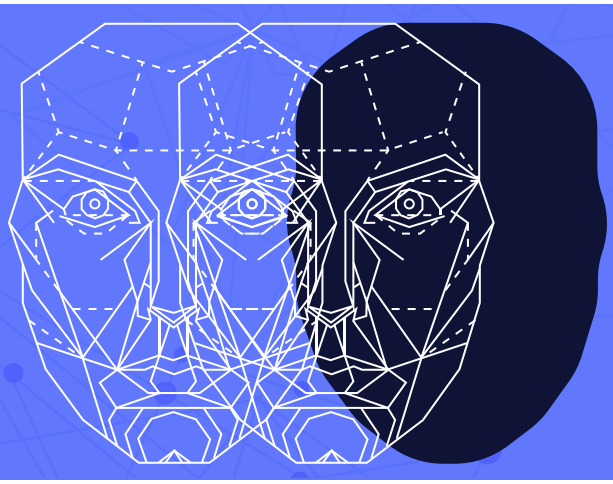
info@democracy-reporting.org
www.democracy-reporting.org

DEMOCRACY REPORTING INTERNATIONAL

DEEPFAKES: A NEW DISINFORMATION THREAT?

Understanding the potential harms and what is being done to address the threat



EXECUTIVE SUMMARY

Highly realistic "deepfakes" (manipulated audio-video content) generated by artificial intelligence (AI) are just around the corner and both experts and social media companies may not be well equipped to detect them. At the same time, it is becoming easier and cheaper for amateurs to manipulate video, and low-grade manipulation such as simply slowing down a video may be effective enough to reach a large audience before platforms or experts can authenticate content. Other factors may accelarate this threat including state disinformation actors with advanced AI capacities, the increasing popularity of video and private-messanging platforms and rising polarisation globally. Some experts warn against hype around deepfakes, believing that detection models may be improved when an influx of deepfakes provides more input data. There is a risk that too much talk about deepfakes may become self-defeating by needlessly undermining public trust in online sources.

That said, deepfake technology may pose particular harm to democratic discourse, individuals and national and private-sector security. At this point in time, political deepfakes are not prevalent, suggesting the threat to democratic discourse has not yet materialised. On the other hand, deepfakes already pose a tangible human rights threat to women in particular. The rights of many women around the world are violated by the use of deepfakes for pornographic content (headshots of women are superimposed on pornography). Deepfake techology may also be used for legitimate purposes (e.g. political satire and entertainment).

To combat deepfakes, stakeholders have focused on technical detection solutions, policies and recommendations for society (e.g. media literacy). Such solutions include machine-learning models capable of determining whether or not a video has been manipulated or using blockchain as a digital footprint. In terms of platform policies, companies define harmful deepfakes in different ways and will respond differently (e.g. by removal or labelling) – although most policies require a high threshold to act. Companies that currently offer deepfake development technology (TikTok and Snapchat) provide no policy (although they promise that such content will be watermarked). In the EU's 2018 disinformation action plan, the European Commission mentioned deepfakes as a threat, but it is unclear how successful the strategy has been in tackling the issue. The EU has several opportunities to act by supporting the recommendations by NATO StratCom and taking further action in the upcoming Digital Services Act and/or European Democracy Action Plan. Outside of the EU, several governments (some U.S. states) have enacted deepfake bans during elections and related to fabricated pornography, although such policies highlight a need to balance free speech and effectiveness.

INTRODUCTION

The saying "seeing is believing" may no longer be true. Artificial Intelligence (AI) and less advanced technology may be used to manipulate video or create so-called deepfakes. This new form of disinformation raises the question of whether we can believe what we see. What does this mean for truth in democratic discourse, holding government accountable and broader society?

Deepfakes have been described as "a new tool for an old problem"[1] – disinformation. But they may also represent a new level of threat as videos are more intuitively credible than text

[1] James Andrew Lewis, "Trust Your Eyes? Deepfakes Policy Brief", Center for Strategic & International Studies, Washington D.C., 23 October 2019, https://www.csis.org/analysis/trust-your-eyes-deepfakes-policy-brief.

and often speak directly to emotions.
Several cases show that the use of deepfake technology is already a threat to society in a number of different areas. Manipulated videos have been used to falsely represent government officials causing confusion among citizens.

