Updated: March 2022

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, WITNESS advocates for responses to current harms and preparation for future threats that centers vulnerable populations globally. In this backgrounder we explore:

- Technologies: What are the key deepfake technologies and what can they do?
- Threats: What are the 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 the tools to create them are increasingly being commercialized. Currently, deepfakes disproportionately impact women because they're used to create non-consensual sexual images and videos with a specific person's face. But there are fears deepfakes will have a broader impact across society, business 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 and 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 ("puppeteering"): 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.
- Create a photo of an event or object from a text description
- Transfer a realistic face from one person to another, the most commonly known form of "deepfake"

[See examples of many of these here]

How does the technology behind them work?

These techniques primarily *but not exclusively* rely on a form of artificial intelligence known as deep learning and the work of 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 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 and create a better and better fake version of the source, for example the face of the person you are mimicking.

As of early 2022, many of these techniques — particularly the creation of realistic face-swap deepfakes — continue to require significant computational power, an understanding of how to tune your model, and often significant CGI post-production to improve the final result. Good examples of a sophisticated deepfake requiring all these inputs are the "Tom Cruise" TikTok videos you may have seen!

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. And recent research found that humans were not capable of spotting realistic faces of people who never existed. We should not assume humans are inherently equipped to detect synthetic media manipulation.

The current deepfake and synthetic media landscape

Malicious deepfakes and synthetic media are — as yet — not widespread outside of non-consensual sexual imagery. Non-consensual sexual deepfakes are unfortunately easily available and generated involving celebrities, porn actresses or ordinary people.

Additionally

- People have started to challenge real content, dismissing it as a deepfake.
- While deepfake satire provides new opportunities for free expression it often treads a fine line with deception.
- Images of 'people who never existed' are being used increasingly to disguise fake accounts in disinformation.

The threats from deepfakes

In workshops led by WITNESS as well as trainings with 500+ people over the past three years, 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 (open source investigation) 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, Sub-Saharan Africa and Southeast Asia expert convenings, and other meetings globally, prioritized their main concerns in relation to how new forms of media manipulation and increasing mis/disinformation will affect their work, societies and communities.

- 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 that national level politicians have to help defend against a synthetic media attack.
- Undermine the possibilities of using video as evidence of human rights abuses and crimes, hampering accountability and justice.
- Already over-loaded and under-resourced journalists will not have the media forensics capacity to capacities of journalists and fact-checkers to fact-check deepfakes
- Human rights, newsgathering and verification organizations will be pressured 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.
- As deepfakes become more common and easier to make at volume, they will contribute to "fire hose of falsehood" strategies that floods media verification and fact-checking agencies with content they have to verify or debunk. This could overload and distract them.
- Deepfakes will intersect with existing patterns of rapid 'digital wildfire', where false images are shared rapidly in WhatsApp, Telegram and Facebook Messenger and other messaging apps.
- Online video conferencing will be vulnerable to manipulation.

In all contexts, the people we consulted 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.

WITNESS research on deepfakes and satire, including a recent report Just Joking! identified a growing usage of deepfakes for powerful social and political criticism. It showed how photorealistic satirical deepfakes lend themselves to:

- Social critique: Parody and satire to critique power that identify societal and political problems and that the audience recognizes as satirical.
- Deliberate misuse: Claims that something is satire when it is disinformation and blaming the audiences for their failure to 'get the joke'
- Accidental misuse: When context is lost, and it is shared as misinformation identifying it as real)

Updated: March 2022

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 existing problems of media manipulation, state violence, gender-based violence and misinformation/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 can 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. If you want to test your ability though go to: https://detectfakes.media.mit.edu/

Platforms like Facebook and independent companies will develop tools that can do some detection, but these will only be providing clues and will not be broadly available in the immediate future. 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 or the SIFT framework.

SHEEP (an acronym in English) suggests that to avoid getting tricked by online misinformation you should "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.

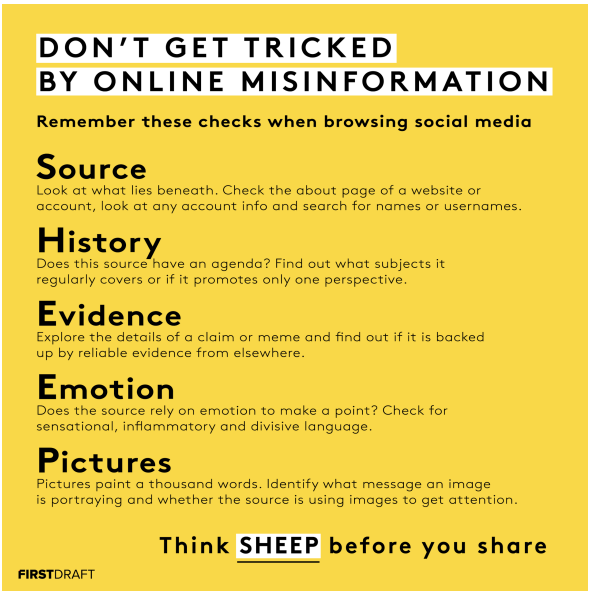
HISTORY: Does this source have an agenda? Find out what subjects it regularly covers or if it promotes only one perspective.

