorism, particularly in the area of combating terrorist financing.

More information can be found here: www.counterextremism.com

Published by:
Konrad-Adenauer-Stiftung e. V. 2020, Berlin

Contact at the Konrad-Adenauer-Stiftung:
Nauel Semaan
Policy Advisor Counter-Terrorism
nauel.semaan@kas.de

Cover page image: © istock by Getty images/posteriori
Kapiteltrenner: p. 6, 15 © istock by Getty images/StudioM1; p. 9, 23
© istock by Getty images/a-r-t-i-s-t; p. 29 © alamy/Dmytro Razinko;
p. 49 © istock by Getty images/Evgeny Gromov
Design and typesetting: yellow too Pasiek Horntrich GbR
The print edition of this publication was climate-neutrally printed by
Kern GmbH, Bexbach, on FSC certified paper.
Printed in Germany.

Printed with financial support from the
German Federal Government.

ISBN 978-3-95721-690-8

The dissemination of fake news as a political instrument has long been an issue in contemporary political discourse. Technological innovations continue to expand the potential for disinformation campaigns, threatening our domestic security. This report explains how the unregulated creation and distribution of so called “deep fakes” – videos and images altered by artificial intelligence – pose a threat to our democratic societies and what policy makers can do to counter it.

Konrad-Adenauer-Stiftung e. V.