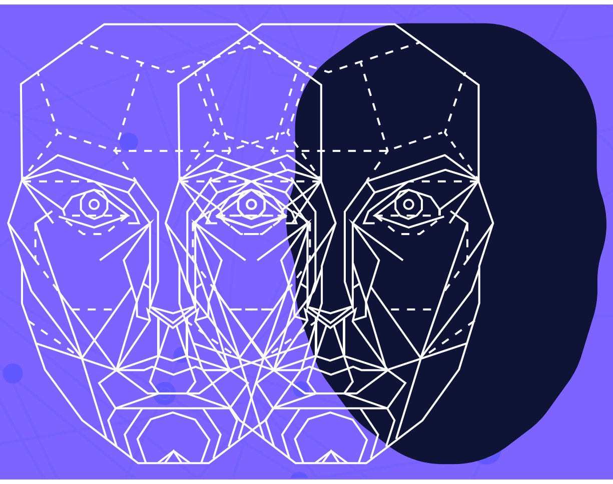
DEEPFAKES AND ELECTIONS: A QUICK GUIDE FOR
ELECTORAL STAKEHOLDERS
DECEMBER 2020
info@democracy-reporting.org
www.democracy-reporting.org

DEMOCRACY
REPORTING
INTERNATIONAL

DEEPFAKES AND ELECTIONS:

A quick guide for electoral stakeholders



Elections are times of high risk, where a perfectly realistic deepfake could have an immediate impact on voters. Although the current state of technology is quite advanced, high-profile cases of deepfakes during elections have yet to materialize.[1] In recent elections, however, less sophisticated video manipulation techniques have been used to spread false information, and these might be just as effective in deceiving the average person. "Manipulated media" will be used as an umbrella term in this guide to refer to both deepfakes and cheapfakes (see definitions below).

Transparent data on the current scope of the problem are lacking, so more monitoring is needed, particularly during elections. Although many social media platforms release some data to promote transparency, such reports do not include information on how they enforce their policies regarding manipulated media. As a result, there is a lack of information on the prevalence of manipulated media materials and the ways in which they are being used to spread false information. More critical eyes are needed, especially in countries where fewer resources are devoted to tackling the spread of false information during elections. Additional research will lead to a better understanding of the problem and improve efforts to hold to account those stakeholders responsible for monitoring the situation.

This guide is aimed at electoral stakeholders, and particularly civil society organisations (CSOs) and election management bodies that monitor social media discourse.[2] It will provide a brief background on the topic and on resources that can aid in monitoring manipulated media during elections:

- How might deepfake technology be used during elections?
- What do we know based on recent elections?
- What can electoral stakeholders do?
- Further resources and tools for civil society election monitors.

For more general information on this topic, see DRI's backgrounder on deepfakes and a deeper assessment of current preparedness measures.

Definitions:

	Manipulated Media[3]	
	Deepfake	**Cheapfake**
Spectrum	High technical sophistication High barriers to entry	Low technical sophistication Low barriers to entry
Technology	AI-based technology used to manipulate AV content	Accessible software or no software used to speed, slow, cut, restage or re-contextualize AV content

[1] Rafael Goldzweig and Madeline Brady, "Deepfakes: How prepared are we? Multi-stakeholder perspectives and a recommendations roadmap", Democracy Reporting International, 30 November, 2020, <https://democracy-reporting.org/wp-content/uploads/2020/11/2020-11-Deepfakes-Publication-No-2-Web-file-1.pdf>.

[2] Madeline Brady, "A new disinformation threat?", Democracy Reporting International, 1 September, 2020, <https://democracy-reporting.org/dri_publications/deepfakes-a-new-disinformation-threat/>.

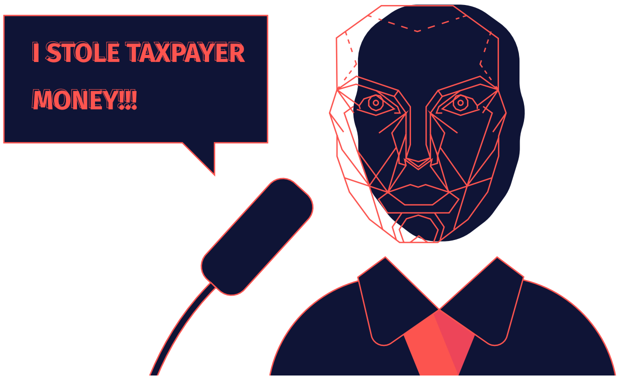
[3] Britt Paris and Joan Donovan, "Deepfakes and cheap fakes: The manipulation of audio and visual evidence", Data & Society, 18 September, 2019, <https://datasociety.net/library/deepfakes-and-cheap-fakes/>.