EVIDENCE: Explore the details of a claim or meme and find out if it is backed up by reliable evidence from elsewhere.

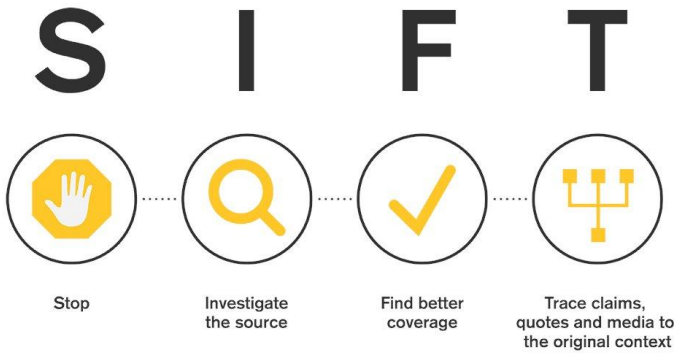
EMOTION: Does the source rely on emotion to make a point? Check for sensational, inflammatory and divisive language.

PICTURES: Pictures paint a thousand words. Identify what message an image is portraying and whether the source is using images to get attention.

Updated: March 2022



SIFT provides another related model of common-sense analysis of suspect information:



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 and combine this with new media forensics tools being developed. Learn more in WITNESS's report on needs, and recent work from the Partnership on AI.

Updated: March 2022

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.

Some tools are starting to be released. An example from Sensity.AI https://platform.sensity.ai/deepfake-detection.

However, we recommend approaching these with extreme caution. Recent broad competitions for deepfakes detection have not come up with models that were effective enough on known techniques or sufficiently applicable to new techniques and most publicly available detectors will be less effective than more closed systems. Detection tools tend to be less reliable if you don't know the technique used to generate the synthetic media, and less reliable on the low-resolution or compressed media we see online. A recent experience of a suspected deepfake in Myanmar shows the challenges of relying on publicly available detectors without accompanying expertise.

Even as robust tools are developed, they will not be made available widely, particularly outside specific platforms and media companies. It is likely that media and civil society organizations in the Global South will be left out and it is important to advocate for mechanisms that enable them to have greater access to detection facilities. WITNESS is arguing for increased equity in access to detection tools, investment in the skills and capacity of global civil society and journalism, and for the development of escalation mechanisms to provide analysis on critical suspected deepfakes.

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 – starting with the moment when they are filmed on smartphones, to when they are edited and then shared or distributed on social media. This 'provenance and authenticity infrastructure' can then show you information on where a photo or video and if/how a photo or video has been changed. This is relevant for both 'shallowfakes' like miscontextualized videos or edited videos as well as deepfakes. You can then use this information to help you make decisions on whether to trust the content. One example of an initiative in this area is the Content Authenticity Initiative led by Adobe, and the recently launched Coalition for Content Provenance and Authenticity (C2PA).

However, there is a risk that tools that are developed to better help track the origins of videos and show how they have 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, or

cannot attribute the photos to themselves for fear of what governments and companies will do with this information. WITNESS has led a Harms, Misuse and Abuse Assessment of the C2PA specifications in order to identify these and other potential harms, and to develop strategies to avert and mitigate them. This 'provenance and authenticity infrastructure' will still be abused, and malicious actors will find loopholes, so the key step moving forward is to bolster a human rights framework, with guardrails against harm, mechanisms for redress, and opportunities to empower critical voices.

What should platforms do?

Social media platforms like Facebook and Twitter have policies on deepfakes and how they will handle them, as well as how they will handle manipulated media more broadly.

WITNESS discusses the Facebook policy here and the Twitter policy here

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 caused by a video?
- Does the intention behind the 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?
- Do they apply to public figures?

Facebook (Meta)

Facebook's policy 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. There have been examples where Facebook has erroneously taken down deepfake satire as misinformation. One example of this occurred in Cameroon when a local academic and activist shared a clearly fabricated video of the French ambassador telling Cameroonians that they never really achieved independence

from France's colonial exploitation. Facebook's third party fact checkers at the French broadcaster France 24 labelled the video partially false, thus nullifying the rhetorical power of the critique.

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

Twitter's policy is available here.

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."

Twitter focuses on three key questions which determine whether they may 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?

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.

TikTok

TikTok "prohibits digital forgeries (Synthetic Media or Manipulated Media) that mislead users by distorting the truth of events and cause harm to the subject of the video, other persons, or society."

Our take

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 ongoing questions that relate to the policies:

- How will both Facebook, Twitter, TikTok and others ensure accurate deepfake detection?
- How will the platforms make a judgement on when a modification is malicious, or whether something is parody, or instead masquerading as satire or parody?
- How will they communicate what they learn to skeptical consumers?
- How will they make sure that any decisions they make are subject to transparency and appeal as they will make mistakes?

What are lawmakers doing?

Governments are just starting to legislate around deepfakes. These laws relate to both non-consensual sexual images as well as usages for deception and mis/disinformation.

In the US a number of laws have been proposed at the State and Federal levels, and in the EU, the AI act endorses labeling of synthetic media for consumer protection.

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'.