Under Discussion

Illustration: © racken.

Democratising Deepfakes

How Technological Development Can Influence Our Social Consensus

An Interview with Dr. Hans-Jakob Schindler,
Senior Director of the Counter Extremism Project

The dissemination of fake news as a political instrument has long been an issue in contemporary political discourse. It is important to react to technological innovations that continue to expand the potential for disinformation campaigns, threatening our domestic security. Nauel Semaan talked to Dr. Hans-Jakob Schindler, Senior Director of the Counter Extremism Project, about the "new superweapon of fake news" – so-called deepfakes.

Ai: *Dr. Schindler, deepfakes have become commonplace and are rapidly spreading online. It is still difficult to differentiate between authentic and unauthentic photos or videos. Deepfakes are, so to speak, the new superweapon of fake news – now it is not only possible to disseminate disinformation, but also to make it look credible. But what exactly are deepfakes, and what is their place in a political landscape already subjected to disinformation?*

Hans-Jakob Schindler: In recent years, the popularity of social media has made fake news and deepfakes part of routine political discourse. Yet, it is important to clearly define the two phenomena. Fake news is false information that is disseminated and then shared further. It is manifested in all forms, including text, audio, images, and video. That is why fake news is fundamentally a social problem that must be combatted in a broad manner. Deepfakes are a subset of fake news. They are electronically modified videos and photographic images that change or simulate people and events and harness the persuasive power of audiovisual media to achieve their effect.

Subsequent changes to audiovisual media are not a fundamentally new phenomenon. In the film industry, electronically modifying videos and recordings has now become an accepted artistic device. For instance, special effects can make actors appear younger, or insert them into old video footage.[1] The movie "Forrest Gump" already pioneered this technology back in 1994. The emergence of deepfakes in political discourse in the form of photo manipulation is not particularly new, either. For instance, already at the start of the 20th century, during Stalin's dictatorship in the Soviet Union, government officials who had fallen out of favour were regularly removed from photos in order to eliminate them, at Stalin's behest, from the official history and national memory. In the past, such operations required a high degree of technical and manual knowledge and skill.

This phenomenon is currently gaining in importance since social media is increasingly exploited as a means of manipulation and, in addition to simple false information, political manipulations occur. The technical development has also ensured that technological obstacles are much easier to overcome, and neither the capabilities of a film studio nor extraordinary computing power are necessary to produce deepfakes. A more playful variation of this technology is the face-swap function that is known best from the social media platform Snapchat.[2]

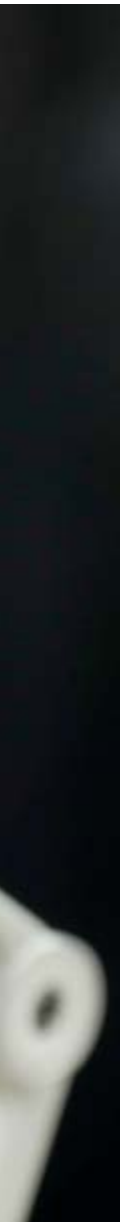
Ai: *Does that mean that you no longer have to be a techie to manipulate images and videos? So anyone could create a deepfake?*

Hans-Jakob Schindler: Professor Hany Farid, Senior Advisor of the Counter Extremism Project (CEP), is currently working on a study on this issue on behalf of the CEP and the Konrad-Adenauer-Stiftung.[3] In this context, he speaks of a "democratisation" of deepfake technology. That means that the production of deepfakes can be performed by a much larger number of actors, and hence result in an increased occurrence and political impact.

Manipulation via app: Today, technological obstacles for producing deepfakes are much easier to overcome.
Source: © Steve Marcus, Reuters.

In principle, there are three methods for creating deepfakes: face swap, lip sync, and puppet master. The face swap method transfers the facial features of one person to the head of another. This enables, for instance, an actor to perform certain actions while the face of the targeted person is added creating the impression that he or she performed the actions.

The more technically elaborate lip sync method only simulates the target person's lip movements so that they adapt to the new spoken words. The latter is either spoken synchronously or provided by synthesising the target person's voice. This allows a recording to be manipulated so that the target person says whatever is necessary for the video manipulation. However, it is sometimes possible to see that the lip movements do not always match the rest of the target person's facial movements.