1 How might deepfake technology be used during elections?

First, as already noted, a very realistic deepfake could have immediate impacts on voters during elections. For example, a deepfake imitating a candidate or news anchor could provide false voting information, causing confusion on election day. Second, deepfake technology may also be used to put false words into a candidate's mouth or to make them appear to do things they have not done, in order or harm their reputation. Fortunately, no such cases of highly advanced imitation for harmful purposes have occurred in recent elections. Third, political actors (candidates, parties, online influencers, etc.) might use the hypothetical threat of deepfakes to call into question factual information harmful to their reputations. For example, political actors might deny the validity of an authentic video, claiming it is a deepfake, in order to avoid responsibility for its contents. Fourth, they might use the hypothetical threat of deepfakes to make unsubstantiated claims to confuse voters. In a real-life example, during the 2020 Georgian parliamentary elections, the ruling party claimed that the opposition would release a deepfake video prior to the election.[4] This claim was made without any providing any real evidence to support it, and there is no evidence that such a deepfake was released.[5]

Scenarios:

1. ELECTION DAY CONFUSION



2. SMEARING A CANDIDATE

3. DENYING RESPONSIBILITY BY CALLING AN AUTHENTIC VIDEO "FALSE"

4. CAUSING CONFUSION THROUGH UNSUBSTANTIATED ACCUSATIONS

[4] Agenda.GE, "Ruling party: opposition has plans to release deepfakes on election day", 26 October, 2020, <https://agenda.ge/en/news/2020/3319>.

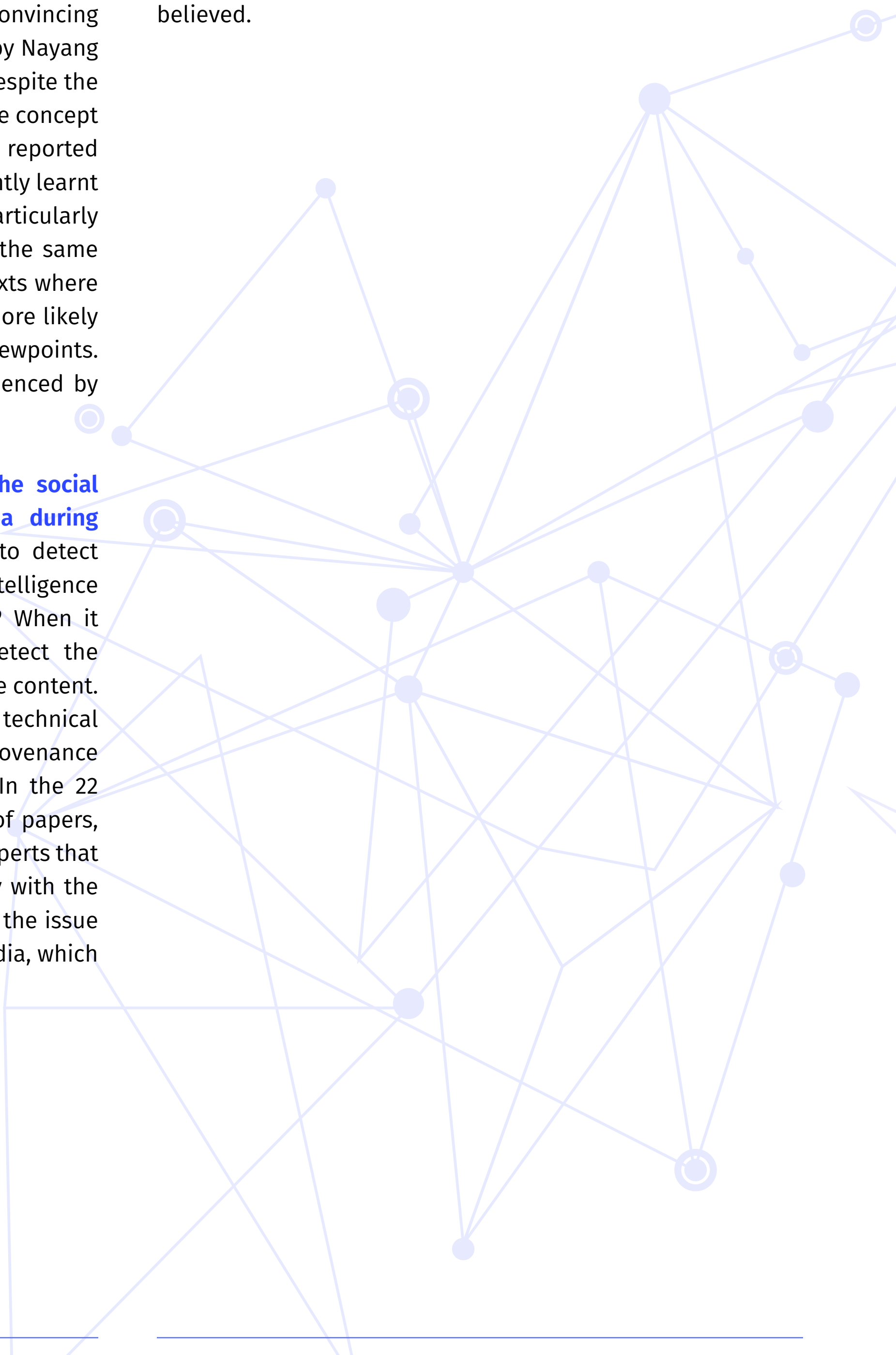
[5] Ibid.