Why are deepfakes a threat? What are the potential impacts of this technology? What is being done? What next steps are needed? These are some of the questions we will answer in this briefing paper.

WHAT ARE DEEPFAKES?

The word deepfake[2] is typically used as an umbrella term to describe all forms of audio-visual manipulation including video, audio or both.[3] This may or may not involve the use of AI technology. It is also possible to use AI to generate text at scale to produce fake news articles and more (so-called "readfakes").[4] However, this paper will focus on the manipulation of audio-visual content referred to here as deepfakes.

Definitions:	
Synthetic media	Any form of media generated by AI, including video, audio or images and text.
Audio-visual (AV) manipulation	Any technical means for influencing the interpretation of audio and/or visual media, which differs on a spectrum of technical sophistication, barriers to entry, and techniques
Deepfake	"Deepfake" is used in this paper as an umbrella term to cover all forms of audio-visual manipulation, although strictly-speaking, deepfakes are highly sophisticated manipulation of audio-visual media using AI-driven technology. This term comes from merging "deep learning" with "fake".
Cheapfake (Shallowfake)	Low-level manipulation of audio-visual media created with accessible software or no software to speed, slow, cut, restage or re-contextualise content.
Readfake	The AI generation of fake text at scale.

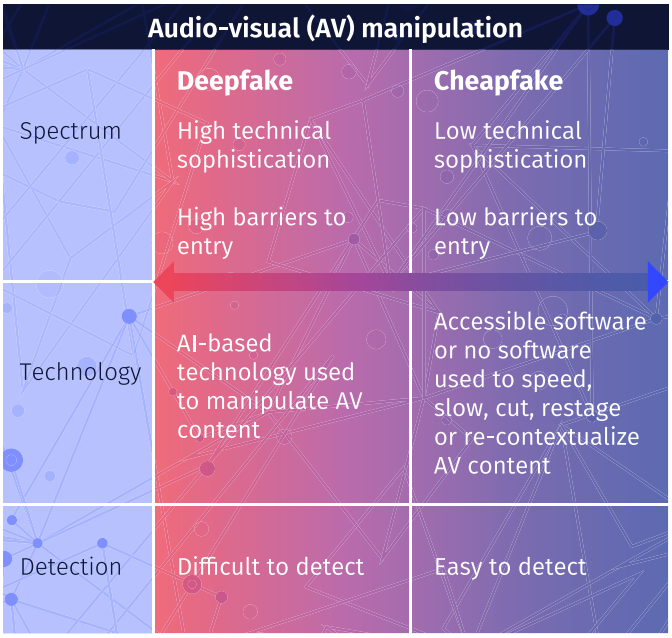
1. THE THREAT

The biggest fear may be the "perfect" deepfake, which experts, companies and government will be unable to tell whether or not a recording has in fact been manipulated. When false content is undetectable, even by experts, people will not be able to distinguish fact from fiction. Even if social media companies or governments enact policies against deepfakes, they cannot enforce them if detection technology does not exist. This means false information may be able to spread at scale before users are informed or content is removed. In October 2019, a private-sector deepfake developer and top detection expert at the University of Southern California, Professor Hao Li, said a "perfectly real" deepfake would be possible in six to 12 months.[5]

The current state of detection has been described as a game of cat and mouse as forensic scientists cannot keep up with the rate of development.[6] University of California Berkely's Professor Hany Farid describes how "The number of people working on the video-synthesis side, as opposed to the detector side, is 100 to 1".[7] It is not entirely clear how much investment big tech companies are making in deepfake detection. Facebook and Microsoft have announced an initiative to create better data sets and incentivise participation in detection challenges, but is has only limited funding ($10 million) and is likely only one part of the efforts.[8]

Video manipulation technology is becoming cheap and easy to use. Manipulating audio-visual material is already possible using less advanced software or no software at all, as described by Britt Paris and Joan Donavon of Data & Society.[9] This means that nefarious actors are able to manipulate video in a convincing way at a low cost with low technical knowledge. Additionally, even a low-tech audio-visual manipulation may be sufficient to convince the public into beliving false or misleading information. Paris and Donovan describe this as a "spectrum" of deepfake to cheapfakes[10]:

[2] The term emerged in 2017 when a Reddit user named deepfakes posted the first deepfake, which was a pornographic video with a celebrity's face integrated using AI technology. Samantha Cole, "We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now", Vice, 24 January 2018, https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley
[3] For more detailed information on how deepfakes are made see Tim Hwang, Deepfakes: Primer and Forecast, NATO StratCom COE, May 2020, https://www.stratcomcoe.org/download/file/fid/82786
[4] Cory Bergman, "11 trends in global disinformation for 2020", FACTAL blog, 17 December 2019, https://blog.factal.com/2019/12/11-emerging-trends-in-global-disinformation-for-2020/
[5] Tonya Mosley, "Perfect deepfake tech could arrive soonr than expected", WBUR, Boston, 2 October 2019, https://www.wbur.org/hereandnow/2019/10/02/deepfake-technology
[6] "Imposter Syndrom", Octavian Report, 2019, https://octavianreport.com/article/hany-farid-fight-threat-deepfakes/2/
[7] Drew Harwell, "Top AI researchers race to detect 'deepfake' videos: 'We are outgunned', The Washington Post, Washington D.C., 12 June 2019, https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/
[8] "Creating a data set and a challenge for deepfakes", Facebook, https://ai.facebook.com/blog/deepfake-detection-challenge/
[9] Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence", Data & Society, 18 September, 2019, https://datasociety.net/library/deepfakes-and-cheap-fakes/
[10] Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence".



Deepfakes are already rapidly on the rise as manipulating audio-visual becomes cheaper and easier. A 2019 study by Deeptrace found that the number of deepfakes doubled within their seven month period of study.[11] As a form of disinformation, manipulated video has the potential to become as ubiquitous as false stories are today. Although, a paper by Tim Hwang of Harvard-MIT (commissioned by NATO StratCom Center of Excellence) points out that as more deepfakes emerge, the data set will grow larger which actually helps AI detection models.[12]

ACCELERATING THREAT FACTORS

Several factors may accelerate the threat that manipulated audio-visual content poses to society:

State actors known to carry out foreign influence campaigns have increasingly sophisticated AI technology. Clint Watts from the Foreign Policy Research Institute notes that both China and Russia, in particular, have strong enough AI capacities to incite fear and distort democracies.[13] That said, domestic threats should not be overlooked.[14]

Video-sharing platforms are on the rise, making deepfakes an attractive medium for disinformation. Social media specialised in video-sharing are highly popular. YouTube is the second most popular social media platform worldwide and TikTok is seventh[15] and on the rise.[16] Both platforms are known to be used for politics,[17] particularly by activists, influencers and politicians themselves. Their popularity provides deepfakes a ready distribution platform.

Private messaging platforms are some of the most popular social media worldwide, meaning disinformation can spread quickly unnoticed and at scale. WhatsApp, Facebook Messenger and WeChat are in the top five social media platforms globally.[18] These platforms provide the perfect distribution means to spread disinformation in a "peer-to-peer" manner[19] making dissemination invisible.

Rising polarisation globally means "partisans can barely agree on facts".[20] A polarised political environment creates the perfect storm for false audio-visual content. Strong attitudes based on deeply-rooted political beliefs may determine people's assumptions about a video before it has been authenticated. Nefarious actors may attempt to exploit these cracks in society to divide people further. In polarised societies, different groups tend to follow different and often highly partisan media, providing less control against deepfakes that one would expect from quality media that folllow journalistic standards of non-partisanship.

IS THE THREAT OVER-HYPED?