The puppet master method is the most technically complex, and has not yet been fully developed and perfected. It involves maintaining the target person's face in the video, but completely electronically manipulating his or her facial movements. This means that not only the voice, but all of the facial movements, can be entirely synthesised. The average viewer no longer notices any manipulation, therefore making it possible to have the target person speak authentically given that facial muscle movements are in complete harmony with mouth movements. The weakness of this method is that the facial movements in the manipulated video do not always correspond to the target person's natural movement pattern – this is an important point for forensic detection and proof of manipulation.

Ai: *Could you give some current examples of deepfakes?*

Hans-Jakob Schindler: At the moment, illegal deepfakes are primarily being used in blackmailing and extortion cases. I recently heard about a case in which an employee received a call that he assumed was from his superior. However, the voice on the phone was synthesised. The employee shared important account data that resulted in financial damage to his company.[4] Deepfakes are also used in non-consensual pornography.[5] This involves blackmail with electronically manipulated recordings that allegedly show the victim in embarrassing situations. Victims pay to ensure that the videos are not distributed.

A subset of deepfakes is the creation of artificial identities. Here, images of existing persons are combined electronically to produce images of a new person that do not match any other living person. Such new, unique electronic identities can be fleshed out with CVs and biographical documents that can be ordered online. This is a new variant of identity fraud. The artificial identities are then used to case a target person for espionage or prepare a spear-phishing operation.[6]

But we are also witnessing an increased use of deepfake videos in the political arena, too. Last year, a video of the Speaker of the US House of Representatives, Nancy Pelosi, circulated in which she was supposedly drunk while giving a speech.[7] Although the video was debunked relatively quickly, it underscored the explosive political potential of combining this technology with the dissemination capabilities of social media.

Ai: *What specific dangers do deepfakes pose to our society? What actors have the intention and capability of using deepfakes as a weapon?*

Hans-Jakob Schindler: Video recordings are extremely credible, since they are assumed to accurately reflect reality. That is why skilful manipulation of such recordings for criminal or political manipulation is extremely problematic. If, as can be expected, the current technical trend of simplification and dissemination of this technology continues, it will further undermine the basic social consensus about what is factually true and what is not. One of the gravest consequences is the liar's dividend.[8] Because it is now possible to manipulate videos almost perfectly, it can always be claimed that videos of embarrassing or illegal actions are actually deepfakes. This has repercussions for both, political discourse and, in some cases, legal procedures. That is why the development of technologies that allow detection of deepfakes is an important societal task.

In the last few years, the political sphere has seen repeated cases in which authoritarian regimes attempted to manipulate political processes and elections in democratic states and erode the basic societal consensus. The 2016 US presidential election is merely the best-known example of this. The progressive dissemination and simplification of this technology allows such actors to dispense with state structures when implementing their strategies. If supposedly private individuals can produce deepfakes on behalf of states, it will be all the more difficult to identify clear political responsibility. This represents a growing danger, especially since major technology companies still refuse to assume any responsibility for the dissemination of such manipulations. We only need to think of the US congressional hearings with Mark Zuckerberg in late October 2019, during which he, as CEO of Facebook, denied any responsibility of Facebook for the distribution of false political information on his global platform.[9]

Effective propaganda machinery: It is assumable that terrorist organisations will employ new technologies in future to support the manipulation of individuals in their efforts to radicalise and recruit. Source: © Dado Ruvić, Reuters.

Ai: *You have been dealing with international security policy for 20 years. Your focus has always been on combatting terrorism. Large jihadist organisations, such as the so-called Islamic State, are known for their effective media and propaganda strategies. How relevant and viable is the use of deepfakes for terrorist organisations?*

Hans-Jakob Schindler: At the moment, there are no known cases of terror organisations producing deepfakes. However, that does not stop the effective propaganda machinery of organisations such as the Islamic State (IS) from employing such technologies in future to support the manipulation of individuals in their efforts to radicalise and recruit.

Specifically, the deepfake phenomenon can play a role in the judicial processing of IS returners from Iraq and Syria. Some current cases in Europe are dealing with serious IS crimes.[10] Images and video material are also being used as evidence. The authenticity of these IS recordings are beyond question, but the liar's dividend could make prosecution much more difficult in less serious cases. If the accused IS members could now credibly assert that images and video and audio recordings of their crimes have been electronically manipulated, the prosecutor would be faced with new technical challenges.