2 What do we know based on recent elections?

The use of less-sophisticated techniques to manipulate video materials for disinformation purposes has already been observed in recent elections. For example, during the 2020 United State presidential election, a video of Democratic Party candidate Joe Biden observing a moment of silence was characterized out of context to support the narrative promoted by his opponent, Republican Party candidate Donald Trump, that Biden was too old to be elected president (Trump regularly referred to Biden as "Sleepy Joe" over the course of the campaign). Such cheapfakes may already be convincing enough to deceive the average user. A recent study by Nayang Technological University, in Singapore, found that, despite the fact that 54 percent of respondents were aware of the concept of deepfakes, "one in three of those respondents reported sharing content on social media that they subsequently learnt was [manipulated media]".[6] Such tactics may be particularly successful when there is large-scale promotion of the same narratives across multiple platforms.[7] Also, in contexts where there is extreme political polarization, people are more likely to believe "information" that confirms their own viewpoints. Thus, they become more susceptible to being influenced by false content.[8]

We are prepared on neither the technical nor the social level to address the use of manipulated media during elections. To date, there are no algorithms able to detect high-quality deepfakes generated with artificial intelligence (AI) technologies with a high degree of accuracy.[9] When it comes to cheapfakes, humans are needed to detect the nuances between satire and truly deceptive and false content. This means AI can't yet provide a quick fix. Other technical solutions to create digital footprints on media (provenance technology) are years away from being in place. In the 22 expert interviews conducted by DRI for this series of papers, there was an overwhelming consensus among the experts that society is not currently prepared to deal effectively with the threat.[10] Most importantly, voters lack awareness of the issue and, additionally, preparedness requires trust in media, which is difficult to guarantee.[11]

It's important to consider unexpected targets during elections. Not only might candidates be the subjects of manipulated media during elections, but journalists, institutions or members of vulnerable groups (e.g., women, ethnic/religious minorities, LGBTQI+ persons, the less-educated) might also be targeted.[12] In some cases, incumbent candidates might themselves be the potential source of such manipulated media.[13] As a result, governments might struggle to make authoritative statements about deepfakes and, even when they do, they might not be believed.



