WITNESS.ORG, UNITED STATES

PREPARING FOR DEEPFAKES AGAINST JOURNALISM

Sam Gregory
WITNESS.org
sam@WITNESS.org

TAKEAWAYS

- Journalists, especially women, may be targets of humiliation attacks with deepfake technologies. Organizations should prepare themselves for this.

- There are a lot of universities, start-ups and even public service companies developing technologies for deepfake detection. However, in the technology race, the fakers have an advantage.

- Journalistic organizations need strategies to defend their role as distributors of trusted, verified content, and know how to use new anti-deepfake approaches in their workflow.

- Often it will be about proving something someone claims is a 'deepfake' is real, rather than verifying that content is falsified

Celebrities and politicians are already victims of deepfake technology, which can be used to produce revenge porn and other manipulating and often humiliating fake videos. The next in the line of targets are journalists. US-based nonprofit WITNESS.org has held workshops with media houses, technologists and academics to build a strategic approach for fighting this phenomenon.

It might sound like a joke: To put words in someone's mouth, in a video, and see if someone believes whether he or she actually said it.

But this is no joke. Deepfake videos and audios are threatening journalism in two opposite but equally harmful ways: you trust in something you should not, and you don't when you should.

With deepfake technology, you can paste a person's face onto somebody else's body and put a person somewhere where he or she has never been, or insert them into an event that never happened. For journalism, the ability to use machine learning to produce fake audios and videos represents a major paradigm shift. So far, 'seeing and hearing' has been 'witnessing and believing'. Television and radio – by broadcasting real-life events – have over time become trusted sources of news.

Therefore, deepfakes need a strategic approach to be countered, concludes US-based non-profit organization WITNESS.org. It has pioneered and initiated collaboration with news media, with whom it shares a common interest in being prepared for the time when both find themselves in the 'eye of a storm' with a massive attack – which could happen at any time.

Sam Gregory, program director at WITNESS.org, underlines the need for awareness and preparedness, and has coordinated action among those who share the same concern. In his opinion, deepfakes are in fact both a new wave of reputation-based attacks targeting journalists as well as a new form of manipulating online content. "People who've worked with UGC or content like that have had a decade at least of dealing with the same kind of challenges."

And how does one prepare? Like the BBC, DW, The Washington Post and The New York Times have done: by familiarizing themselves with the deepfake technologies. The BBC, for example, has tested faking their own BBC World News presenter, Matthew Amroliwala, speaking Spanish, Hindu and Mandarin Chinese – languages he does not speak. These results, featuring a familiar newsreader's face, were frighteningly good and reminded the organization itself about the dangers of deepfakes.

> "DON'T MOVE TOO FAST TO A 'YOU-CAN'T-BELIEVE-ANYTHING-APPROACH', TO A CONVENTION THAT WILL FUNDAMENTALLY CHALLENGE BOTH TRUST IN JOURNALISM AND TRUST IN COMMUNICATION THAT IS NOT YET JUSTIFIED." (SAM GREGORY)

Sam calls for a common approach involving sharing training datasets and technologies among researchers, platform companies and journalists. Detection should be built into daily processes in newsrooms. In practice, that means toolkits, browser extensions and access on platforms to systems like reverse search for videos. Add lots of training and stress-testing of newsroom processes.

While there are effective technologies for detecting each type of deepfake, often built on useing the same technologies with which the fakes were created, it's already a cat-and-mouse game in which the fakers have an advantage. While journalists have to build credibility and trust, fakers don't have that burden and often also have the technological upper hand.

As detection technologies get better, so do faking technologies. Eye-blinking, for example, was found in summer 2018 to be unnatural in fake videos. Once that finding became public, fake algorithms were shortly thereafter taught to blink naturally.

In the near future, it will become impossible to the naked eye or ear to distinguish real video or audio from synthetic versions.

That's why shared detection and authentication technologies are needed, providing strong signals for human reasoning and good journalistic practice and for sorting through content as it gets easier and easier to make fakes at volume. Journalistic organizations should also prepare themselves for attacks on their teams and journalists who are attacked should be supported and protected.

"Non-consensual sexual images are already a problem – an issue which is often underplayed. This has been used against journalists in a number of cases already," says Sam Gregory.

Sam underlines the role of audio in fakes. "Audio synthesis has been improving, I think, more rapidly than people expected a year ago. And it's more vulnerable in some senses because it has less semantic clues around it," says Gregory.

The key message for journalists from him?

"Don't move too fast to a 'you-can't-believe-anything-approach', to a convention that will fundamentally challenge both trust in journalism and trust in communication that is not yet justified." "It's not that every image is fake."

"We have to avoid playing into the hands of people who want to call everything 'fake news' and to technology solutions that will completely substitute a technical signal for human judgement, rather than complement human judgement. Yet we do have to prepare."

LINK TO WEBSITE

https://blog.WITNESS.org/2019/06/deepfakes-synthetic-media-updated-survey-solutions-malicious-usages/

https://lab.witness.org/projects/osint-digital-forensics/

https://lab.witness.org/projects/synthetic-media-and-deep-fakes/

https://blog.WITNESS.org/2018/07/deepfakes/

http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf

https://arxiv.org/pdf/1806.02877.pdf

AN INTERESTING EXAMPLE OF BBC FAKING THEIR OWN NEWSCASTER, INCLUDING VIDEO OF HIM SPEAKING DIFFERENT LANGUAGES:

https://www.bbc.co.uk/blogs/internet/entries/814eee5b-a731-45f9-9dd1-9e7b56fca04f