DEEPFAKES: PREPARE NOW

PRACTICAL SOLUTIONS AGAINST THE MALICIOUS USES OF AI-MANIPULATED MEDIA

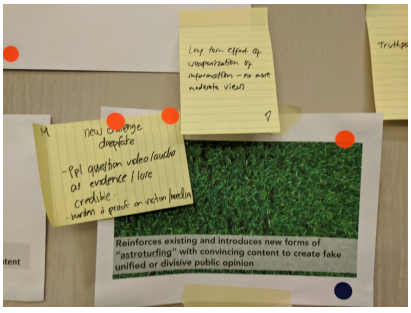
Report on the first SOUTHEAST ASIA-centered expert meeting on deepfakes preparedness and solutions Organized by WITNESS, 3 March 2020, Kuala Lumpur

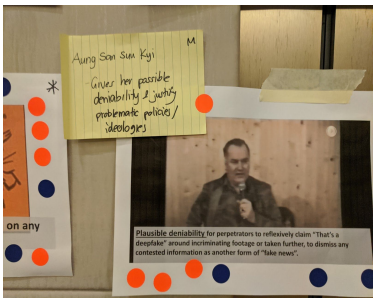
SOUTHEAST ASIA EDITION

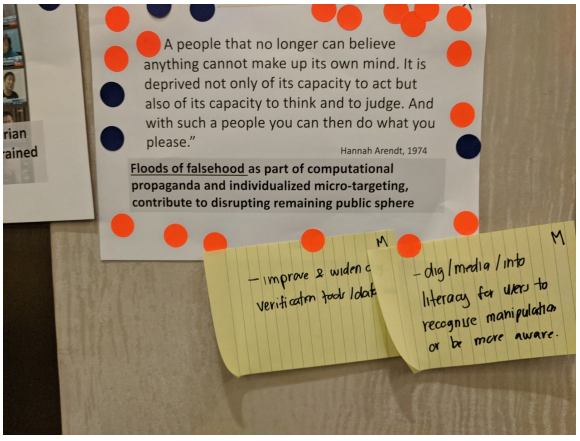
DEEPFAKES: PREPARE NOW

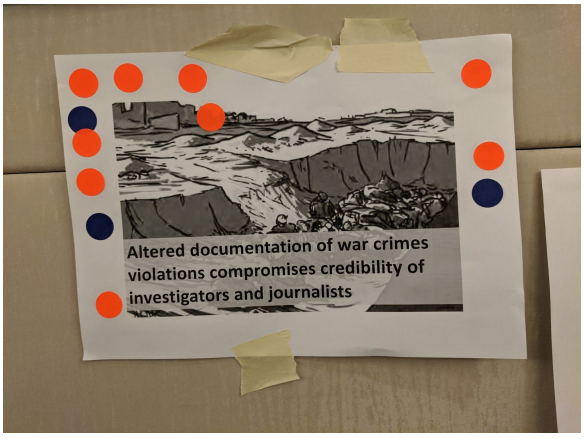
Practical solutions against the malicious uses of AI-manipulated media

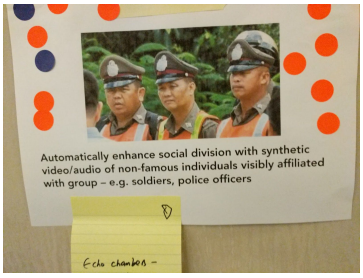
Report on the first Southeast Asia-centered expert meeting on deepfakes preparedness and solutions, organized by WITNESS (Asia-Pacific) 3 March 2020, Kuala Lumpur

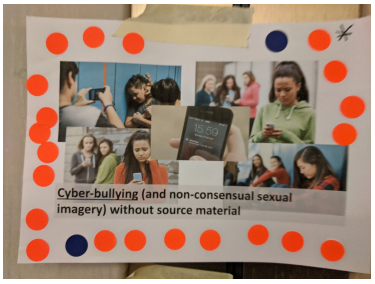












Threat concerns identified by participants (see Group Exercises)

Table of Contents

1. Executive summary

On Tuesday, 3 March 2020, WITNESS and WITNESS Asia-Pacific hosted a one-day experts workshop in Malaysia to increase understanding of the problem of deepfakes and synthetic media in Southeast Asia (SEA) and the broader Asia region, as well as the priority threats and solutions from a regional context, feeding into global efforts. The inaugural SEA workshop was the third such session by WITNESS, coming after those held in Brazil and South Africa last year, in July and November respectively.

