REPORT ON A ONE-DAY EXPERT WORKSHOP ON UNDERSTANDING DEEPFAKES AND OTHER FORMS OF SYNTHETIC MEDIA IN SUB-SAHARAN AFRICA

ORGANISED BY WITNESS IN COLLABORATION WITH THE CENTRE FOR HUMAN RIGHTS, FACULTY OF LAW, UNIVERSITY OF PRETORIA

Report compiled by: Ade Johnson

with additional editorial support from Corin Faife

MEETING: 26 NOVEMBER 2019

Table of Contents

1. Executive summary 3

2. Workshop overview, rationale and key objectives 8

2.1. Target participants and experts 9

2.2. Workshop structure and methodology 9

3. Morning sessions: Background context, threat prioritization 10

3.1. Preliminary Session: Welcome remarks and introductions: 10

3.2. Session 1: WITNESS - Brief introduction and background 10

3.2.1. WITNESS — Introduction to deepfakes and rationale for advocacy 11
3.2.2. WITNESS — Identifying prevalence of deepfakes 12
3.2.3. WITNESS — Action plan on the deepfakes problem 12
3.2.4. WITNESS — Background on existing video manipulation 13

3.3. Session 2: Technical perspectives on deepfakes 14

3.4. Session 3: Gender-based violence (GBV) and deepfakes 19

3.5. Session 4: Deepfakes in the context of mis/disinformation in South Africa – Thandi Smith, Media Monitoring Africa 21

Interlude: Group exercises 24

Exercise A: Prioritisation of threat models and vulnerabilities 24

Exercise B: Challenges for media/fact-checkers, social movements, disinformation specialists 25

Afternoon sessions: Solutions, reflections and feedback 28

3.6. Session 5: Solutions and interdisciplinary responses discussed globally 28

3.7. Session 6: Technical perspectives on deepfakes detection — Francesco Marra, University of Naples Federico II 31

3.8. Session 7: Group exercise: Discussion of solutions and needs from a Global South perspective. 34

3.9. Session 8: Feedback on Twitter manipulated media policy 40

4. Workshop outcomes 43

4.1. Way forward and action plan 43

4.2. Immediate next steps 44

1. Executive summary

Deepfake videos emerged in late 2017, and in the intervening time have been labelled as a "looming crisis for national security, democracy and privacy." Deepfakes and other 'synthetic media' are new forms of media manipulation that make it easier to make someone look like they said or did something they never did in a video or audio recording, or to manipulate objects and scenes within a video. Response to deepfakes at the level of legislation and social media platform policy, as well as discussion on the right solutions, has also largely been confined to Europe, North America and China.

However, in the long run there is nothing to suggest that deepfakes will target primarily Western audiences, and it is important that researchers, civil society, journalists and politicians in Asia, Africa and Latin America take an interest in the emerging threat and begin to respond accordingly. This is particularly true as well-meaning "solutions" start to be developed without adequate consultation on their relevancy to threats prevalent in the majority world, or their and efficacy for most countries and situations.

It is against this backdrop that for some time, WITNESS has been involved in advocacy and engagements that focus on deepfakes as an emerging problem. At a time when a range of possible responses are being proposed — from the technical to the political — it is important to promote rights-respecting solutions that protect marginalized communities while also safeguarding freedom of expression online. Therefore, one strand of WITNESS' advocacy work has involved consultation with representatives of groups who are vulnerable to disinformation or actively involved in combating it (e.g. movement organizers, specialists on gender-based violence, journalists, fact-checkers, academics, civil society groups, human rights activists) to solicit input on policy responses to deepfakes and synthetic media.

On Tuesday, 26 November 2019, WITNESS in collaboration with the Centre for Human Rights, University of Pretoria hosted a one-day expert workshop focused on increasing understanding of the problem of deepfakes as well as prioritizing threats and solutions. To the best of our knowledge, this was the first intensive workshop to be held in Sub-Saharan Africa in order to kickstart discussions on deepfakes and explore possible responses in the

African context, and based on African experiences of existing problems. Over the course of the day, the workshop would first build a common understanding of the threats presented by deepfakes and other forms of synthetic media, and then encourage a prioritization of possible interventions from a Sub-Saharan African perspective. This convening followed on from an earlier meeting in Brazil.

The conference was attended by 30 stakeholders and experts involved in digital verification, fact checking, digital rights, human rights advocacy, gender-based violence, movement leadership, technology, journalism and media. Participants that attended the workshop were drawn from different countries in the African continent and Europe. Countries represented from Africa included Ghana, Kenya, Nigeria, South Africa, Uganda and Zimbabwe. The European countries represented included the United Kingdom and Italy.

Over the course of the workshop, participants first learned about the history of deepfakes, including the clear link to nonconsensual pornography and gender-based violence, and their technical characteristics including the state of automated detection software. (Contents of the workshop sessions are outlined in detail in Section 3.)

With this baseline understanding established, participants were given the opportunity to engage in group discussion around possible threat scenarios, and to identify the most plausible and harmful in their contexts. Subsequently they focused on analyzing the technical, policy and educational responses that are currently being proposed. This included sharing their input on a draft synthetic media policy that had been released by Twitter a few weeks before the date of the workshop.

Besides contributing to greater understanding, one of the key outcomes of the workshop was a mapping of the areas where participants' concerns and recommendations differed from those commonly expressed in the US context. As one example, one of the most notable differences was the level of threat assigned to internal vs external actors: In the US, threat perception around deepfakes tends to imagine high-level political interference from foreign actors, i.e. a video attributing false statements to a senior government official. However, representatives of grassroots groups imagined a more pressing threat

coming from agents of their own state, such as videos manipulated to show justification for police activity, to discredit prominent movement leaders, or to scapegoat activists for actions they had not committed in order to provoke a violent response from opposition groups.

In another contrast to discussion in the US, there was a real concern about the potential of deepfakes to incite violence rather than just spread misinformation. These concerns sometimes focused on the potential for rumours to spark mob violence in areas with political and/or ethnic, communal tensions, and also for deepfakes to be used as cover for state violence in either a police or military context.

Many participants identified low levels of media literacy as a problem in combating deepfakes and misinformation more generally. This led to a concern around the use of fake audio and video in ongoing health misinformation campaigns, like anti-vaxxing. As in other countries, media organizations were concerned about the challenge of 'doing more with less' in their journalistic work given low staff numbers and falling revenues.

(Full results of the threat prioritization exercise are presented here.)

In identifying possible solutions and mitigations against the emerging threat, participants emphasized a range of possible interventions spanning from technical to educational to policy based. On the technical side, there was a request for better documentation to outline the range of available algorithmic detection techniques along with their uses and limitations, and to more closely integrate some already existing detection solutions into social platforms. (For example, while visual media spread via private WhatsApp channels cannot be easily searched and debunked by fact-checkers, built-in tools could enable reverse image search functionality directly from the app.) Having seen the range of advanced detection techniques available to computer science researchers, there was a concern that it would be a long time before such techniques were made available to grassroots or indigenous groups or even media outlets, and that efforts were not being made to bridge the gap in the technical sophistication needed to implement them and interpret them.

Media professionals stressed the need for more collaboration and resource sharing in order to respond to the threat effectively and with an efficient use of limited funds. Journalists and fact-checkers identified highly technical fields like media forensics as being an area where resources could be shared between teams and organizations. Building clear channels of communication between actors ahead of time was also highlighted as an area where improvement was needed, and would lead to a more effective response.

Stakeholder groups were broadly agreed on the need for improved public media literacy as a precursor to developing an understanding of sophisticated media manipulation from deepfakes. A call for the translation of training materials into local languages emerged as a key demand, along with a recognition of the importance of working with trusted figures (from community leaders to social media influencers) in promoting a critical approach to online news.

There was also agreement that social media platforms could play a more active role in promoting media literacy by using videos, games and news articles. Finally participants suggested that children should be advised on critical media consumption habits from a young age, with material on disinformation and media manipulation incorporated into school curricula.

