ACCURATE as of April 1 2020

WITNESS helps people use video and technology to protect and defend human rights – witness.org. For more on our work on deepfakes and preparing better: wit.to/Synthetic-Media-Deepfakes

Deepfakes

Deepfakes make it easier to manipulate or fake real peoples' voices, faces and actions as well as the ability to claim any video or audio is fake. They have become a critical concern for celebrities and politicians, and for many ordinary women worldwide. As they become easier to make, vulnerable communities worldwide need to prepare. In this backgrounder we explore:

- **Technologies**: What are the key technologies and what can they do?
- **Threats:** What are key threats identified globally?
- **Solutions**: What are the potential technical and policy solutions?

What are deepfakes and synthetic media?

Deepfakes are new forms of audiovisual manipulation that allow people to create realistic simulations of someone's face, voice or actions. They enable people to make it seem like someone said or did something they didn't or an event happened that never occurred. They are getting easier to make, requiring fewer source images to build them, and they are increasingly being commercialized. Currently, deepfakes overwhelmingly impact women because they're used to create nonconsensual sexual images and videos with a specific person's face. But there are fears deepfakes will have a broader impact across society and politics as well as in human rights investigations, newsgathering and verification processes.

Deepfakes are just one development within a family of artificial intelligence (AI)-enabled techniques for synthetic media generation. This set of tools and techniques enable the creation of realistic representations of people doing or saying things they never did, realistic creation of people/objects that never existed, or of events that never happened.

Synthetic media technology currently enables these forms of manipulation:

- **Add and remove objects within a video more easily**
- **Alter background conditions in a video.** For example, changing the weather to make a video shot in summer appear as if it was shot in winter
- **Fake face or body movements:** Simulate and control a realistic video representation of the lips, facial expressions or body movement of a specific individual (for example to make it appear they were drunk).
- **Fake lip-sync:** Match an audio track to a realistic manipulation of someone's lips to make it look like they said something they never did
- **Fake voice**: Generate a realistic simulation of a specific person's voice

- **Change a voice's gender or make it sound like someone else**: Modify an existing voice with a "voice skin" of a different gender, or of a specific person
- **Create a realistic but totally fake photo of a person who does not exist.** The same technique can also be applied less problematically to create fake hamburgers, cats, etc.
- **Transfer a realistic face from one person to another, the most commonly known form of "deepfake"**

These techniques primarily but not exclusively rely on a form of artificial intelligence known as deep learning and what are called Generative Adversarial Networks, or GANs.

To generate an item of synthetic media content, you begin by collecting images or source video of the person or item you want to fake. A GAN develops the fake — be it video simulations of a real person or face-swaps — by using two networks. One network generates plausible re-creations of the source imagery, while the second network works to detect these forgeries. This detection data is fed back to the network engaged in the creation of forgeries, enabling it to improve.

As of early 2020, many of these techniques — particularly the creation of deepfakes — continue to require significant computational power, an understanding of how to tune your model, and often significant postproduction CGI to improve the final result. However, even with current limitations, humans are already being tricked by simulated media. As an example, research showed that people could not reliably detect current forms of lip movement modification, which are used to match someone's mouth to a new audio track. This means humans are not inherently equipped to detect synthetic media manipulation.

The current deepfake and synthetic media landscape

Deepfakes and synthetic media are — as yet — not widespread outside of nonconsensual sexual imagery. DeepTrace Lab's report on their prevalence as of September 2019 indicates that over 95% of the deepfakes were of this type, either involving celebrities, porn actresses or ordinary people. Additionally, people have started to challenge real content, dismissing it as a deepfake.

The threats from deepfakes

In workshops led by WITNESS, we reviewed potential threat vectors with a range of civil society participants, including grassroots media, professional journalists and fact-checkers, as well as misinformation and disinformation researchers and OSINT specialists. They prioritized areas where new forms of manipulation might expand existing threats, introduce new threats, alter existing threats or reinforce other threats. They also highlighted the challenges around "it's a deepfake" as a rhetorical cousin to "it's fake news."

Participants in our Brazil/South Africa/Southeast Asia expert convenings and other meetings globally prioritized their main concerns that new forms of media manipulation and increasing mis/disinformation will:

- **Journalists, community leaders and civic activists will have their reputation and credibility attacked**, building on existing forms of online harassment and violence that predominantly

target women and minorities. A number of attacks using modified videos have already been made on women journalists, as in the case of the prominent Indian journalist Rana Ayyub.

- **Public figures will face nonconsensual sexual imagery and gender-based violence as well as other uses of so-called credible doppelgangers.** Local politicians may be particularly vulnerable, as they have plentiful images but less of the institutional structure around them as national-level politicians to help defend against a synthetic media attack.
- **Undermine the possibilities of using video as evidence** of human rights abuses
- **Overload the under-resourced capacities of journalists and fact-checkers** to fact-check deepfakes as they lack the media forensics capacity
- As deepfakes become more common and easier to make at volume, they will **contribute to a fire hose of falsehood** that floods media verification and fact-checking agencies with content they have to verify or debunk. This could overload and distract them.
- **Pressure will be on human rights, newsgathering and verification organizations to prove that something is true, as well as to prove that something is not falsified**. Those in power will have the opportunity to use plausible deniability on content by declaring it is deepfaked.
- **Intersect with existing patterns of rapid 'digital wildfire' where false images are shared rapidly** in WhatsApp and Facebook Messenger and other messaging tools

In all contexts, they noted the importance of viewing deepfakes in the context of existing approaches to fact-checking and verification. Deepfakes and synthetic media will be integrated into existing conspiracy and disinformation campaigns, drawing on evolving tactics (and responses) in that area, they said.