The workshop was attended by 35 stakeholders and experts involved in journalism, freedom of expression, human rights advocacy, fact-checking, digital rights, digital verification, filmmaking, movement leadership, international justice, platforms, research and technology. They represented a wide range of fields, namely the media, human rights, fact-checking, documentary, technology, platforms, research, academia, international law, as well as civil society. All were understood to be experts with perspectives relevant to discussing how to better prepare for emerging threats from new forms of media manipulation. Participants who attended the workshop were drawn from different countries in SEA—Cambodia, Indonesia, Burma (Myanmar), Singapore, Thailand, and the host country Malaysia. There were also participants from India, Sri Lanka and Taiwan.

Over the course of the day, the workshop established a common understanding of the threats presented by deepfakes and other forms of synthetic media, then solicited from participants a prioritization of possible regional and global interventions from a SEA perspective.

Starting with the history of deepfakes, including the strong link to gender-based violence (GBV) and an introduction to their technical characteristics, the workshop then took a look at the context of existing visual misinformation and disinformation in the region, in particular Indonesia. (Contents of the workshop sessions are outlined in detail in Section 3.) Next, participants discussed in groups about threat models and vulnerabilities, and identified the most plausible and harmful threats in their contexts. An eye-opening technical briefing via Skype by Francesco Marra of the GRIP team at University Federico II on detection methods helped to inject more thoughtful input into the discussion that followed on SEA prioritization of the

solutions being proposed and rapidly driven forward at a global level and participants' feedback on policies and approaches.

The biggest priority threat in their view is the 'zero trust' world that deepfakes would engender, in particular the threat to the democratic processes of elections and journalism, with public figures as either a target or perpetrator. More specifically, gender-based attacks on human rights activists and journalists were their greatest concern, alongside the related issues of cyberbullying and non-consensual sexual imagery without source material. They feared the threat of violence posed by credible doppelgangers of real people inciting rights abuses or conflict, as well as floods of falsehood.

Spelling out the threats further, they were concerned about the 'truthpocalypse' that would result from the weaponization of information, and the impact on media at various levels. Deepfakes that hijack media brands would erode the trust on which media functions, and their capacity to function as purveyors of truth is also hampered by the lack of detection tools and capacities to counter deepfakes. This relates to their other fear—the lack of preparation to face the threat of deepfakes.

Participants were also wary of the state, not only with regard to policy and legislation concerning the issue, which are driven by vested interests and blighted by disproportionality, but also in its capacity to spread disinformation to spark mob violence and as subterfuge for state violence, particularly in conflict areas such as Burma (Myanmar), West Papua, Sri Lanka, and Southern Thailand.

They were also mindful that deepfakes would further threaten already vulnerable groups, and social division would deepen from deepfake-generated echo chambers and confirmation bias.

They thought the alarm should also be sounded about the transparency and accountability surrounding data that is obtained by deepfake apps.

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

With regard to possible solutions and mitigations against the emerging threat, participants identified several educational and technical actions.

Noticing a growing apathy about truth, the media literacy group mooted an awareness-raising campaign on its importance.

Generally, participants agreed that media literacy is the preventive supplement to fortify the public from falling for deepfakes. It was suggested that context mattered so focusing, for example, on young people and housewives or children in school could be appropriate choices. They cautioned with some audiences of making unnecessary distinctions around how a fake was made - i.e. deepfake versus other forms of manipulation.