Ai: *But with a little time, it is often possible to prove whether an image or video has been manipulated or not. Should a rebuttal and explanation of the disinformation not be sufficient to combat the effects of deepfakes?*

Hans-Jakob Schindler: It is now possible to electronically detect deepfake videos, yet it takes a great deal of technical effort. The University of California, Berkeley is currently working on developing such methods.[11] In principle, these detection methods are based on the creation of typical movement patterns for individuals. Each person has a number of idiosyncratic head, mouth, and muscle movements that match their spoken words and result in a speech and movement pattern unique to that individual. A relatively precise pattern can be calculated from this combination of various factors. This speech and movement pattern is then compared to the recognisable patterns in the video. Since manipulation necessarily involves changing these patterns, it can thus be proven with a high degree of mathematical probability.

However, this method works only if there are enough reliable original recordings of the person shown in the suspicious video for a pattern to be calculated. Hence, it is currently available only for people in the public eye. Fortunately, a high-quality deepfake video also requires a large number of original recordings, so this method is effective at providing evidence.

Such forensic methods are especially useful in the judicial area, where collecting evidence provides sufficient time for effective forensic proof to be collated. Such methods are also helpful in effectively combatting deepfake videos used for political manipulation, but are not sufficient. Social science research has shown that merely debunking fake news is not enough to greatly reduce its impact.[12] Corrections of false information are not as influential on consumers of fake news as the original story. The same effect can be assumed for deepfake videos. That is why forensic evidence of such manipulation only form one part of a wider range of measures.

Ai: *How else can the threat of deepfakes be effectively combatted?*

Hans-Jakob Schindler: Effectively combatting political manipulation owing to deepfake videos will require a range of solutions. First of all, it is important to raise social and political awareness of the capabilities and dangers arising from this technology. It must be emphasised that not every video disseminated via social media is credible. Questions of trust in the system continue to be important here. If trust in the effectiveness and credibility of the political system is undermined, manipulation becomes easier and the damage caused by deepfake videos greater.

In addition to raising awareness, there are technical options for limiting the effect of deepfake videos. If the industry could agree upon the automatic inclusion of an electronic signature in the data set when the video is originally recorded, originals could be certified in this way. This technology, called "hashes", has been around for a long time, and is successfully used in various applications such as data transmission.[13] Whenever there is a change in the original file, the hash also changes, which could provide an initial indication of potential manipulation.

Ultimately, we will also need to take a closer look at the dissemination mechanisms for deepfake videos. This primarily involves the large social media platforms and companies. There is no way to control global dissemination mechanisms with hundreds of millions of users, and in the case of Facebook even billions, without some sense of corporate social responsibility. The huge impact of targeted political manipulation is significantly increased when manipulation is widely distributed. For several years, the Counter Extremism Project has argued that the adoption of regulatory and legislative measures is inevitable. Germany's Network Enforcement Act (Netzwerkdurchsetzungsgesetz, or NetzDG) represents a trailblazing first step towards more responsibility for platform operators.[14] Deepfake videos that are used for criminal or politically manipulative purposes can be defined as a violation of the victim's right to his or her own image. They thus constitute illegal content within the meaning of Paragraph 1 (3) of the NetzDG, and are potentially already covered by the law.[15] The Counter Extremism Project will actively support the law's amendment, which is to take place in 2020, from its new office in Berlin.

Ai: *At KAS, we are taking a multi-faceted approach to the issue of deepfakes in such formats as our Facts & Findings, in which economic journalist Norbert Lossau discusses necessary action and solutions for dealing with deepfakes. We are working with CEP to publish a joint study in mid-2020, which specifically addresses deepfake security concerns. In this context, I would like to ask you a question regarding your forecast for the situation in Germany: Do you think that the upcoming Bundestag elections might become a target for deepfake attacks?*

Hans-Jakob Schindler: There is no doubt that, in the last few years, external actors have attempted to influence the political process within Germany.[16] There is current evidence that deepfake videos were used to spread political disinformation during Britain's House of Commons elections.[17] There is no reason to assume that such actors will not try again, using all the technical means at their disposal, to achieve their goal. Deepfakes are a

potentially extremely effective new technical weapon in this context. That is why it will be important to raise public awareness and employ technical and legislative measures to enhance the defensibility of the political process in Germany. A certain degree of manipulation will remain possible in any system. The question is, however, whether the effectiveness of such attempts and thus the harm to political and social discourse can be contained.

There is still enough time to counteract the manipulative potential of deepfakes. Nevertheless, the social debate surrounding the issue should start now, since, as I have pointed out, a suite of measures will be necessary. Decisions about how and to what extent new structures are to be created, technology innovations implemented, or regulatory interventions made, will certainly take longer than the technical refinement of deepfake technology. CEP and Konrad-Adenauer-Stiftung will publish the results of their joint study in mid-2020. The study will also include initial specific recommendations for actions for political decision-makers in Berlin.