(Further discussion of public literacy solutions is presented in Section 3.8)

The final workshop exercises was a feedback session in which participants were presented with questions from Twitter's draft policy on synthetic media, and invited to discuss their preferences in terms of potential responses the platform could take.

For the most part attendees expressed a preference for misleading content to be clearly labelled but only removed in the most severe cases, since removal makes content more difficult for fact-checkers to debunk, and can draw attention to stories that would have faded from public view. Reservations were expressed over the use of "likely to cause physical harm" as the primary test for content removal, since it was unclear how this would be defined, and why other forms of harm (e.g. mental distress) would be ignored.

Lastly participants were sceptical of Twitter's ability to make correct labelling calls in light of social, cultural and linguistic contexts, and worried that under political pressure the company would not always be able to act as an impartial judge of misleading content.

The workshop ended with a call to develop (in the words of one participant) a "careful awareness" of the problem, in which sober appraisal of the threat was used to drive well considered responses rather than any knee-jerk reactions that could have unintended consequences further down the line.

Feedback on specific steps to take in moving forward included:

- Strengthen communication channels between the participant journalists, academics, civil society groups and grassroots organizers ahead of time in order to more effectively debunk deceptive videos when they arise.
- Update journalism training curriculum to include more information on deepfakes and other AI-driven manipulation.
- Look for funding bodies that could cover translation costs for material concerning digital disinformation.
- Initiate further surveys into media forensics capability in journalistic organizations, and begin to develop a plan for creating specialist facilities that could be shared across media outlets.
- Lobby politicians to raise awareness of disinformation as a social problem to be tackled, and to which resources must be allocated.
- Continue to address existing problems with 'shallowfakes' - i.e. mis-contextualized videos and lightly edited content

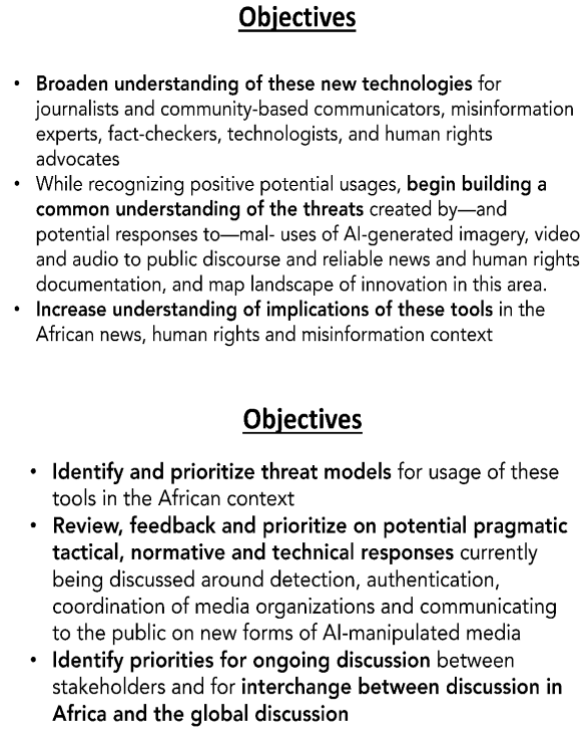
For further reading based on the proceedings of the workshop, see WITNESS blog posts here:

- In Africa, Fear of State Violence Informs Deepfake Threat
- To Fight Deepfakes Build Media Literacy, Say African Activists
- Twitter Released A Draft Policy on Synthetic Media. Here's What Stood Out to the Activists We Consulted.

2. Workshop overview, rationale and key objectives

On Tuesday 26 November 2019, WITNESS in collaboration with the Centre for Human Rights, University of Pretoria hosted an expert workshop that focused on increasing understanding of the problem of deepfakes and other forms of synthetic media, and on prioritizing threats and proposed solutions. The workshop was held at the Centre for Human Rights, University of Pretoria. This is the first workshop to be held in Sub-Saharan Africa on this topic, with the goal of initiating discussions and conversations that focus on the threat from deepfakes and other forms of synthetic media in the Sub-Saharan African context.

The main aim of the workshop was to identify the threats that deepfakes and other forms of so-called synthetic media (a range of ways to modify media using new forms of artificial intelligence) pose in order to proffer solution driven interventions particularly from a Sub-Saharan African perspective.



*(To access all PowerPoint slides from the workshop, click **here**)*

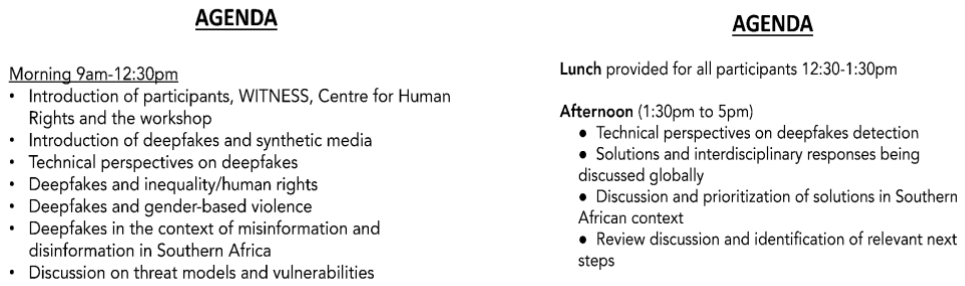
Summarised objectives of the workshop include:

- *Increased understanding*
- *Preparation for the threat rather than panic*
- *Gathering ideas for solution-driven interventions*
- *Network building*
- *Identifying the way forward and next steps*

2.1. Target participants and experts

A total of 30 stakeholders and experts specifically involved in digital verification, fact checking, digital rights, human rights advocacy, gender-based violence prevention, movement leadership, technology, journalism and media attended the workshop. The stakeholders consisted of representatives and experts from different and wide-ranging fields including the media, academia, human rights, technology as well as civil society. Participants that attended the workshop were drawn from different countries in the African continent and Europe. Countries represented from Africa included: Ghana, Kenya, Nigeria, South Africa, Uganda and Zimbabwe. The United Kingdom and Italy were also represented.

2.2. Workshop structure and methodology



AGENDA

Morning 9am-12:30pm

- Introduction of participants, WITNESS, Centre for Human Rights and the workshop
- Introduction of deepfakes and synthetic media
- Technical perspectives on deepfakes
- Deepfakes and inequality/human rights
- Deepfakes and gender-based violence
- Deepfakes in the context of misinformation and disinformation in Southern Africa
- Discussion on threat models and vulnerabilities

AGENDA

Lunch provided for all participants 12:30-1:30pm

Afternoon (1:30pm to 5pm)

- Technical perspectives on deepfakes detection
- Solutions and interdisciplinary responses being discussed globally
- Discussion and prioritization of solutions in Southern African context
- Review discussion and identification of relevant next steps

The workshop was divided into two main sessions: The morning session consisted mainly of theoretical and technical presentations. The afternoon sessions were more interactive and participatory, with participants taking part in discussions and group exercises around threat and solutions prioritization.

3. Morning sessions: Background context, threat prioritization

3.1. Preliminary Session: Welcome remarks and introductions:

The workshop commenced with opening remarks given by the Director, Centre for Human Rights. In response to the opening remarks, the WITNESS team and specifically the Programme Director thanked the Centre for Human Rights for co-hosting and collaborating to hold the workshop.

Following participants' introductions, the workshop began with a spectrum ice breaker exercise. First, participants were asked whether they understood what deepfakes meant. For the second question, participants were asked whether they felt that AI- manipulated media such as deepfakes was something to be concerned and worried about. Although there were a few participants that felt that deepfakes are not an urgent issue currently in Sub-Saharan Africa, it was clear that most participants felt that deepfakes are a matter of urgency.

3.2. Session 1: WITNESS - Brief introduction and background

The objective of the session was to share insights on WITNESS' vision, mission and specifically the work that it does in relation to preparing for the emerging deepfakes problem so as to ensure that participants gain a common understanding of the emerging threat.