Media professionals stressed the need for more interdisciplinary collaborations and resource sharing in order to respond to the threat effectively and with an efficient use of limited funds. In particular, it would be helpful to have a database of experts as a source of reference for fact-checkers globally. Given the gap in the technical capacities, a collaborative training on media monitoring and harm reduction is needed for stakeholders of diverse backgrounds. Platforms could provide them with tools and metadata for deepfake detection.

A group that chose to focus on platforms' roles agreed that content moderation was necessary (without indicating a clear preference on takedown or labelling), but noted the need for supporting smaller platforms as well as focusing attention on functions like a reporting button in WhatsApp. They also noted a need for a mechanism to stop the spread of non-consensual images and better ways to address existing mis-contextualized 'shallowfake' videos and images.

(Further discussion of solutions is presented in Section 4.3)

The final exercise was a feedback session regarding authenticity issues. Participants grappled with dichotomous and contradictory results from the same action pertaining to the extent and limits of tracking the authenticity of media, such as balancing between privacy and accountability needs.

The workshop ended with suggestions on specific steps to take moving forward. These included:

- Simplifying the vocabulary for public education purposes, after which an awareness campaign that includes simple, brief, multi-lingual videos can be held.
- Providing accessibility to detection systems
- Build capacity for shared media forensics
- A database of experts who can help journalists identify synthetic media.

- Updates on what is being done to counter deepfakes, and sharing of best practices around the world, which will require reporting and translation work.

A backgrounder developed for participants in the workshop is available here.

2. Workshop overview, rationale & key objectives

On 3 March 2020, WITNESS held an experts workshop at Armada Hotel, Petaling Jaya, Malaysia, for key experts in Southeast Asia (SEA) and the broader Asia region. Its purpose was to increase their understanding of the problem of deepfakes and other forms of synthetic media, and to ascertain their priority threats and proposed solutions. Coming after two other WITNESS workshops, in Brazil and South Africa, this workshop was the first to be held in SEA on this topic, with the same goal as the aforementioned ones held in the South American and African continents—i.e. to facilitate discussions and conversations on the threat from deepfakes and other forms of synthetic media in the regional contexts,, and to ensure that global discussions occurring largely in the US and Europe on technical and policy solutions are more globally-informed and lead.

The main aim of this workshop was to identify the threats that deepfakes and other forms of synthetic media pose, in order to proffer solution-driven interventions, particularly from a SEA perspective.

Objectives:

- Broaden journalists and community-based communicators, misinformation experts, fact-checkers, technologists, and human rights advocates' understanding of these new technologies.
- While recognizing positive potential usages, begin building a common understanding of the threats created by—and potential responses to—mal-uses of AI-generated imagery, video and audio to public discourse and reliable news and human rights documentation, and map landscape of innovation in this area.
- Increase understanding of implications of these tools in the Asian news, human rights, and misinformation context.
- Identify and prioritize threat models for usage of these tools in the Asian context.

- Review, give feedback and prioritize potential pragmatic, tactical, normative and technical responses currently being discussed around detection, authentication, coordination of media organizations and communicating to the public on new forms of AI-manipulated media.
- Identify priorities for ongoing discussion between stakeholders and for interchange between discussion in Asia and the global discussion.

To sum up, the objectives were to:

- increase understanding of deepfakes and synthetic media;
- build connections and community towards addressing the problem;
- ensure that WITNESS' global advocacy on this issue is informed by local stakeholders' perspectives/needs; and
- identify next steps on this in SEA.

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

2.1. Target participants & experts

A total of 35 stakeholders and experts specifically involved in journalism, freedom of expression, human rights advocacy, fact-checking, digital rights, digital verification, filmmaking, movement leadership, platforms, international justice, research and technology attended the workshop. The stakeholders consisted of representatives and experts from different and wide-ranging fields, namely the media, human rights, film, technology, platforms, research, academia, international law, as well as civil society. Participants who attended the workshop were drawn from different countries in SEA, i.e. Cambodia, Indonesia, Burma (Myanmar), Singapore, Thailand, and the host country Malaysia. There were also participants from India, Sri Lanka and Taiwan.