The interview was conducted by Nauel Semaan, Policy Advisor for Counter Terrorism at the Konrad-Adenauer-Stiftung.

–translated from German–

1 Osteried, Peter 2019: Wie Hollywood die Totenruhe stört, golem.de, 13 Nov 2019, in: https://glm.io/144938?t [10 Feb 2020].
2 Goss, Tricia 2019: How to Do a Face Swap, lifewire, 11 Nov 2019, lifewire, in: https://bit.ly/38ga0x3 [10 Feb 2020].
3 Counter Extremism Project, in: https://counterextremism.com [12 Feb 2020].
4 Stupp, Catherine 2019: Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, The Wall Street Journal, 30 Aug 2019, in: https://on.wsj.com/31LyThL [10 Feb 2020].
5 "Non-consensual pornography" refers to the distribution of private images and video recordings against the will of the affected person. See Scott, Alexandrea 2017: What is Nonconsensual Pornography?, National Council of Juvenile and Family Court Judges, 12 Feb 2017, in: https://bit.ly/2tUh43A [10 Feb 2020].
6 Satter, Raphael 2019: Experts: Spy used AI-generated face to connect with targets, AP News, 13 Jun 2019, in: https://bit.ly/37fnaJm [10 Feb 2020]; Swinhoe, Dan 2019: What is spear fishing? Why targeted email attacks are so difficult to stop, CSO, 21 Jan 2019, in: https://bit.ly/31UfaMY [10 Feb 2020].
7 Winkler, Peter 2019: Ein Video zeigt eine betrunkene Nancy Pelosi – und führt uns vor Augen, was mit Deepfakes heute alles möglich ist, Neue Zürcher Zeitung, 25 May 2019, in: https://nzz.ch/ld.1484614 [10 Feb 2020].
8 Harwell, Drew 2019: Top AI researchers race to detect 'deepfake' videos: 'We are outgunned', The Washington Post, 12 Jun 2019, in: https://wapo.st/2UKvb6A [10 Feb 2020].
9 Rodriguez, Salvador 2019: Watch video of AOC grilling Zuckerberg on Facebook allowing lies in political ads, CNBC, 24 Oct 2019, in: https://cnb.cx/2OLgM64 [10 Feb 2020].
10 United Nations Meetings Coverage and Press Releases 2019: Victims' Testimony Steering United Nations Team Investigating ISIL / Da'esh Atrocity Crimes in Iraq, Special Adviser Tells Security Council, 26 Nov 2019, in: https://bit.ly/31UfGdS [10 Feb 2020].
11 Manke, Kara 2019: Researchers From the I School and Engineering Use Facial Quirks to Unmask 'Deepfakes', Berkeley School of Information, 18 Jun 2019, in: https://bit.ly/38gRggS [11 Feb 2020].
12 Chan, Man-pui Sally / Jones, Christopher R. / Jamieson, Kathleen Hall / Albarracín, Dolores 2017: Debunking, A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation, in: Psychological Science 28:11, pp. 1531–1546.
13 Schmitz, Peter 2017: Was ist ein Hash?, Security Insider, 23 Aug 2017, in: https://bit.ly/2HhVdGc [10 Feb 2020].
14 Echikson, William / Knodt, Olivia 2018: Germany's NetzDG: A key test for combatting online hate, Counter Extremism Report 2018/09, Centre for European Policy Studies (CEPS), Nov 2018, in: https://bit.ly/2tUh8Ak [10 Feb 2020].
15 It would be appropriate to make special mention of § 1 (3) of the NetzDG in conjunction with § 201a of the German Criminal Code: see Federal Ministry of Justice and Consumer Protection 2017: Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), 1 Sep 2019, in: https://bit.ly/2SDlMej [10 Feb 2020].
16 Pörzgen, Gemma 2017: Informationskrieg in Deutschland? Zur Gefahr russischer Desinformation im Bundestagswahljahr, Federal Agency for Civic Education (bpb), Aus Politik und Zeitgeschichte (APuZ), 19 May 2017, in: https://bpb.de/248506 [10 Feb 2020].
17 The Soufan Center 2019: IntelBrief: The Use of Disinformation in the British Election, 13 Dec 2019, in: https://bit.ly/2UJTCkt [10 Feb 2020].