There are counter opinions to the idea of an imminent flood of social media with deepfakes. Hwang argues that the threat is real although more narrow than often portrayed. In his view the biggest risk would be one-off operations by sophisticated actors targeting high-level politicians (think Macron "leaks"). He warns that focusing only on deepfakes may distract from related threats: for example, AI can also be used to create mass false identities. In DRI we have stressed that discussion on disinformation should not over-focus on the "message" (content) but understand that any anaylsis of disinformation also needs to consider the "messenger" and the "messenging/disemination" (the 3M approach).[21] Hwang rightly points at the risk of reducing the debate too much to the question of content.[22]

[11] Giorgio Patrini, "Mapping the Deepfake Landscape", Deeptrace, 7 October 2019, https://deeptracelabs.com/mapping-the-deepfake-landscape/
[12] Tim Hwang, "Deepfakes – Primer and Forecast", NATO StratCom, May 2020, https://www.stratcomcoe.org/deepfakes-primer-and-forecast
[13] Limarc Ambalina, A Threat to Individuals and National Security", Lionbridge, 21 June 2019, https://lionbridge.ai/articles/deepfakes-a-threat-to-individuals-and-national-security/
[14] Seva Gunitsky, "Democracies Can't Blame Putin for Their Disinformation Problem", Foreign Policy, 21 April 2020, https://foreignpolicy.com/2020/04/21/democracies-disinformation-russia-china-homegrown/
[15] "Most popular social networks worldwide as of July 2020, ranked by number of active users (in millions)", Statista, July 2020, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[16] "The Incredible Rise of TikTok – [TikTok Growth Visualization], Influencer Marketing Hub, 14 May 2020, "https://influencermarketinghub.com/tiktok-growth/
[17] Chris Cillizza, "YouTube is 10 years old. Here's how it has changed politics forever", Washington Post, Washington D.C., 23 April 2015, https://www.washingtonpost.com/news/the-fix/wp/2015/04/23/you-tube-started-10-years-ago-today-its-fundamentally-changed-politics/
[18] "Most popular social networks worldwide as of July 2020, ranked by number of active users (in millions)", Statista, July 2020, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[19] Samuel Wooley, "Encypted messaging apps are the future of propagana", Brookings, 1 May 2020, https://www.brookings.edu/techstream/encrypted-messaging-apps-are-the-future-of-propaganda/
[20] Danielle Citron, "Deepfakes and the New Disinformation War", Stanford Law School: The Center for Internet and Society, Palo Alto, 11 December 2018, http://cyberlaw.stanford.edu/publications/deepfakes-and-new-disinformation-war
[21] Michael Meyer-Resende and Rafael Goldzweig,
"BP100: Online Threats to Democratic Debate: A framework for a Discussion on Challenges and Responses", Democracy Reporting International, 26 June 2019, https://democracy-reporting.org/dri_publications/bp100-online-threats-to-democratic-debate/
[22] Tim Hwang, "Deepfakes – Primer and Forecast"

2. POTENTIAL IMPACTS AND HARMS

WHEN IS THE USE OF DEEPFAKE TECHNOLOGY HARMFUL TO SOCIETY?

Deepfakes are harmful when people cannot tell whether the content is real or not or if they tangibly harm the subject through exploitive imagery in the absence of consent, and this information cannot be confirmed by authorities or media outlets. Harmful deepfakes are created by nefarious actors with the intention of manipulating society and/or targeting specific individuals through intential misrepresentation.

However, not all manipulated audio-visual content is harmful to society, and deepfake technology can have legitimate uses. Citizens should be able to freely express their political opinions online, which includes satirising and criticising politicians or other people. Additonally, deepfake technology is used widely by the entertainment industry for special effects in movies.

However, when manipulated political content is clearly labelled, particularly during elections, citizens can tell between fiction and reality and therefore view a satirical deepfake for its intended purposes. Authoritarian regimes may attempt to completely ban deepfakes as a means of restricting free speech.

Case 1: Satirical Deepfakes in Brazil

The Brazillian journalist Bruno Satori uses deepfake techniques (open source libraries, tutorials, Adobe Premire and Photoshop) to satirise Brazilian politicians, particularly President Bolsonaro.[23] Satori labels his posts clearly indicating that he has created the content.