2.2. Workshop structure & methodology

Morning (9am–12.30pm)	**Afternoon (1.30pm–5.30pm)**
• Introduction of participants, WITNESS and the workshop • Introduction of deepfakes and synthetic media • Technical perspectives on deepfakes and synthetic media • Deepfakes/synthetic media and gender-based violence	• Technical perspectives on deepfakes/synthetic media detection

• Deepfakes/synthetic media in the context of misinformation and disinformation in SEA • Discussion on threat models and vulnerabilities • Lunch 12:30-1:30pm	• Solutions and interdisciplinary responses being discussed globally • Discussion and prioritization of solutions in Southeast Asian 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 by WITNESS. The afternoon sessions were more interactive and participatory via discussions and group exercises around threat and solution prioritization.

3. Morning session: Background context, threat prioritization

3.1. Preliminaries: Welcome remarks & introductions

The workshop commenced with a brief welcome by Arul Prakkash, Senior Program Manager, Asia-Pacific of WITNESS. This was followed by a logistical briefing from workshop facilitator and WITNESS Program Director Sam Gregory, who highlighted healthcare measures in view of the flu season and COVID-19 outbreak. Sam then introduced the purpose of the workshop, went through the agenda, and got participants thinking about their one expectation of the workshop. He set the Chatham House Rule for discussions and social media, and privacy rules regarding photographs. Participants were informed of a rapporteur who is at the workshop to document it for internal records. The other WITNESS staff helping to facilitate small group discussions were Dia Kayyali, Program Manager—Tech+Advocacy, and Tanya Karanasios, WITNESS Director, Global Programs.

Before participants introduced themselves, they made a human spectrogram on the following:

- whether they understood what deepfakes meant;
- how worried they were about them; and
- how worried would they be in 10 years' time.

Those currently worried cited concerns about the negative impact on marginalised communities such as the Rohingyas; one on the opposite end found deepfakes to be no worse than the existing manipulated media. When it comes to the long term, most participants feared that education/literacy would not be able to keep up with the technological advancement that would ease the creation of deepfakes. But though currently worried, a few participants were optimistic that content verification will improve as the seed planted from those present in this workshop grows in the future. The spectrogram produced results that not only showed diverse attitudes and thoughts but also the commonly shared belief of increased worry over time.

3.2. Session 1: WITNESS—brief introduction & background

Slides available here

WITNESS helps human rights activists, journalists and media practitioners all over the world use video and technology to protect and promote human rights. It was formed in the United States (US) after a national conversation on police brutality was sparked from a bystander's video recording of a brutal beating of an African American. Now a global network with team-members in Europe, Latin America, SEA, Middle East and Africa, WITNESS has been working with individuals and organisations for 25 years to provide trustworthy content and powerful narratives on human rights, including providing video evidence of war crimes, police violence, and land rights issues.

Over the past 10 years, WITNESS has also been working with social media platforms to address the prevalence of manipulated media and the issue of accountability. Its focus is on strategic ways to ensure that the technology infrastructure meets the needs of users so that the latter are able to use technology effectively. How it does this is by listening to the grassroots to identify their challenges, facilitating learning and sharing between communities, and engaging with tech companies towards this end.

3.2.1. Introduction to deepfakes

Deepfakes are artificial-intelligence-based synthetic media that falsely depict events or people. Surfacing 18 months ago, they were initially thought to be a harbinger of a 'technocalypse' or 'infocalypse' and other such alarmist headlines proclaiming the end of truth. Realising that such reactionary responses are not useful in the quest for truth, WITNESS began its advocacy work with experts, technology companies, civil society and government authorities to better prepare for deepfakes.