WITNESS helps human rights activists, journalists and media practitioners all over the world use video and technology to protect and promote human rights. WITNESS works globally with about 15 team members in the United States and 20 team members across all continents in Europe, Latin America, South East Asia and Africa. WITNESS works with individuals and organisations in supporting the documentation of human rights violations and abuses. This includes providing video evidence of war crimes, police violence, land rights issues etc.

As the world has evolved, WITNESS now works with social media. With this increased volume of evidence comes the problem of an increase in manipulated media and non-accountable social media platforms. The goal of WITNESS is to listen carefully to identify critical challenges and problems in the video-as-evidence field, then advocate for better strategies and approaches that will inform interventions to protect human rights and the integrity of trustworthy information.

3.2.1. WITNESS — Introduction to deepfakes and rationale for advocacy

Deepfakes first emerged around two years ago. Initial rhetoric suggested they would be responsible for an infopocalypse, indicating a collapse of trust in traditional media and the 'end of truth'. Such rhetoric is not beneficial and is potentially concerning for organisations such as WITNESS that rely on trustworthy content as video evidence in their human rights advocacy. As a result WITNESS has begun to engage in advocacy around deepfakes response.

In response to the emerging problem, WITNESS has been involved in a number of key engagements globally. These include: talking to different stakeholders including lawmakers, social media platforms as well as technologists on the deepfakes problem; collaborating with journalists and media practitioners and organisations to identify needs and solutions; as well as its active involvement in international coalitions such as Partnership on AI in efforts to develop solution-driven interventions from different perspectives.

For more information and reporting on these engagements and the different perspectives can be found on https://wit.to/Synthetic-Media-Deepfakes

So far discussion on deepfakes has been largely dominated by the United States and the Global North. The rationale for this workshop is to explicitly counter US dominance and to begin to pursue interventions relevant to the Global South perspectives on deepfakes. To this end WITNESS has already initiated two workshops in Brazil. One workshop was focused on grassroots and community perspectives while the other focused on

perspectives from experts from both grassroots and community organizing as well as technology, journalism, and fact-checking.

The Pretoria workshop will inform this process by ensuring that there is an inclusion of African perspectives in the discussion of potential interventions on deepfakes. An upcoming workshop to take place in Malaysia will similarly highlight Asian perspectives on the problem.

3.2.2. WITNESS — Identifying prevalence of deepfakes

It is clear that deepfakes are not widespread yet. What are more common are shallowfakes where video titles are changed or a false context is added; or simple alterations are made to video like changing speed or a small edit. However, there is a likelihood that deepfakes will be used more in the future as the technology becomes more readily available, and at moments that are particularly susceptible to media manipulation like elections.

While deepfakes in politics are not a problem yet, it is more common to see deepfakes used as an attack on women. In fact, research suggests that 96% of deepfakes online are targeted at harassing, bullying and violating women usually manifesting in non-consensual sexual and pornographic content.

A prominent Indian journalist for example was attacked with a fake video where her face was used on a different body and placed in a compromising sexual encounter that was then shared and went viral on social media and network platforms. There are other examples of applications that have been developed with the capability to simulate nude women, while this same technology is not applied to men. Thus, the gender imbalance in the use of deepfakes becomes apparent.

3.2.3. WITNESS — Action plan on the deepfakes problem

- Understand the problem: There is a clear public benefit in taking steps in ensuring that people are aware of the nature of deepfakes and the threat they pose.

- De-escalate deepfakes rhetoric: This involves downplaying the fearful rhetoric that surrounds deepfakes while still preparing for potential harms that arise. It also includes taking pre-emptive action in order to mitigate potential harms and threats that may develop at a future time. The critical question is therefore, how do we prepare and what is needed in order to be better prepared?
- Identify and share the solutions: Possible solutions to the deepfakes problem must be identified, tested, and shared with groups that may be affected. There is significant momentum for solutions to deepfakes but many relevant stakeholders are ignored and not heard. For instance, there is a need to ensure the inclusion of relevant perspectives and voices of lawyers, human rights activists, technologists, media practitioners and journalists and community members in the interventions. Additionally most of the policy and technical solutions are being proposed in Silicon Valley or Washington DC, Beijing or Brussels. The critical question here is: How do we ensure that the right stakeholders are involved and included in solution seeking initiatives?

3.2.4. WITNESS — Background on existing video manipulation

Deepfakes are a new aspect of an existing problem: Video and audio manipulations have always been part of media, and the ability to manipulate at scale has developed alongside social media. These forms of manipulation have manifested in different ways. Examples include:

- Miscontextualised videos: The vast majority of existing fake videos are in this category. Images and videos in one context are falsely placed in another context with a different caption. Participants were shown a number of examples of videos to depict this.
- Edited videos: This involves removing or rearranging scenes in a video to send a different message. An example is a faked child abduction video in India that became a source of rumours and led to a series of lynchings of innocent people, simply because the educational message at the end was removed. Similar videos have also been seen in South Africa.

- Manipulated videos: This is where videos are deliberately altered in order to be deceptive. An example is where a video of US House Majority Leader Nancy Pelosi was deliberately slowed down in order to make her appear to be impaired or drunk. Participants were asked to give examples of slowed down videos in South Africa. Although there were no immediate examples shared, it was highlighted that unfortunately because people learn from these kinds of manipulation, it would not be surprising if manipulated and slowed down videos of high-profile people begin to show up in South Africa soon.
- Staged videos: They are relatively uncommon, and are videos where people employ actors to enact incidents that are not real. Deepfakes are potentially the new forms of staged videos.
- Flood or firehose of falsehood: This is a general misinformation strategy where torrents of contradictory media are broadcast about the same event, making it difficult to know what to believe. It is usually a more sophisticated strategy involving state actors, and was 'pioneered' by Russia in Ukraine.

3.3. Session 2: Technical perspectives on deepfakes

The objective of this session was to give a simplified understanding and explain the technicalities behind deepfakes, also illuminating the possible uses and misuses.
*Participants were encouraged to ask questions and to share specific examples especially from a South African and Sub-Saharan Afircan perspective to reinforce knowledge. There was a deliberate effort to simplify the technical underpinnings of deepfakes so that participants had a common understanding. (To download PowerPoint slides, click **here**.)*

Deepfakes: How do they work?

Deepfakes go beyond face-swaps. The term includes other forms of synthetic media manipulation, which are produced from what is known as *training data*. This is input from the real world, for example the faces that are shared online and on social media platforms. These images go through a machine learning algorithm used to create a representation of faces and voices. A generative adversarial network (GAN) is a type of deep learning

network that engages in a *cat and mouse* game between neural networks to develop effective forgeries. While one network tries to generate a realistic forgery of for example, someone's face, the other one tries to detect the fake. In other words, one network creates a representation of a face and the second network competes to identify the facs. The one attempts to improve at detection and the other attempts to improve the forgery until a very convincing end product is made.

Deepfake quality is improving daily

To understand the levels of progress of synthetic media over the last five years for example, participants were shown a range of fictional faces. (See below.)

These faces were described as pictures of someone that never existed, generated by a computer. This illustration shows the steady progression of improvements that have been made in face generation to become increasingly convincing. Given this progress it was apparent that participants found the recent results convincing enough to mistake for a real human.

What could be done with these techniques?

Altering videos:

Commercially available tools now have the ability to alter a video by removing images and objects easily from within the frame with a *content-aware fill application* (as exists in applications produced by Adobe and others).

To demonstrate the efficacy of the tool, a video developed by the New York Times Visual Investigations team was shown. Participants were asked to identify the number of policemen in the video. However, one policeman had been digitally removed from the original video in a way that was almost imperceptible.

In another example participants were shown two videos that depict the same scene with changes in weather conditions, one a summer day, the other winter. Participants were asked to identify which of the two videos was the real image. Most participants were uncertain and could not immediately tell the real from the fake weather condition. Again, this exercise showed the sophisticated manipulations that could be made using AI.