[6] Nanyang Technological University. "One in three who are aware of deepfakes say they have inadvertently shared them on social media." ScienceDaily, 24 November 2020, <www.sciencedaily.com/releases/2020/11/201124092134.htm>.
[7] Joe Pierre M.D., "Illusory truth, lies, and political propaganda: Part 1", Psychology Today, 22 January, 2020, <https://www.psychologytoday.com/us/blog/psych-unseen/202001/illusory-truth-lies-and-political-propaganda-part-1>.
[8] Democracy Reporting International, "Deepfakes and elections: Should Europe be worried?", 11 November, 2020, <https://democracy-reporting.org/deepfakes-and-elections-should-europe-be-worried/>.
[9] The Facebook deepfake detection challenge led to a model with an accuracy of 65 per cent. See: Facebook AI, "Deepfake detection challenge results: An open initiative to advance AI", 12 June 2020, <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>.
[10] Goldzweig and Brady, "Deepfakes: How prepared are we? Multi-stakeholder perspectives and a recommendations roadmap", op. cit., note 1.
[11] Ibid.
[12] Ibid.
[13] Sam Gregory, "What you see is what you trust? How educational initiatives can boost media literacy & fight disinformation", Alliance of Democracies, via YouTube (46:20), 20 May, 2020, <https://www.youtube.com/watch?v=dvPUT9A-lmQ>.

3 What can electoral stakeholders do?

In the short-term, electoral stakeholders must be prepared to react quickly. What can be done?

Recommendations

CONDUCT REAL-TIME MONITORING DURING ELECTIONS

Election management bodies, election observers and civil society:

Look beyond non-text posts when monitoring. Increase their knowledge of tools and resources for monitoring video content. Produce clearer guidelines to help identify and define malicious manipulation.

□ How do manipulated media materials spread across multiple social media platforms?

□ What tactics are being used (e.g., is this highly sophisticated technology?)

□ Are platforms labelling video identified as manipulated and, if so, how?

Campaigns:

Establish “war rooms” to monitor candidate-specific disinformation.

□ How can campaigns counter false narratives about candidates?

CREATE AN ACTION PLAN

Electoral management bodies:

Draft protocols and communications channels for rapid reaction.

□ Who should election officials contact? Should they engage with the public, media or specific governmental agencies?

□ What if government itself is the source?

Media Organisations:

Prepare journalistic practices in the event a deepfake is released during elections.

□ How should the organization report on the topic to avoid amplifying false information further and confusing the public?

□ What if your organization is the target of an attack?

Actions in the long-term are needed to prepare for the threat of deepfakes. See DRI’s recent report for a survey of current actions and further recommendations.

4 Further Resources for monitoring manipulated media

The following sections will provide an overview of how to spot media manipulation, of possible coding categories for monitoring, and of how to communicate your findings.

4.1. Spotting manipulation

Monitoring video content will almost certainly require more time and resources than monitoring text-based posts. For example, watching a full two-minute YouTube video will take longer than reading a 280-character Tweet. Additionally, further technical tools might be needed to successfully verify the video. Here are some resources and tools that can help in getting started:

Spot a deepfake[14]	Resources and tools
A. MAKE A VISUAL CHECK	Eight recommendations from MIT Media Lab:[15] • Check whether the subject is a face – usually this is the case; • Check the cheeks and forehead for overly smooth or aged skin; • Check the eyes and eyebrows for unexpected shadows; • Check the person's glasses for any unusual glare; • Check whether any typical facial hair is missing; • Check whether facial moles look real; • Check whether the person blinks abnormally; and • Check size and color of the person's lips.
B. USE AN ALGORITHMIC DETECTION MODEL	Companies such as Sensity or FakeNetAI offer paid application programming interface (API) services to check whether videos might have been manipulated using AI detection algorithms.[16] They may offer trial accounts and discounts for non-profit customers.

Spotting less sophisticated manipulation	Resources and tools
A. TRY SEARCHING THE WEB FOR ORIGINAL OR RELATED VIDEOS	• Do you see any unique features in the video that you can try searching the web for? For example, background logos behind the speaker to identify where a speech is being made. • If you cannot find the video from the original source, can you find different video angles of the same moment?
B. REVERSE SEARCH SCREENSHOTS FROM THE VIDEO[17]	Try reverse image searching tools, such as: Tineye,[18] Yandex[19] or Google Reverse Image Search
C. WATCH THE VIDEO IN SLOW MOTION OR FRAME-BY-FRAME	Try a tool such as watchframebyframe.com,[20] which allows you to enter YouTube or Vimeo links and watch frame-by-frame or in slow motion
D. TRY USING THE INVID VERIFICATION PLUGIN TO GATHER MORE INFORMATION[21]	InVID allows users to upload a video or post a link to apply reverse image search, to retrieve available metadata (i.e., location), to use a magnifying lens on video and to use other helpful filters to identify manipulation.[22]

[14] Note that such solutions are not absolute. As deepfake technology becomes more advance, such detection techniques may not be 100 per cent accurate.