However, much of the conversation has been in the US, where solutions are being developed. Hence, to incorporate the voices from the global south and ensure that their needs are being addressed as well, WITNESS initiated this meeting in SEA, preceded by two others in Brazil and South Africa. All are critical countries in their respective regions and have a strong media and an established civil society, as well as problems with misinformation and disinformation – and in the case of South Africa and Malaysia, these meetings have invited a range of regional stakeholders.

Deepfakes have not appeared much in elections or politics but have been used in gender-based violence. They are not widespread yet, which gives us a window period to prepare for the fight ahead that must include diverse societal voices and particularly people already disproportionately harmed by existing media manipulation. This inclusion is a crucial factor to solving any problem, as past experience has shown.

3.2.2. Background on existing video manipulation

Slides available here.

The workshop participants were shown examples of existing video and audio manipulations, or shallowfakes: the common miscontextualised video or photo; the slightly edited video, affecting messaging and even brands; the manipulated video, such as on Philippine senator Leila de Lima's 2019 speech purportedly supporting drug lords; the staged one, which is rare due to the massive effort required, such as the Rohingya allegedly burning their own villages (photos); and memes. These are increasingly integrated into firehoses of falsehood and other disinformation tactics, with the intention of creating a breakdown of trust.

3.3. Session 2: Technical perspectives on deepfakes

Slides available here

The objective of this session is to establish a common language among the participants.

Deepfakes are produced drawing on training data, i.e. visual and auditory content that are shared online which are used in a machine-learning algorithm to create a representation of faces and voices. An example of such an algorithm uses two competing, adversarial neural networks that function like a cat-and-mouse game to develop effective forgeries: one sets out to build realistic forged visual-auditory content while the other sets out to detect the forgery, providing feedback to the first network to improve the product until the second network is done.

What could be done with these techniques?

- **Altering videos like photos,** i.e. pulling out visual elements with a content-aware fill application, which is the simplest form of manipulation. The tool is commercially available, produced by Adobe (as "Content Aware Fill" for video) and others.

  To demonstrate the efficacy of the tool, two examples were given. The first is a video developed by the New York Times Visual Investigations team showing a scene with a number of policemen. Participants were asked to guess how many policemen there were, but no one could spot one who had been digitally removed.

  The second example was of a landscape in two different weather conditions: summer and winter. Asked to identify which of the two videos was the actual one, participants were split. Similar first example, a visual element, i.e. the snow, was added to the original video.

  Tip: To verify such content, check it against the factual context of the identified location.

- **Creating a realistic voice or human face that never existed**—there are now websites that offer as many as 100,000 faces of people who do not exist, free for downloading. These computer-generated representations of a real person of any age, ethnicity, and even gender, or any object, such as a cat.

Participants were shown how sophisticated the technology has become over four years, producing faces that are increasingly convincing. One implication is that fake accounts can be created without stealing other people's pictures anymore. Participants were shown a recent example of a network using such pictures for fake social media accounts. It is also used in fun social media apps to show what one could look like at a different age or as a different gender.

- **Simulating and manipulating a representation of a real individual's facial and voice movement**—there is a tool that matches the video image of the actual person's lips to the words on another audio soundtrack. Participants were shown a manipulated video of footballer David Beckham's speaking in seven languages, including Swahili and Yoruba, about the dangers of malaria. A more recent example shown was of Indian politician Manoj Tiwari's two deepfaked videos of his speaking in English and a local dialect, which was the first widespread application in politics that caused much confusion and controversy. The tool is available via open source and commercially for corporate clients. One useful application for it would be in the movie industry for dubbing dialogue.

  Participants were also shown a project that makes synthetic models of the faces of famous figures such as Isaac Einstein and Salvador Dali, giving them new expressions. Of concern is that the progress in research and development is creating techniques that require increasingly fewer images of a face as the training data used to generate a convincing fake of it, from 500 images previously to a mere 16 images now.

  Next was a video showing a PhD student's project that simulated a professional dancer's movements and transferred them onto her own body, producing a new and realistic video of herself dancing like an expert. The idea behind this deepfake is akin to using actual people as puppets. The negative application would be to simulate an unflattering or even criminal behaviour on certain people.