Creating a realistic voice or face of a human that never existed:

Participants' attention was drawn to how realistic representations of for instance a cat, a hamburger and a human face can be generated. From the discussions, a comment was raised about the fact that because the representations looked so convincing, it created an emotional response even knowing that the person did not exist. This facial manipulation could be done on a large scale but not all the faces looked as convincing. There were also discussions of how common these application-based manipulations are and how participants might have used them.

Participants gave examples of applications that they had used to manipulate faces and pictures. A common example was a popular app on Facebook. With this application, users were able to use manipulation techniques to see what they would look like when they are older. The facilitator suggested that these fun examples that resonate with people could be useful in kick-starting conversations and building public literacy on deepfakes and other forms of synthetic media.

Simulating and manipulating a representation of a real individual's facial and voice movement:

These tools use face expression modification to map expressions from one face onto another. A video example showed a woman using this technology to make a synthetic

model of her own face with new expressions. The emphasis here was that research is moving towards techniques that require fewer images of an individual face to generate a convincing fake.

Another video depicted real people as 'puppets': by capturing the body movements of a figure in one source video and transferring it to another target, a new and realistic video can show the target person performing actions carried out by the source. It was also mentioned that such techniques are starting to be commercialized.

Next participants were shown a form of deepfake in which the lip movements of one person are matched to the words in a new audio soundtrack. The video example showed how famous footballer David Beckham's lip movements had been matched to an audio discussing in seven different languages, including Swahili and Yoruba, the ill-effects of malaria.

One participant commented that this example showed *deepfakes for good*. Another participant shared that there were similar examples of video dubbing in South Africa where for example, foreign celebrities were made to look like they were singing local songs. Through this participants were urged to begin thinking about the balance between positive uses and serious emerging threats that these techniques could pose.

In summary, what the foregoing discussions prove is the emerging possibilities that synthetic media pose which include:

- The ability to more easily alter video just like with photo editing
- The ability to create a realistic voice or face of a human that never existed
- The ability to simulate and manipulate a representation of real individual's voice, face, movement
- The prospect of an interplay with enhanced micro-targeting, affective computing and other AI-based content creation/activity at volume including text
- Deepfakes could soon be done at greater volume especially because of the tendency for computing power to decrease in cost

Deepfakes: An urgent problem, or not really?

The key question posed to participants was how to respond to the perception that deepfakes do not present an urgent problem. In response it was emphasised that while deepfakes might not appear an immediate problem, there are good arguments that now presents a window of opportunity to act. These include:

- *Easier to* use: The idea that the techniques and tools are getting steadily easier to use, cheaper and adaptable, while quality is improving.
- Likely to be deployed at scale: The future of deepfakes is that they are likely to be used by a large number of people rather than just skilled programmers and will be easier to make at scale rather than requiring lots of detailed, intensive work. Deepfake-like tools are starting to be integrated into apps. At such a time synthetic media will be much more prevalent if we do not act now .
- Economic incentives for use: Along with increased access comes the idea of 'deepfakes as a service,' where the ability to create synthetic video is a skill for hire. As markets develop around the creation and distribution of deepfakes, more people (especially women) will be likely to be targeted without intervention.
- Intersectionality with misinformation and disinformation: Deepfakes intersect with existing mis/disinformation problems and tactics/strategies. Social media has proven to have a politically polarizing effect, with platforms liable to amplify emotionally resonant falsehoods. Realistic manipulated video could easily exacerbate this situation.

Some reflections and questions that emerged from the session's discussions include:

- Deepfakes open up the question of control and the protection of images. Are there pro-active measure that we should take to prevent images online from being used as the input to synthetic videos (e.g. adversarial pixel perturbations)?
- How would children's rights be protected with the emergence of fake images? Does the law cater for images of children to be protected from synthetic media? (This question was prompted by an illustration from Germany, where the police have been allowed to use synthetic images of children in order to have access to child

exploitation sites. Such proposals run contrary to proactive efforts to prevent synthetic media but also present a compelling use case.)

- Some participants were concerned about deepfakes' potential to incite violence as well as spread misinformation — e.g. to be used as cover for state violence particularly in the context of an authoritarian regime. It is possible to imagine an outcome in which the burden is on proving that something is real rather than proving it is fake.
- Another participant drew an analogy between fake currency and deepfakes: With fake currency, it only becomes a concern to everyday people at the point where it cannot be spent. Thus, the tendency is not to be concerned while the fake currency has not been rejected. By analogy, right now it is politicians that are worried about deepfakes, but everyday citizens do not encounter scenarios that impact them so it is hard to create a feeling of urgency. How do we instigate that?
- An additional question asked was how to make tools for detection accessible and what would be the trade-offs, given concerns that detection tools will be used also to improve forgeries?

These questions and reflections served as a useful background to the practical discussions on identifying the threats that deepfakes pose, and the potential interventions.

3.4. Session 3: Gender-based violence (GBV) and deepfakes

*The objective of the session was to expose to the links between gender-based violence and deepfakes. (To download PowerPoint slides, click **here**)*

Tracing the roots of deepfakes

A Google search of the term *deepfakes* exposes how its use has centred mainly on pornography and non-consensual sexual images. The technology was originally developed to superimpose the faces of female celebrities onto pornographic material without consent, and so has always been tied to gender-based violence.

The implications of deepfake activities

Unfortunately, a market has been created where these videos have become accessible, cheaper and easier to develop to the extent that producing deepfakes can be sold as a commercial service. There are examples of a number of applications such as *DeepNude* that use deepfakes in a sexualized context, as well as video hosting sites such as *PornHub* that profit from the distribution of deepfake videos.

A corollary is that deepfakes are often deployed to damage the reputation of women. The example of Rana Ayyub, a female investigative journalist in India that had a simulation of her face pasted into a pornographic video reinforces the point. This example shows that deepfake videos can be used to silence and humiliate women for various reasons including politically motivated harassment.

Research undertaken by *Deeptrace*, a website involved actively in identifying, tracking and investigating deepfakes videos online estimates that about 96% of deepfakes video online is targeted at women.

A question was raised here over what non-consensual pornography/ revenge pornography means. In responding to the question, the distinction was made that some genuine videos are made consensually but shared non-consensually, whereas in the case of deepfakes the target never consents. A point was also made that "revenge porn" is a somewhat misleading term, since the purpose of such videos is usually to shame and intimidate rather than for sexual enjoyment.

In closing, there was discussion of efforts to mitigate the threat of non-consensual pornographic deepfake videos. This includes providing guidelines or step-by-step information/channels on how to remove sexualised videos from the public space and search engines, and to provide information on organizations that provide support.

Participants' reflections and feedback

These discussions precipitated a participant's reflections on work done on gender-based violence in the South African context. It was explained that when it comes to gender-based violence and deepfakes, the intention is to shame women while reinforcing patriarchal tendencies. Their insight was that in the near future deepfakes might be targeted at *slay queens*, who are black women projecting a glamorous lifestyle through social media, and more generally people of colour who are generally more susceptible to harassment.

One participant expressed that the session linking deepfakes to gender-based violence has broadened her understanding of the negative effects of deepfakes technology on women. The examples of *#metoo* and the *total shut down movement* led to women breaking the silence and sharing their stories anonymously online and on social media platforms about the incidences of violence they suffered. Yet, there were cases in the *#metoo* movement where identified perpetrators took the opportunity to open criminal cases against the women. This demonstrates that what had been anticipated as a supposed safety net through which to break the culture of silence can also be turned to a weapon to shame and control.

Participants were also interested in learning more about legal responses to deepfakes. This brought to the fore discussions on legal loopholes that exist where a synthetic representation of an individual is not treated in the same way as a real representation of the individual — although the dangers might be the same as if the real image was shared. An area of possible discussion in this respect is whether South Africa's recent legislative framework in this area offers sufficient protection from synthetic images versus real images.

3.5. Session 4: Deepfakes in the context of mis/disinformation in South Africa – Thandi Smith from Media Monitoring Africa

*The objective of the session was to expose to participants the reality of disinformation and misinformation from a South African context. (To download PowerPoint slides, click **here**.)*

What is Media Monitoring Africa's interest in disinformation?