POTENTIAL AREAS OF HARM

Three potential areas of harm[24] may be considered when evaluating deepfake techology's use by nefarious actors: harm to democratic discourse, individuals and national/private security:

Democratic Discourse

- Citizens unable to decipher between fact and fiction
- Politician's culpable deniability
- Loss of citizen trust

Individuals

- Defamation and libel
- Fake pornography
- Blackmail and revenge porn
- Harm to both public and private individuals

Security

- National security
- Cybercrime (targeting governments and private companies)
- Civil conflict

Non-harmful:

Political Satire

- Citizen's right to speak freely and criticize government

Film Production

- Use by the entertainment industry for TV/film

[23] Fernanda Canofre, "A Brazilian journalist uses deepfake to make political satire", Global-Voices, 12 August 2019, https://globalvoices.org/2019/08/12/a-brazilian-journalist-uses-deep-fake-to-make-political-satire/

Harm to Democratic Discourse

The Brookings Institute notes that harmful deepfakes may impact democratic discourse in three ways:[25]

- **Disinformation:** citizens may believe and remember online disinformation, which can be spread virally through social media.
- **Exhaustion of critical thinking:** if citizens are unable to know with certainty what news content is true or false, this will exhaust their critical thinking skills leading to the inability to make informed political decisions.
- **The Liar's Dividend:** politicians will have the power of culpable deniability to suggest that true audio or video content is false, even it is true (in the way that "fake news" has become a way of deflecting media reporting).

  When citizens cannot distinguish between false or potentially false information and facts, this not only impacts their ability to form political opinions but their overall trust in democratic institutions.

Case 2: Nancy Pelosi Cheapfake in the U.S.

One of the most well-known manipulated videos appears to show U.S. House Speaker Nancy Pelosi drunk or ill by slurring her words during an interview in May 2018. The video was first posted by a news Facebook page with 35,000 followers,[26] and identified as "far-right" by some news outlets.[27]

Facebook stated that it would not remove the content as it has no policy calling for the removal of false content, although it was de-ranked after rated as "false" by a third party fact checker.[28]

President Trump reinforced this drunk Pelosi narrative days later by tweeting a different re-contextualised video of Pelosi edited by Fox News with the text "PELOSI STAMMERS THROUGH NEWS CONFERENCE."[29]

This case shows that low tech video manipulation or re-contextualisation — cheapfakes — already has the ability to appear incredibly real and spread rapidly before action is taken by social media platforms.

Image Source: NBC News[30]

[24] Note that these areas may be inter-related. For example, a misleading deepfake about an individual politicians may harm both democratic discourse and national security.
[25] Alex Engler. "Fighting deepfakes when detection fails", The Brookings Institute, 14 November 2019, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/
[26] Joan Donovan and Britt Paris, "Beware the Cheapfakes", Slate, 12 June 2019, https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html
[27] Kathryn Watson, "Trump tweets heavily edited video of Pelosi played by Fox Business", CBS News, 24 may 2019, https://www.cbsnews.com/news/trump-tweets-heavily-edited-video-of-pelosi-played-by-fox-news/
[28] Makena Kelly, "Distored Nancy Pelosi videos show platforms aren't ready to fight dirty campaign tricks", The Verge, 24 May 2019, https://www.theverge.com/2019/5/24/18637771/nancy-pelosi-congress-deepfake-video-facebook-twitter-youtube
[29] Donald J. Trump, Twitter, https://twitter.com/realDonaldTrump/status/1131728912835383300?s=20
[30] Jason Abbruzzese, "Doctored Pelosi videos offer a warning: The internet isn't ready for 2020", NBC News, 24 Mat 2019, https://www.nbcnews.com/tech/tech-news/doctored-pelosi-videos-offer-warning-internet-isn-t-ready-2020-n1010011

Case 3: Alleged Ali Bongo Deepfake in Gabon

Amid a fragile political and economic sitution, Gabonese President Ali Bongo was reported ill and disapeared from public life for several months during the autumn of 2018.[31] Rumours spread that Bongo had died following inconsistant health reports from officials.[32]

Upon his first public video address in January 2019, Bongo appeared visibly different, leading activists and media to allege the video was a deepfake and the president had in fact died.[33] About one week later, a military coup was attempted, although the effort ultimately failed.