- There is also **audio simulation** with a similar concept of placing training data in the algorithm and fine-tuning to create a realistic voice

- A recent research is on **text-based manipulation to a talking-head video**, i.e. changing the words and lip movements without having to edit the video.
- It was stressed that most deepfakes are **lip-sync dubs**, not face swaps, though the latter are more attention-grabbing and well-known. An example of the former is ex-US president Barack Obama's double-edged warning about deepfakes; examples of the latter are Russian president Vladimir Putin's face being superimposed on US First Lady Melania Trump's and techpreneur Elon Musk as a baby.

- **Fake texts** are another development, alongside experiments of submitting AI-generated comments to calls for feedback regarding laws, which somewhat alarmingly went undetected.

In summary, the types of deepfake are:

- Alteration within a video, as has been possible with photos
- Creation of a realistic voice or face of a person who never existed
- Simulation and manipulation of a representation of a real individual's voice, face, movement etc.
- In interplay with other technological trends such as enhanced micro-targeting (of minority communities) and affective computing (i.e. understanding emotions and responding to them).

Why worry?

In the spectrogram at the start of the workshop, some participants had thought that deepfakes were not an immediate problem or no worse than the current manipulation of visual content or had more negative implications on gender-based violence than politics. However, Sam argued that there should be cause for serious concern now, as deepfakes:

- are getting easier to use, as increasingly improved technology meant less training data and lower technical skills are required;
- are getting better in terms of quality of simulation;
- are moving to mobile platforms, thus increasing accessibility, and following that:
- are becoming an 'as a service';

- are likely to be used at scale, as opposed to being one-off, highly invested efforts associated with specialised high-tech products (Sam showed a video shared by a participant at another workshop in Myanmar, which used a popular Chinese app called Zao to render the Chief Minister of Yangon performing a K-pop dance late last year);
- will be an additional challenge to people's cognitive ability to deal with misinformation. Going by the trend of online social media, deepfakes will also be weaponized as part of misinformation and disinformation strategies.

Participants' questions and comments

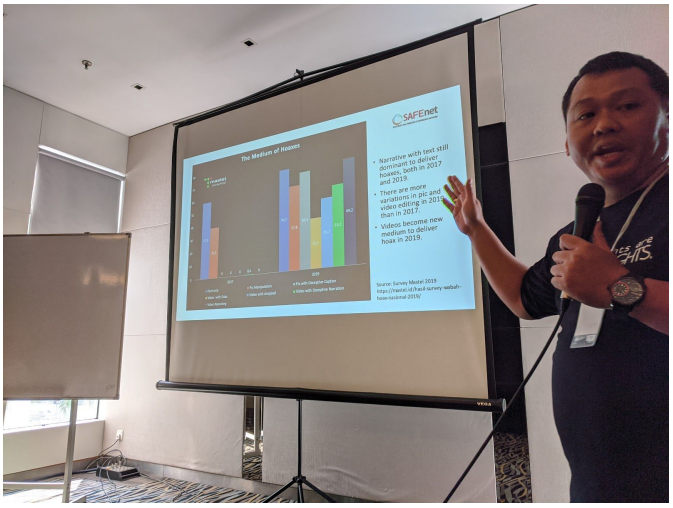
To a question about the minimum number of images to create a deepfake, Sam said it depends on what function is being asked of it: To manipulate faces, researchers are experimenting with a single shot; in tested work, it requires 15-20 images. Face swaps still need a lot of images, though the trend is towards less images. The products will include models that can be easily applied on different peoples.

A participant commented that an emerging threat is how middle-eastern governments are taking advantage of the suspicion over deepfakes to manipulate people's perception and to win them over to the side of governments as part of a whitewash or cover-up of their own scandals.

3.4. Session 3: Deepfakes in the context of misinformation in SEA—Damar Juniarto, SAFEnet

The objective of the session was to provide participants with the baseline on disinformation and misinformation in SEA. (To download PowerPoint slides, click here.)