Media Monitoring Africa is a civil society organisation based in Johannesburg set up in 1993 to analyse the coverage of the first democratic election in South Africa. The organisation analyses news content for issues of diversity, representation, ethics etc. using the evidence for advocacy on a number of issues such as freedom of expression, access to information and most recently digital rights.

Distinctions: disinformation and misinformation

Efforts were made during the session to clarify the distinctions between key concepts such as misinformation and disinformation. It was highlighted that there were deliberate efforts by the organisation to stay away from the term *fake news* because of the contentions that focus on the idea that if it is news, it is not fake and if it is fake it is not news. On one hand, *misinformation* refers to the unintentional spreading of false information. On the other hand, *disinformation* is defined as the deliberate spreading of false information to cause public harm or to gain profit.

Examples of disinformation and misinformation within the South African context

Certain examples from South African were used to support the prevalence of misinformation and disinformation in the country. These examples include entrenched disinformation campaigns about South African Airways. Another popular example cited was the reported fake news on Comprehensive Sexuality Education which was causing undue anxiety to concerned parents.

These examples demonstrated that misinformation can easily shift public narratives based on false information and sensational reporting.

Another example of disinformation shared was the recent xenophobic attacks in the country. These attacks were partly precipitated by the sharing of mis-contextualised videos, for example a burning building that happened in India but was claimed to be in South Africa. Another example were rumours shared on WhatsApp about foreigners kidnapping children in the CBD (Central Business District) in Johannesburg.

Thinking about solutions to disinformation and misinformation

- There is no *one size* fits all approach. It would require a multifaceted and multi-stakeholder approach that incorporates a variety of solutions and interventions.
- It would require fast and consistent fact-checking. The present challenge is that most of the fact checking must be done after the false information has reached the public space. In this respect, advance warning on trending topics from social media platforms would be useful.
- Literacy initiatives and increased awareness on the dangers of misinformation and disinformation is pivotal.
- Media Monitoring Africa has a monitoring platform called Platform 411. A platform that was originally designed as a complaints platform during elections but it now has a broader reach to include issues such as incitement to violence, journalist safety etc.
- There is a wide range of legislative and regulatory framework on these issues but the question is whether it sufficiently covers the threats that deepfakes pose.
- Media organisations and practitioners must also build trust and be credible.

Participants' questions and reflections

Some questions centred on automatic checking, and whether the media can make use of machine learning as a tool for verification and to detect disinformation. However, the challenge with machine learning would be the ability of the machines to contextually differentiate between false claims and claims that were simply misleading or deceptively framed. It is likely that machines would have the same problems that human beings have, for instance distinguishing tabloid news from disinformation.

Interlude: Group exercises

Exercise A: Prioritisation of threat models and vulnerabilities

Objective: Participants were asked to look at threats identified in other workshops and indicate which they felt were the easiest to tackle, the most important, and those that required most collaboration.

As part of a working break, participants took part in a prioritisation exercise. Potential threat scenarios and models identified from previous workshops were placed in a corner board. Each participant was given green, red and yellow coloured dots with which to categorize the threats.

The results suggested that participants generally considered *credibility-based attacks on public figures, human rights defenders and journalists; poisoning the well in a leak with a few well faked videos;* and *gender-based attacks on credibility of human rights defenders and journalists* as priority threats that required the most collaboration and coordination to address.

Threats that were judged important but straightforward to address included *integration of faked audio/video into ongoing public health or conspiracy campaigns (e.g. anti-vaxxing); swamping newsroom operations with unverifiable media; non-consensual revenge porn* as well as *extortion and cyber-crimes*.

Exercise B: Challenges for media/fact-checkers, social movements, disinformation specialists

Objective: in groups, participants begin to brainstorm threats the emerging deepfakes problem poses from a Southern African and Global South perspective. These threats would complement the existing threat map from previous workshops.

For this session, participants were split into three stakeholder groups based on their areas of interest. These groups were:

Disinformation

Media /fact checking

Human rights and social movements

The instructions for the group exercise

Participants were asked to brainstorm threats that deepfakes pose to the particular groups they belong to. Specifically they were asked to discuss:

- Deepfakes threats that participants are most worried about
- How existing challenges reinforce other threats
- Priority threats
- Ignored or otherwise missing threats

Participants in the *disinformation* group discussed the following perceived threats:

- *Potential amplification of social ills, and hate crimes:* Though deepfakes and disinformation, there is the likelihood for existing social problems (e.g. hate speech) to become amplified.
- *Weaponisation by the powerful and influential:* Increased tendency for powerful state actors and even religious leaders to use disinformation to build influence. In other words, ability of the powerful to control and/or impose their narratives and views on the media.
- *Vulnerability of youths and children*: Increased susceptibility of digitally active youths and children to disinformation spread via social platforms.
- *Targeting of marginalized groups*: Other perceived threats include how disinformation narratives can be targeted at marginalized groups, for instance LGBTI activists, to justify violent actions. These groups are more vulnerable and can be targeted with disinformation, while false stories can be used as the rationale for increased surveillance.
- *Environment and climate change*: Climate change inflames tensions by causing conflict over resources, while also being an issue subject to widespread misinformation now it has become a topical issue in South Africa.

Participants in the *media/fact-checking group* discussed the following perceived threats:

- *Reinforcing apathy and low trust in the media*: The emerging deepfakes problem deepens the crisis for journalism, because such threats undermine already weak trust in the media.
- *Loss of resources in news rooms*: Decline in journalism revenue puts pressure on resources in newsrooms, making it difficult to dedicate appropriate time and skill to investigating manipulated media. In many African countries newsrooms must operate across multiple languages and such content requires more resources to verify.

- *Mainstream media competing with social media:* Media organizations now get a lot of their content from social media, for both budgetary reasons and due to the importance of social media for culture. This increase in user-submitted content creates more channels for deliberate hoaxes.
- *Speed at which false information is generated versus speed of debunking:* Even when stories are debunked, the original image or video has usually been widely shared. It can be difficult to get similar reach for a debunk.
- *Misinformation within closed sharing networks*: If the messages aren't on the public internet there is no chance for fact-checkers to see and debunk. Growing usage of e.g. WhatsApp, Telegram, Messenger and other message apps is a problem for fact-checking in general.

Participants in the *human rights and social movement group* discussed the following perceived threats:

- *The intersection between disinformation, propaganda and criminalisation*: State actors' perpetuation of deliberate disinformation acts as the fuel and justification for surveillance, tracking, and criminalization of activists that are already happening as part of the phenomenon of closing civic space.
- *The role of the state as an actor of violence:* Activists already face state violence from police and military actors. Disinformation spread through deepfakes could directly lead to physical harm – and in many places activists are already killed with impunity.
- *Disinformation as a means to push negative narratives*: Synthetic media can be used to push false narratives into the public sphere, which could then undermine support for human rights causes.
- *Accountability vs amplification:* There is a constant challenge to hold bad actors accountable without in the process amplifying the false news. In some cases deepfakes can be amplified through the debunking process.

Afternoon sessions: Solutions, reflections and feedback

The afternoon session began with reflections and feedback from two participants providing perspectives on community activism and digital rights in a Sub-Saharan context. The following points were highlighted:

- *The need for media literacy & local translation:* The use of appropriate terminologies, wording and language is very important to reinforce grassroots understanding of the deepfakes conversation. There is a need to translate the technical terminology into local and indigenous languages that speaks to ordinary people and grassroots groups, and for programs that build general media literacy.
- *Understanding the interrelationships of deepfakes with propaganda, surveillance, criminalisation and vulnerability of activists*: False narratives created by deepfakes can be used as a weapon to incite violence, propaganda and surveillance in ways that make activists vulnerable to attacks. Cities are militarising, people are being killed because of disinformation, misinformation and propaganda. Already incidents are rife of state and government sponsoring of propaganda against verified, factual and authentic information.