Since this alleged deepfake emerged, President Bongo has appeared in a live setting[34] and experts have confirmed the video was not in fact a deepfake.[35]

This case exemplifies the risks that deepfakes pose to the credibility of government, which may result in questioning real sources of information. In this case, the discussion on a possible deepfake may have contributed to the failed coup d'état.

Image Source: Mother Jones[36]

Harm to Individuals

The most real and "urgent threat" of deepfake technology today may not be politics but pornography fabricated without the consent or knowledge of subjects, whose images are superimposed onto video-content.[37] DeepTrace found that **96% of online deepfake videos include pornographic content.**[38]

All pornographic videos identified in DeepTrace's report featured women, particularly from the entertainment or news and media industry.[38] This shows that **women are disproportionately targeted by such harmful deepfake content,** which may have psychological effects on its victims. American actress Zoe Saldana noted this is "violating, offensive and insulting for women."[40]

Other potential uses of deepfake technology to falsely target both public and private individuals may include **defamation, blackmail or revenge porn.**

Case 4: Deepfake porn victim and activist in Australia

18-year-old Australian law student Noelle Martin reverse-image searched herself to find photos of her face photoshopped onto unfamiliar explicit photos. Since she has spoken out, unknown actor(s) have only multiplied their posting.[41] Martin received mixed results when contacting sites, mostly overseas, to remove content. When sites would remove content, new content would pop up again.[42] Martin has since become an activist shedding light on this issue, which has contributed to new laws although no real change has been seen regarding content removal.

Source: ABC News Eliza Laschon

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

[32] Sarah Cahlan, "How misinformation helped spark an attempted coup in Gabon"

[33] Simon Adler, "Breaking Bongo", Radiolab, 27 November 2019, https://www.wnycstudios.org/podcasts/radiolab/articles/breaking-bongo

[34] "Gabon's Ali Bongo makes first live public appearance after stroke", Aljazeera, 16 August 2019, https://www.aljazeera.com/news/2019/08/gabon-ali-bongo-live-public-appearance-stroke-190816173931861.html

[35] Sarah Cahlan, "How misinformation helped spark an attempted coup in Gabon"

[36] Ali Breland, "'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/

[37] Cleo Abram, "The most urgent threat of deepfakes isn't politics. It's porn.", Vox, 8 June 2020, https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell

[38] Giorgio Patrini, "Mapping the Deepfake Landscape"

[39] Giorgio Patrini, "Mapping the Deepfake Landscape"

[40] Emma Kelly, "Avenger star Zoe Saladana slams 'cowards' who posted fake nudes of her online", Metro, 2 December 2019, https://metro.co.uk/2019/12/02/avengers-star-zoe-saldana-slams-cowards-posted-fake-nudes-online-11255317/

[41] Kirsti Melville, "The insidious rise of deepfake porn videos – and one women who won't be silenced", ABC National Radio, 29 August 2019, https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774

[42] Ruby Harris, "How it Feels to Find Your Face Photoshopped Onto Internet Porn", Vice, 18 April 2019, https://www.vice.com/en_au/article/gy4p47/how-it-feels-to-find-your-face-photoshopped-onto-internet-porn

Harm to national/private-sector security

Cybercriminals may attempt to impersonate government officials or CEOs to gain information or financial resources from government and private sector companies. In analysis of the financial sector, Jon Bateman from the Carnegie Endownment for International Peace notes that the current threat of such extortion is low, although this problem will likely increase, noting new cases of deepfakes used for fraud and extortion in recent months.[43]

Case 5: 222,000 Euro Fraud Scheme in Europe

In a recent EU case in March 2019, the CEO of a U.K.-based energy company received a call from fraudsters using an AI-generated voice impersonating the CEO of the German-based parent company requesting an urgent transfer of €222,000. This demonstrates how current deepfake technology is sufficient and available for cybercriminals to use, while society is not entirely prepared.

Image Source: pixel2013 via Pixabay

From a national/domestic security perspective, unverifiable audio-visual content impersonating officials (e.g. threatening invasion) or controversial actions (e.g. burning a Koran or a Bible) has the potential to spark conflict. In this way, it may **"challenge the basis of trust across many institutions".**[44] As an example scenario, UC Berkley Professor Hany Farid suggests: "What if somebody creates a video of President Trump saying, 'I've launched nuclear weapons against Iran, or North Korea, or Russia?' We don't have hours or days to figure out if it's real or not." Adding to this, AI may be used to forge signatures (i.e. readfakes).