Southeast Asia Freedom of Expression Network (SAFEnet) is a regional network protecting digital rights in terms of access, expression and safety. Based in Bali, Indonesia, the organisation has published a report, "How "Hoax" Hysteria Used to Justify Tighter Internet Laws and Repress Free Expression in Southeast Asia," whose main point was that hoaxes or



disinformation in the region operate within a framework where existing laws already inhibit freedom of expression. Repressive governments are exploiting concerns regarding hoaxes to control social media, which is the last bastion of freedom of expression to the people in the country, via censorship and surveillance.

A mapping of regulations and government measures in SEA shows the following in place:

- Brunei: Article 88 and 258, Criminal Code; Decree 72
- Cambodia: Regulation on online web and social media, 2018
- Indonesia: Article 27:3, Electronic Information and Transactions Act (in its Bahasa Indonesia acronym, UU ITE); Article 28:2 UU ITE; Article 14-15 No. 1 Act, 1946; Cyber Drone 9, Ministry of Communication and Information Technology; National Cyber and Encryption Agency (in its Bahasa Indonesia acronym, BSSN)
- Myanmar: Section 66(d) of the Telecommunications Law
- Myanmar: Social Media Monitoring Team, 2018
- Philippines: Social Media Regulation Act, 2017
- Singapore: Protection from Online Falsehoods and Manipulation Act (POFMA), 2019
- Thailand: Article 112 of Criminal Code (lèse-majesté)
- Malaysia's Anti-Fake News Act 2018 was repealed in 2019.

Focusing on Indonesia, the top 3 false news topics are politics, race/religion/tribe, and health, according to findings from the two national surveys in 2017 and 2019 conducted by ICT Society (MASTEL). What changed in between the two periods was the format: while texts and manipulated pictures remained dominant, *in 2019, videos emerged as the new medium for false information*. Social media is the dominant channel and, increasingly, instant messaging, followed by websites.

How have the different stakeholders in Indonesia managed such false information? The government has a fact-checking website, stophoax.id, and another one for digital resources, siberkreasi; it also requests takedowns from platform providers, creates new laws, and enforces internet shutdowns. The media fraternity has a fact-checking hub, cekfata.com. Tech companies

either voluntarily or upon requests take down accounts on their platforms. Civil society organizations have a fact-checking website, turnbackhoax.id, and promote ethics, digital literacy network, and digital rights.

The following are some notable online false information cases:

- Stand-up comedian Ge Pamungkas faced online threats and a police report was made against him in 2017 after an edited video of his performance that allegedly insulted Islam was shared on a website catering for Islamic radicals. SAFEnet provided legal help, advocacy as well as protection from online persecution.
- #boikotindosat 2017 was an online campaign urging for the boycott of giant telecommunications company Indosat for allegedly supporting the Christian then governor of Jakarta, Ahok. Muslim cyber troopers used fake accounts to generate support for the campaign.
- An initially real Twitter account, Otalapaw, started to send spam; it was later discovered to be a dead person's hijacked account operating as a spam bot.
- In 2018, pilot Pribadi Alisudarso faced online threats and was grounded following a tweet by an influencer, Yusuf Muhammad, which asserted that the former had not only allowed a passenger to speak on the intercom of a plane that he allegedly flew but also declared himself as desiring a martyr's death, and was a stalwart of the radical cleric Felix Siauw. SAFEnet issued a corrective statement and helped to reinstate the pilot's position.
- An online gender-based violence (GBV) case in 2019 involved a woman activist who is also an Ahok supporter. She received 300 to 5,000 requests per day for 'services' after her cell phone number and fake pictures of her in compromising poses were non-consensually shared on dating apps and WhatsApp groups catering to men looking for women companions.
- Another GBV case in 2019 was that of West Papuan lawyer Veronica Koman, who was alleged by Ministry of Communication and Information Technology to have spread misinformation and was accused of being an agent provocateur. The government actually shut down the internet in West Papua twice over the storm of opinions on this case. Civil society fought back and embarked on a solidarity campaign. *Tempo*, one of the cekfakta groups, showed the government was wrong. Despite the ministry's apology, however, attacks against her by state media are still going on today. One influencer doxed

her. A past picture of her with Hong Kong activist Joshua Wong was viralized to vilify her.