3.6. Session 5: Solutions and interdisciplinary responses discussed globally

The objective of the session was to expose participants to available solutions and interventions that are being developed in response to the emerging deepfakes threat. The object was also for participants to think about interventions that are possible, and then contribute ideas on possible solutions.

A brief introduction was provided to initiate the discussions on how to address the deepfakes threat. The following questions were posed to participants as solution areas that could be considered:

- *Can we teach people to spot deepfakes?*

- *How do we build on existing journalistic capacity and coordination?*
- *Are there tools for detection — and who has access?*
- *Are there tools for authentication — and who is excluded?*
- *Are there tools for preventing our images from being used as training data?*
- *What do we want from commercial companies producing synthetic media tools?*
- *What should platforms and lawmakers do?*

What do the possible solution areas mean?

- *Can we teach people to spot deepfakes?* Tips on advising humans on how to spot deepfakes tend to rely on heuristics, such as the idea that "deepfakes don't blink." However, once research had been published which made this claim, soon after deepfake creation algorithms were tweaked to add more blinking. Overall we should try not to rely on a current algorithmic *Achilles heel* as a detection tip, and instead look to multiple signals for detection and ground technical signals in other media literacy paradigms (e.g. SHEEP)

- *How do we build on existing journalistic capacity and coordination?* Collaborations are needed between fact checkers, journalists, human rights investigators and verification specialists in the Global South to find cross-disciplinary solutions building on existing practice. It is therefore crucial to identify the tools that matter and what practices will lead to their adoption. Generally, there is a gap between the cutting edge of the research and the technical skills and resources available to journalists. Participants who work in this area were urged to explicitly name what is needed — for example better tools for finding earlier versions of shared shallowfakes. This would help facilitate a more targeted focus from perceived to actual needs.

- *Are there tools for detection — and who has access?* This question is discussed extensively in the next session. See below (3.7: Session 6)

- *Are there tools for authentication — and who is excluded?* Authentication tools involve tracking the source of an image or other media by providing additional data, tracking if an image or video has been tampered with or edited, and linking it to particular creators. One of the difficulties with tracking is the question of how willing stakeholders would be to give additional information when uploading images. It also triggers pertinent questions such as, would it be safe to divulge additional information considering the sensitivity of work done by specific stakeholders, and would it give too much power to technology in a way that excluded certain groups? These discussions precipitated a question on whether the removal of metadata by social media platforms is intentional. The response indicated that this removal was intentional for privacy and legal liability reasons. However, platforms such as Twitter, Adobe and New York Times are starting to think about shared standards for content attribution and authentication. *(NOTE: WITNESS has a recent report on this area: 'Ticks or It Didn't Happen': Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia')*

- *Are there tools for hiding our images from being used as training data?* This question centres on protection and how to protect individuals from malicious attacks. There are certain experimental methods through which images can be 'protected' from being deepfaked. It involves introducing an adversarial distortion to an image that is imperceptible to the naked eye but prevents a computer vision system from identifying the image. However, the challenge is the ability to introduce such distortions on every image available in the public space and constantly updating such images to make it effective. Thus, it might be better to explore legal as well as technical options with regards to protection.

- *What do we want from commercial companies producing editing and synthesising tools, in other words toolmaker responsibility?* What obligations do commercial companies who make synthetic images hold? As a starting point, the toolmaker should be obliged to ask individuals for their consent to be incorporated into a deepfake. The toolmaker could also show how images built with synthetic tools

are made and also ensure that images made with a synthetic tool can be easily detectable. (for further detail on options see this paper)

- *What should platforms do?* Social are currently in the process of making decisions on what should be done with regards to deepfakes. Using Twitter as a case study, there is an attempt to delve critically into platform responsibility to the deepfakes threat. This question is discussed extensively in a later session. See below (3.8: Session 7).

3.7. Session 6: Technical perspectives on deepfakes detection — Francesco Marra

*The objective of the session was to share insights with participants on technical perspectives that underpin deepfakes detection. This is coupled with responding to the question of whether there are tools to detect deepfakes and who has access. (To download PowerPoint slides, click **here**)*

The session began with a test to gauge participants' ability to detect videos that were not deepfaked but had been modified with facial manipulation applications. Participants were instructed to look carefully at the video clips depicting President Obama, and were asked whether they could spot the fake clip. In fact, all the videos clips shown were fake.

The presenter used photographs and doctored versions of photos of famous political figures to show that the art of doctoring photographs is as old as photographs itself. Before the advent of deep learning, tools to convincingly edit videos were usually reserved for expert users. However, with the advent of AI-based image manipulation, a non-expert user can now learn to create a photorealistic video in ways that were previously reserved for movie VFX studios.

What can be done?

Generally, media forensics tries to provide detection tools that focus on forgery detection as well as forgery localisation tools. Detection tools use various methods in detection including:

- *Physical integrity:* This involves identifying and detecting the deepfakes by exploring physical features. In other words, looking carefully for shadowing or illumination inconsistencies.
- *Visual integrity*: This involves relying on clues for example different eye colours, very smooth areas, artefacts on edges.
- *Semantic integrity:* This involves extracting some kind of semantic information for instance looking out for an inaccurate or wrong connection in time or weather conditions. For instance, a fake image depicting summer and the individual in a particular tropical location when other sources suggest that the event took place in winter and the individual was in a different location.

Challenges with tools and methods that rely on physical visual and semantic integrity

- Physics and visual based techniques rely on features, characteristics and traces that would potentially disappear in the next few years. The improvement in generation quality means that synthetic images will become more photorealistic and convincing.
- Semantic integrity analysis is not always applicable, for example the real image and video might never be posted and there may be no other spatial–temporal information. This means that there is no basis for comparisons and thus it cannot be relied on.

Digital integrity: Digital integrity is considered the most robust detection method. Digital integrity analysis relies on traces of what is often referred to as *digital fingerprints* that are left by a camera or image editing software. All images including videos encode such digital fingerprints and in essence, when an image is modified, the image fingerprint is changed. To try and explain/ reinforce this point simply, participants were shown a clip from the film '*Beyond a reasonable doubt' (2009)*.

Because of the uniqueness of the fingerprint, if an image is modified, the manipulation can usually be highlighted. However, how to reliably extract a fingerprint remains a research

problem. Recently, there has been a proposal to use a technique called a *noise print extractor* as a method to extract to detect deepfakes. However, once an image is compressed and is no longer in high resolution, the fingerprint disappears and becomes impossible to detect (although research in this area is ongoing).

Detection based on a GAN architecture

Apart from extracting fingerprints, another digital integrity technique relies on a database and collection/compilation of deepfakes videos in order to be able to detect a similar deepfakes video using a detector. Deepfakes of famous people and celebrities are more common and are easier to detect because there is likely to be a database of faces for celebrities and the more examples of a face that there is access to, the easier it is to train for detection. The problem still remains that for the faces to be detected, the image needs to be in high resolution and not compressed.

Digital integrity limitations include:

- *Compression*: This is common when sharing images on social network platform which makes it difficult to trace fingerprints
- *Generalisation:* This refers to the difficulty of applying detection methods that work for one generation of algorithm to videos that have been made with another.

Is there a universal deepfake detector?

In ending the session, it was emphasised that unfortunately there is no *universal deepfake detector*. What exists is a fusion of different detectors that help in robustness and accuracy and it more difficult to fool a variety of detectors. A survey of detection approaches is available at https://arxiv.org/abs/2001.06564.

Participants' questions and reflections

One question centred on whether there are easier ways to detect a deepfake without going into technicalities. The response indicated that there are no easy methods of detection because even the technology for detection is still very much in the embryonic stages. This is coupled with the fact that detection tools are not easily accessible.

Another question was whether there could be a solution to the deepfakes problem that worked like an antivirus. However, it was noted that it is difficult to prevent what is unknown. In the same way that people are constantly updating virus source code, deepfakes are constantly evolving and improving.