What are the available solutions?

There is a considerable amount of work going on to prepare better for deepfakes. WITNESS is generally concerned that this work on 'solutions' does not adequately include the voices and needs of people harmed by misinformation and disinformation in the Global South and in marginalized communities in the Global North.

Can we teach people to spot these?

It is not a good idea to teach people that they should be able to spot deepfakes or other synthetic media manipulations. Although there are some tips that help spot them now – for example, visible glitches – these are just the current mistakes in the forgery process and will disappear over time. Platforms like Facebook and independent companies will develop tools that can do some detection, but these will only be providing clues. It is important that people also focus on understanding deepfakes within a broader media literacy frame such as the SHEEP approach of the organization First Draft.

SHEEP (an acronym in English) suggests that to avoid getting tricked by online misinformation you should look at. Think **SHEEP** before you share.

SOURCE: Look at what lies beneath. Check the about page of a website or account, look at any account info and search for names and usernames
HISTORY: Does this source have an agenda? Find out what subjects it regularly covers or if it promotes only one perspective.
EVIDENCE: Explore the details of a claim or meme and find out if its backed up by reliable evidence from elsewhere.

EMOTION: Does the source rely on emotion to make a point? Check for sensational, inflammatory and divisive language.
PICTURES: Pictures paint a thousand words. Identify what message an image is portraying and whether the source is using images to get attention.

How do we build on existing journalistic capacity and coordination?

Journalists and human rights investigators need to develop a better understanding of how to detect deepfake using existing practices of OSINT (open-source investigation) and combine this with new media forensics tools being developed.

Are there tools for detection? (and who has access?)

Most of the major platforms and many start-ups are developing tools for detection of deepfakes. However none have yet been released for public easy usage. And there is a challenge that even if tools are developed they are not likely to be made available widely, particularly outside the platforms and media organizations. It is likely that media and civil society organizations in the Global South will be left out of access and it is important to advocate for mechanisms that enable them to have greater access to detection facilities.

Are there tools for authentication? (and who's excluded)

There is a growing movement to develop tools to better track where videos and images come from – from the moment when they are filmed on smartphones, to when they are edited and then shared or distributed on social media. One example of an initiative in this area is the Content Authenticity Initiative from Adobe, Twitter and the New York Times. However, there is a risk that tools that are developed to better help track the origins of videos and authenticate they have not been manipulated may also create risks of surveillance and exclusion for people who do not want to add extra data and information to their photos and videos for fear of what governments and companies will do with this information. These tools will also likely work best with only the newest technologies. WITNESS explores more about these in a recent report on 'authenticity infrastructure': https://lab.witness.org/ticks-or-it-didnt-happen/

What should platforms do?

Social media platforms like Facebook and Twitter have recently released new policies on deepfakes and how they will handle them. WITNESS discusses the Facebook policy at: https://blog.witness.org/2020/01/pros-cons-facebooks-new-deepfakes-policy/ and the Twitter policy here: https://blog.witness.org/2020/01/twitter-facebook-synthetic-media-policy-activist-feedback/.

Key elements of these policies include

- Do they cover just deepfakes or also other forms of manipulated media (e.g. a slowed-down video, or a video that is miscontextualized?)
- How do they define harm from a video?
- Does the intention of sharing matter?
- Do they take down an offensive video? Label it? Provide context on the manipulation? Make it less visible on their site or less easily shared?

Facebook's policy: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ is specific to deepfakes rather than other forms of video or photo manipulation.

Facebook will remove manipulated media when

- "It has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say.
- It is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.

This policy does not extend to content that is parody or satire, or video that has been edited solely to omit or change the order of words. Other misleading manipulated video can be referred to/or picked up by their third-party fact-checkers.

Audio, photos or videos, whether a deepfake or not, will be removed from Facebook if they violate any of our other Community Standards including those governing nudity, graphic violence, voter suppression and hate speech."

Twitter's policy is available here: https://help.twitter.com/en/rules-and-policies/manipulated-media

They indicate you may not deceptively share synthetic or manipulated media (not just deepfakes but also other forms of manipulation) that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context.

They focus on three key questions which determine whether they mau or will label the content or remove it.

1. Is the content synthetic or manipulated?
2. Is the content shared in a deceptive manner?
3. Is the content likely to impact public safety or cause serious harm?

ACCURATE as of April 1 2020

Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled.
✗	✓	✗	Content **may** be labeled.
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

Platforms should be proactive in signaling, downranking – and in the worst cases, removing – malicious deepfakes because users have limited experience of this type of invisible-to-the-eye and inaudible-to-the-ear manipulation, and because journalists don't have the ready tools to detect them quickly or effectively. But addressing deepfakes does not remove the responsibility to also actively address other forms of 'shallowfake' video manipulation like mislabeling a real video or lightly editing a real video

Some questions that relate to the policy: How will both Facebook and Twitter ensure that it is accurately detecting deepfakes? How will it ensure it makes good judgements on when a modification is malicious or whether something is masquerading as satire or parody? How will it communicate what it learns to sceptical consumers? How will it make sure that any decisions it makes are subject to transparency and appeal because of the inevitable mistakes?

What should lawmakers do?

Governments are just starting to legislate around deepfakes. In the US a number of laws have been proposed at the State and Federal level. In the Asia-Pacific region two examples are the laws in the People's Republic of China that ban deepfakes and other 'fake news' and recently proposed legislation in the Philippines. One caution about these laws is when they make a very broad definition of audiovisual forgery and include important forms of free expression like satire, or give broad discretion and power to governments to decide what is 'fake'.