[43] Jon Bateman, "Deepfakes and Synthetic Media in the financial System: Assessing threat Scenarios", Carnegie Endowment for International Peace, 8 July 2010, https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237

[44] Greg Allen an Taniel Chan, "Artificial Intelligence and National Security", Harvard Kennedy School: Belfer Center, July 2017, https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

[45] "Creating a data set and a challenge for deepfakes", Facebook

[46] "Creating a data set and a challenge for deepfakes", Facebook

3. COMBATTING HARMFUL DEEPFAKES

WHAT IS BEING DONE?

All relevent actors appear to be investing in **technical detection solutions** via their own capacities and/or funding crowdsourced detection challenges. Companies and some governments have enacted their own **policies** to address this problem. However, policies can only be implemented if technical solutions for quick detection are available.

Academia, Industry and Civil Society

Technical experts from academia are leading the way on detection solutions and are collaborating with companies.[45] Their goal is to develop technology that can successfully identify whether or not a video has been manipulated. Such solutions involve using **machine learning algorithms** to learn from large video datasets to make predictions.

Using **blockchain** technology has also been proposed, although this would require "a significant amount of time and resources" for companies.[46] Startups like Ambervideo.co propose imprinting "cryptographic fingerprints" on every video and then storing "hashes every 30 seconds" using blockchain so it becomes clear when any changes are made.[47]

Human-rights group WITNESS has been highly active on this issue and has recommended some responses. First Draft has created a guide on How journalists can responsibly report on manipulated pictures and video. DeepTrace has mapped the deepfake landscape through a content analysis to better assess the threat.

Social Media Companies

Companies have invested in technical detection challenges to crowdsource new ways of detecting deepfakes on their platforms. Facebook, Amazon Web Services, Microsoft and Partnership on AI hosted a Deepfake Detection Challenge across 3.5 months with 2,114 participants. They are investing $10 million into this "industry-wide effort".[48]

Since early 2020, some companies have enacted new policies for manipulated media on their platforms.[49] However, platforms define and respond in different ways.

Depite these policies some social media platforms, namely **TikTok and Snapchat, actually develop and introduce deepfake technology to the mainstream.** Forbes and Tech Crunch report that TikTok has built a deepfake maker where users can scan and superimpose their face to pre-selected video scenes.[51] Snapchat recently aquired a new startup called AI Factor that helped develop deepfake technology for Snapchat's Cameo function.[52]

Platform	Policy or position?	
Facebook Instagram	Yes – Manipulated Media Policy in Facebook's Community Standards	Facebook will remove media that has been manipulated in a way "not apparent to an average person" AND is the product of artificial intelligence techniques. It does not include videos including satire, omitted words or word order changes.
Twitter	Yes – updated Twitter Rules following an open comments period where they received 6,500 responses.	Twitter will assess the degree to which (a) media is significantly and deceptively altered or fabricated (b) shared in a deceptive manner (c) content is likely to impact public safety or cause serious harm, and can decide to label or remove the content.
YouTube	Yes – deceptive practices policy.	YouTube bans use of manipulated media defined as: "content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm."
TikTok	No specific policy – relevant language in community guidelines on misleading information.	"We also remove content distributed by disinformation campaigns."
Reddit	Yes – Impersonation policy with a form to report cases of impersonation.	"Reddit does not allow content that impersonates individuals or entitles in a misleading or deceptive manner...as well as deepfakes". However, the platform permits satire.
Pinterest	No specific policy – statement to Forbes.[50]	"Misinformation and disinformation aren't allowed on Pinterest, and we remove it if we find it regardless of whether it's manipulated media."
Snapchat	No policy	
WhatsApp	No policy	