This discussion triggered comments and reflections from participants. For example, one shared that part of her job involves identifying problematic content on social media platforms — but that is difficult to remove such problematic content because of the fast proliferation rates. Another shared the example of the Christchurch shooting video and the perceived failure to quickly remove this content. This raised the question of the political will of Facebook to detect deepfakes and whether it was commercially viable for platforms to detect and remove deepfakes. The response pointed to the fact that although social media platforms are interested in detecting and removing deepfakes because of the risk they present, the resources dedicated to this detection project are questionable.

Generally, participants' reflections showed that although they found these automated detection techniques interesting at a theoretical level, the concern was that such techniques required a level of technical expertise and sophistication that put them out of the reach of indigenous people and grassroots communities, or to the skills available to journalists.

Participants were also interested in knowing whether technologists and researchers working on detection receive financial support and interest among policy makers. It was pointed out that there was a funding gap between research and journalism. This is worsened by the reality that a higher level of resources went into new forms of video manipulation than detection.

3.8. Session 7: Group exercise: Discussion of solutions and needs from a Global South perspective.

The objective of the session was for participants to begin to explore solutions and needs-driven approaches that could be employed in the mitigation of threats that deepfakes currently presents.

For this group exercise, participants were given the option to choose the groups they would like to work on based on their interests. The groups were:

- *Public media literacy*: Discussing the precursors necessary to build media literacy in order that social media users would be less susceptible to deepfakes.
- *Detection, authentication and journalism:* Discussing the needs of media practitioners and journalists including tools, skills and practice together with detection/authenticity solutions.

From the groupings, it was obvious that a significant number of the participants were interested in the *public literacy* group, with more than half of all participants joining this group. Remaining participants joined the *detection, authentication and journalism* group.

Feedback from the public literacy group

The public literacy group reflected on these questions:

- What approaches would be most useful in helping people understand deepfakes?
- What new or old media literacy skills are required to support public literacy on the deepfakes problem?
- What could different stakeholders do to better support a diverse audience in order to be better prepared for new forms of manipulation?

What approaches would be most helpful in helping people understand deepfakes? Answers listed in order of priority (indicated by number of stars assigned by participants).

Participants' suggestions

- Translate the terminology into local and indigenous languages. There is a need for the localisation of deepfakes awareness campaigns in the communities in order for it to have a far-reaching impact. The importance of demystifying the deepfakes term into simple,

simplified and indigenous language was underscored. This means that deepfakes as a technical term needs to be unpacked so as to be easily discernible to the grassroots populace. **(5 stars)**

- Try to identify and use existing structures of leadership and gatekeepers in the community to educate the community. (**4 stars)**
- Get social media platforms involved in public literacy of deepfakes and other forms of synthetic media **(3 stars)**
- Use short videos and games as a form of educational tool to educate people on the deepfake problem. (**2 stars)**
- Provide education through traditional media: There is need to ensure that traditional and local media such as television, radio and the local newspapers and tabloids are employed so that these traditional media can educate the local community through a bottom-up approach **(2 stars)**
- Incorporate teachings on media manipulation into the school system and the curriculum. It was pointed out that there was a need to empower the youth with the necessary information on the deepfakes problem. **(2 stars)**
- The need to educate the public to develop a critical mindset that questions online content in order to be able to decipher what is factual and verified material and what is not. **(1 star)**
- More generally it was suggested that useful lessons could be drawn from the public information campaign that was done on HIV/AIDS in South Africa.

Following these discussions, the group discussed how to ensure a balance between getting people informed about the threats that deepfakes pose and yet ensuring that cynicism does not override trust when authentic information is shared. It was noted that there is a lack of trust in the media and this raises the question of how they are monitored. There was agreement in the group that there was a need to hold traditional media accountable, but a question over where this accountability should be enforced: should this regulation be done for instance, by the government, private companies, or should media be self-regulated?

Some participants suggested using international standards that are usually developed by the United Nations. However, the argument was that these standards are usually normative but not legally binding. In addition, it was pointed out that media regulation should be localised and not internationalised, though this raises a possible problem in that local regulation can be restrictive to rights. For example, certain African nations have implemented a social media tax and suggested that bloggers should register with the government.

What new or old media literacy skills are required to support public literacy on deepfakes?

Participants' suggestions

- Try to develop simple tools that people can use to understand and check for deepfakes
- Compile pamphlets and easy to read documents that can help to explain deepfakes
- Create short films and videos that can assist in understanding deepfakes
- It was noted that because there is a lack of locally relevant examples of deepfakes in countries in the Global South, public literacy might be difficult. Deepfakes could be seen as a Western problem that is not yet a concern in Africa.
- Participants emphasised the need to deregulate knowledge on detection of deepfakes, so that local experts could contribute to the process
- Media literacy must first be improved to deal with the problem of shallowfakes, which are prevalent in the Global South
- Participants underscored the need for the creation of an indigenous and local resource centre on deepfakes and other synthetic media that could explain concepts in local languages.
- Participants mentioned the need for countries to enact laws that would fight the deepfakes problem. This would require sensitising governments using a two-pronged approach involving literacy building as well as the policy aspects at the same time.

What could be done to support a diverse audience to be better prepared for new forms of manipulation?

Participants' suggestions

- Initiate frank conversations on how the deepfakes threat makes them feel, then turn such conversations into solution-based interventions
- Develop training on detection tools, and push for these insights to be included in coding workshops to introduce these tools to young people.
- Increase the level of research by the academic community, and how the findings are shared.
- Encourage naming and shaming of bad actors as well as whistle blowing
- Engage traditional media such as television and radio in the education of the communities in their local and indigenous languages.
- Get social media platforms involved in deepfakes detection, compile and then widely promote guidelines on detection.
- In general, dedicate resources for research on detection
- One suggestion was to develop detection applications similar to True-Caller. With such an application, people are able to see the identity of a caller. This application is widely used by even non-technical people, and perhaps could be generalized to manipulated videos. However, the indication is that such an application for deepfakes might be difficult to develop because it would require a constantly updated database. Equally, there was also a wariness about using technological solutions to resolve social problems — the argument is that deepfakes are a human problem that requires a human solution.

Feedback from the detection authentication and journalistic group

From the perspective of journalism and fact-checking participants were asked:

- What collaborations are needed around communicating the detection of a deepfake?
- What existing coordination can we build on?
- What tools and skills do journalists, media activists and fact-checkers need to deal with this threat?

- What are the key vulnerabilities and high perception risks surrounding established processes for finding and verifying videos, audios and images?
- What would be most useful for fact-finding / truth-finding practitioners in the issue of image and video manipulation detection?

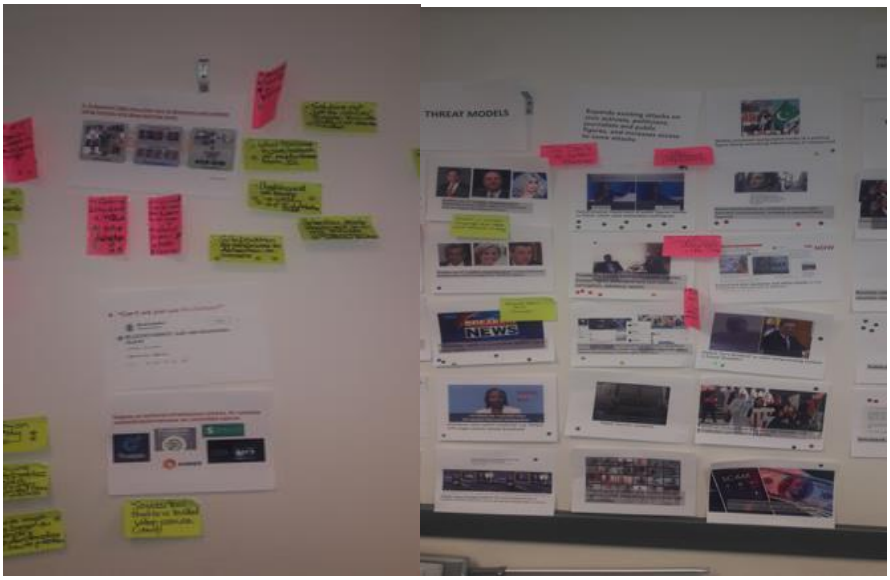
For the detection-authentication aspect, participants were asked:

- What kind of detection solutions would you like to see available?
- What is missing from existing tools to detect manipulated audio and video?
- What do you want platforms to provide in terms of detection and authentication?