[15] "Detect deepfakes: How to counteract misinformation created by AI", MIT Media Lab website, <https://www.media.mit.edu/projects/detect-fakes/overview/>.

[16] See: Sensity, Sensity detection API (1.), <https://sensity.ai/api-2/>; and FakenetAI, "Are you prepared for fake media attacks?", <https://www.fakenetai.com/>.

[17] Aric Toler, "Guide to using reverse image search For investigations", Bellingcat, 26 December, 2019, <https://www.bellingcat.com/resources/how-tos/2019/12/26/guide-to-using-reverse-image-search-for-investigations/>.

[18] See: Tineye, "Reverse image search", <https://tineye.com/>.

[19] See: Yandex, <https://yandex.com/images/>.

[20] See: ramebyframe, Watch YouTube and Vimeo videos frame by frame and in slow motion", <http://www.watchframebyframe.com/>.

[21] See: InVID, "InVID Verification Plugin", <https://www.invid-project.eu/tools-and-services/invid-verification-plugin/>.

[22] Ibid.

4.2. Categories for your monitoring

When conducting your research on manipulated media, coding categories may be used to classify social media posts. As a starting point, we have put together the categories listed below. For further resources, it may be useful to review Facebook's rating options for fact-checkers.[23] For further definitions, DRI has also assembled an overview of social media platform policies on manipulated and synthetic media.[24]

Category	Definition
1. Synthetic	Use of AI technology to create original, highly realistic manipulated audio and video. May be extremely difficult or impossible to corroborate.
2. Manipulated	Use of technology (AI or less-sophisticated) to manipulate a video. May be possible to corroborate with original video or slowing down video to spot manipulation. The intent of manipulation may be harmful or satirical.
3. Manipulated and false or misleading	Video falls into category 1 or 2, and may be deceiving and harmful to the average user.
4. Out of context and misleading	Real video clip re-posted with a misleading and out of context title or description. No manipulation technology used.

*Categories 1 and 2 may be combined, as it might not be possible to determine the exact means used to manipulate the video.

4.3. Communicating findings to the public

When making findings public, technical terminology might not be accessible enough to effectively correct the record. The above categories might be helpful for research purposes, but monitors and platforms should consider the appropriate language and terms used to communicate their findings to a general audience. For example, "deepfakes" might confuse or worry people, so using a more easily understood term, such as "digital forgery", should be considered.[25]

Partnership on AI and First Draft are leading the way on research into best practices to label manipulated media.[26] They recommend 12 principles for labeling content effectively, including: avoiding attracting more attention to mis/disinformation, making labels noticeable, and encouraging emotional deliberation and skepticism.[27]

4.4. Additional resources and training

- DRI's backgrounder on deepfakes
- DRI's assessment of deepfakes as a disinformation threat
- Up-to-date developments from WITNESS
- Online challenge from MIT Media Lab to test your visual detection skills
- Online media literacy tool from Microsoft to spot deepfakes

BEYOND THIS PAPER

This paper is the last of a three-part series, in which DRI is exploring deepfakes as an emerging disinformation threat. In the first paper, we provided an overview of the deepfake threat. In a second paper, DRI interviewed 22 experts from civil society, tech companies and academia to understand how prepared we are for this threat, and presented a recommendations roadmap.

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

[23] Facebook, "Rating options for fact-checkers", <https://www.facebook.com/business/help/341102040382165?id=673052479947730>.
[24] Goldzweig and Brady, "Deepfakes: How prepared are we? Multi-stakeholder perspectives and a recommendations roadmap", op. cit., note 1.
[25] Democracy Reporting International, "Deepfakes and elections: Should Europe be worried?", op.cit., note 8.
[26] First Draft, "Partnership on AI & First Draft begin investigating labels for manipulated media", 22 April, 2020, <https://firstdraftnews.org/latest/partnership-on-ai-first-draft-begin-investigating-labels-for-manipulated-media/>.

[27] Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz and Claire Wardle, "It matters how platforms label manipulated media. Here are 12 principles designers should follow.", Partnership on AI, 9 June, 2020, <https://www.partnershiponai.org/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/>.