[47] "Creating a data set and a challenge for deepfakes", Facebook
[48] "Creating a data set and a challenge for deepfakes", Facebook
[49] Michael Nunez, "Snapchat and TikTok Embrance 'Deepfake' Video Technology Even as Facebook Shuns It", Forbes, 8 January 2020, https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/
[50] Michael Nunez, "Snapchat and TikTok Embrance 'Deepfake' Video Technology Even as Facebook Shuns It"
[51] Michael Nunez, "Snapchat and TikTok Embrance 'Deepfake' Video Technology Even as Facebook Shuns It"; Josh Constine, "ByteDance & TikTok have secretly built a deepfakes maker", Tech Crunch, 3 January 2020, https://techcrunch.com/2020/01/03/tiktok-deepfakes-face-swap/
[52] Michael Nunez, "Snapchat and TikTok Embrance 'Deepfake' Video Technology Even as Facebook Shuns It"

European Union

The EU highlights deepfakes as a disinformation threat in its 2018 disinformation action plan.[53] To tackle online disinformation, including deepfakes, the European Commission committed to several steps including: digital literacy education, investment in fact checking and monitoring capacity, harnessing new technologies, increased election security and media literacy.

In evaluating this strategy one and a half years later, Carnegie's Sarah Bressan asks: **"Can the EU prevent deepfakes from threatening peace?"**[54] She argues that the EU's plan "falls short on the action to prevent potential harm" from this technology. Bressan continues that the EUvsDisinfo challenge was a commendable start to developing detection mechanisms in Europe but considers the programme understaffed. **More resources and government investment in "research and verification tools" are needed** to counteract threats, particularly from Russia. She cautions that such detection tools will only work in EU countries where trust in government is high. In this sense, acountable companies with the proper detection technology are critical.

To improve detection efforts in Europe, **NATO StratCom recommends:** (a) bridging media forensics & stratcom (2) accelerating detection democratisation (3) investing in cognitive research of deepfakes (4) investing in next generation detection.[55]

The EU's upcoming European Democracy Action Plan and Digital Services Act will provide opportunities for regulation and other action to address deepfake threats.

Beyond the EU

Some governments have enacted **"deepfake bans"** or are considering regulation, although such actions must balance adequate measures to protect free speech. Such action may be abused by countries to intentionally limit free speech protections. These bans also fail to address the underlying problem of weak detection mechanisms. The country with the most recorded deepfakes, the United States, has legislated on this issue at the state level, particularly on deepfakes during elections and for pornographic purposes (e.g. Virgina, Texas and California). Legislation has been drafted in other U.S. states and four bills have been drafted but not been adopted at the federal level.[56]

[53] Action Plan against Disinformation, European Union, https://ec.europa.eu/commission/sites/beta-political/files/eu-communication-disinformation-euco-05122018_en.pdf
[54] Sarah Bressan, "Can the Eu Prevent Deepfakes From Threatening Peace?", Carnegie Europe, 19 September 2019, https://carnegieeurope.eu/strategiceurope/79877
[55] Tim Hwang, "Deepfakes – Primer and Forecast"
[56] David Ruiz, "Deepfakes laws and proposals flood US", MalwarebytesLabs, 23 January 2020, https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/

BEYOND THIS PAPER

This paper is part of a three-part series exploring deepfakes as an emerging disinformation threat. In the second paper, Democracy Reporting International will dive deeper by interviewing stakeholders to learn more about their perception of the problem and what needs to be done to minimise this threat. In a final paper, DRI will make recommendations, particularly in relation to disinformation threats in elections.

Auswärtiges Amt

This paper is part of a project funded by the German Federal Foreign Office. Its contents in no way represent the position of the Foreign Office.

ABOUT DEMOCRACY REPORTING INTERNATIONAL

Democracy Reporting International (DRI) strengthens democracy by shaping the institutions that make it sustainable. We support local ways of promoting democracy with impartial analysis and good practices, bringing international standards to life.

The belief that people are active participants in public life, not subjects of their governments, guides what we do. We work with local actors to protect and expand our shared democratic space in a polarised world, regardless of political opinions or personal beliefs.

Find out more at: http://www.democracy-reporting.org

Author: Madeline Brady (with contributions by Michael Meyer-Resende)
Date: 31 July 2020