Participants' suggestions

- More clarity on tools: Works needs to be done to educate on the type of detection tools that exist and how they can be used. Better documentation of media forensics processes would be helpful here.
- Integrate existing tools into social media: Solutions like TinEye image search could be made directly accessible inside social media platforms. Messaging services like WhatsApp could also do a better job at providing tools to check for old/duplicate images being shared as news, given that journalists can't search through and debunk information in private groups.
- Foster collaboration networks for threat sharing: Developing strong links between organizations will help for early warning on emerging threats and false stories. Crowd-sourcing fact-checks might be appropriate in some cases (and faster).
- Expose how manipulation works: Better documentation of the tactics that are used to influence people via social media would help to inoculate against some mis- and disinformation.

Illustrations of the prioritisation exercises

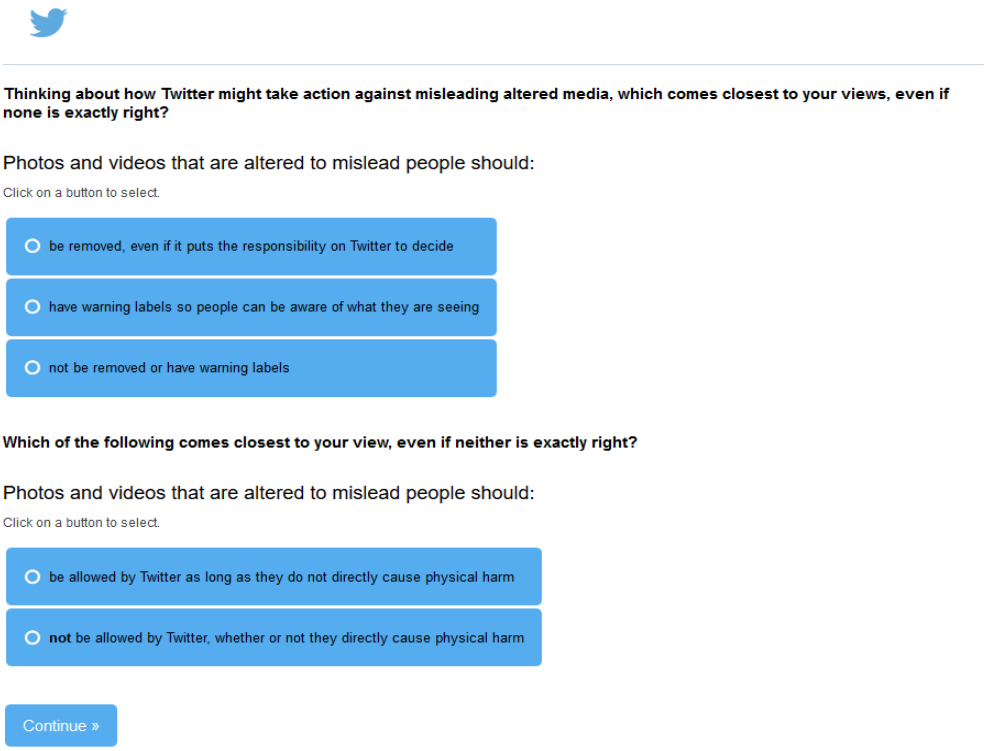


3.9. Session 8: Feedback on Twitter manipulated media policy

The objective of the session was to share with participants a survey by Twitter. The survey was made to assist the platform develop a policy on misleading, altered content including photos and videos that have been falsified with the intention of deceiving and confusing people. Participants were required to critique and offer constructive feedback on the survey. The information and comments from participants would be documented and conveyed to Twitter by WITNESS.

Twitter's focus in the policy questions was on misleading altered content including photos and videos, which includes synthetic media and more. Examples of misleading altered content include: videos that make a person look sick or impaired, videos that add people that were not present, videos that delete people that were present and videos that show events that never happened.

Participants were asked whether it is a good idea to think about misleading altered content as broadly as possible, as Twitter has done. Overall participants felt that it was a good idea to be general in order to anticipate future problems.



1. *The first question was in relation to actions that Twitter should take with regards to misleading altered videos and photos*

   Participants were asked about possible suggestions on what Twitter should do to photos and videos that have been altered to mislead.

   a. No action should be taken
   b. Removed at Twitter's discretion
   c. Have warning labels.

Most participants felt that the photos and videos should have warning labels as removal of these videos could give increased exposure. Other participants felt that such photos and videos should be removed. Some suggested that if a photo or video is deceptive, there is need for people to be informed about the intent to deceive them.

Participants expressed concern over who decides what is misleading and what criteria is used when making the decision. The need to be wary of making gatekeepers of social media platforms like Twitter was underscored.

2. *The next question focused on the relationship to physical harm. In other words, is physical harm an important factor in making decisions to remove videos or warn people?*

Participants were asked whether photos and videos that have been altered to mislead people should:

a. Be allowed by Twitter as long as they do not cause physical harm
b. Disallowed whether or not they do not cause physical harm

There was broad consensus among participants that "physical harm" was a narrow description, and that mental and emotional harm could also lead to physical harm in the future. Some participants questioned why a rule needed to cover physical harm explicitly since incitement to violence is already covered by law.

3. *Another question asked was whether Twitter has the right to alert people when misleading altered media is about to be shared.*

Generally, participants agreed that people should be alerted, but again raised the question of who holds the responsibility for identifying and defining manipulated media. Another point was the challenge of automated detection when content is in local languages, especially indigenous languages. There was also the observation that what is considered misleading media could be contextual and not the same globally.

4. *General reflections and questions from participants on Twitter's policy proposals on manipulated media include:*

- The policy proposal is a good idea, but presents a risk that Twitter could either fail to act or overreact on certain issues because of underlying biases.
- It was noted that the Twitter policy could be clearer, and particularly from the options provided, there is confusion over how physical safety is defined and how a causal link will be identified.
- There was a question of how to deal with propaganda when a fact-check is politicised, and whether Twitter will be able to resist pressure from repressive governments to either censor or re-frame critical content.
- Relevant legislation and definitions from existing human rights law should be used where possible, rather than novel and platform-specific definitions.
- In light of the lack of Twitter employees in the African continent, Twitter users did not feel prioritized for support, and so questioned the company's ability to implement a policy that addressed sensitive issues of harm, truth/falsehood, freedom of expression and local context.

WITNESS subsequently shared this blog post summarizing these concerns, and provided similar feedback directly to Twitter.

4. Workshop outcomes

4.1. Way forward and action plan

As a way to end the workshop and to identify a way forward, participants were asked to identify and reflect on what they considered beneficial to them and to their particular context in the Global South, for example the South African context.

Some feedback on specific steps to move forward included:

- Strengthen communication channels between the participant journalists, academics, civil society groups and grassroots organizers ahead of time in order to more effectively debunk deceptive videos when they arise.
- Update journalism training curriculum to include more information on deepfakes and other AI-driven manipulation.

- Look for funding bodies that could cover translation costs for material concerning digital disinformation.
- Initiate further surveys into media forensics capability in journalistic organizations, and begin to develop a plan for creating specialist facilities that could be shared across media outlets.
- Lobby politicians to raise awareness of disinformation as a social problem to be tackled, and to which resources must be allocated.

4.2. Immediate next steps

To conclude the workshop, the following next steps were outlined:

- *A questionnaire on coordination and feedback would be shared with participants*
- *Participants were asked to engage with resources that WITNESS has produced, participants can identify what is useful and where the gaps lie*
- *Participants to evaluate what interventions are worth working on and WITNESS to give suggestions*

WITNESS would like to thank all who attended the workshop for their participation and valuable contributions to the discussion around deepfakes and synthetic media.

www.witness.org

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