Participants were encouraged to use the 15-minute coffee break to think about threats—examples of which were pasted on a wall—and what they would prioritize, which will be explored in the next session.

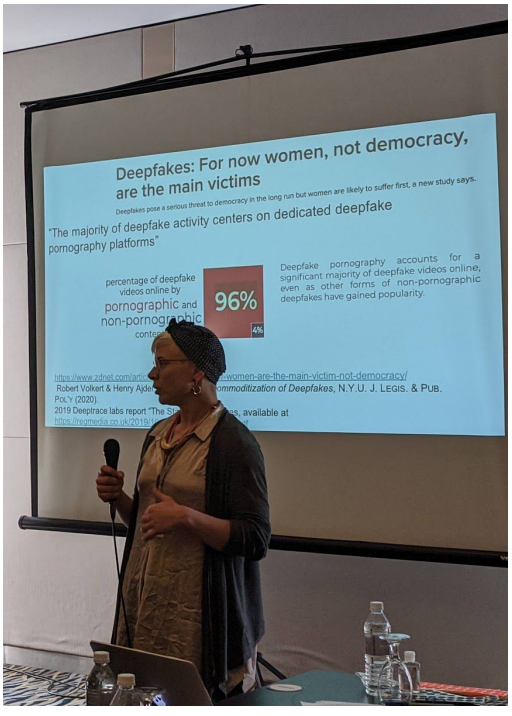
3.5. Session 4: Gender-based violence (GBV) & deepfakes - Dia Kayyali, WITNESS

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

Online gender-based violence (GBV) is not new. The usual manifestations are hateful speech, threats of physical violence, doxing, photoshopped images, and fake profiles of women and gender non-conforming (GNC) people which are used to blackmail, discredit, or embarrass them. However, these are now joined by deepfakes.

An example of a common online GBV, besides verbal abuse and threats, is the creation of fake social media accounts of the target victims, with damaging purported pictures of them, as had happened to Indian activist Preetha G Nair, who had criticised caste discrimination on her social media. Deepfake entered the picture in the case of Indian journalist Rana Ayyub, whose face was inserted in a porn video that was shared widely on WhatsApp, with her phone number as well, for speaking about gangrape in her community.

Women—not democracy—are most vulnerable to deepfakes thus far. A 2019 study by Deeptrace, a Netherlands-based initiative detecting deepfakes, estimates that 96% of deepfake videos online consisted of pornography targeted at women. Also, it is those who are already vulnerable who experience the negative side effects of new technologies, and those vulnerable in multiple ways are the first to experience them, like women



journalists or women human rights defenders of an ethnic/religious minority.

The availability of such harmful content was demonstrated in a DuckDuckGo search (rather than Google, which filters some of such content) in Malaysia on deepfakes, which resulted in links to pornographic and non-consensual sexual images of celebrities. A search in Brazil revealed similar results.

While the technology to make realistic deepfake videos remains exclusive, the barriers to wider usage are increasingly reduced. Meanwhile, there are already plenty of technological solutions to create less realistic deepfakes that still manage to harm women, especially if they are celebrities. FakeApp is one such example; its creator intends to 'democratise' its technology, which centres on creating pornography with celebrity faces, and there are message board how-to lessons as well. GitHub also contains message boards even contains payment offers for deepfakes of certain celebrities.

Navigating takedowns is not easy as the steps given are vague. The biggest porn site, Pornhub, does not have an ongoing policy statement on "deepfake"; it does not even mention the term. It promised to ban non-consensual porn, but an infamous deepfake of US actor Daisy Ridley is still available on the site, easily found via DuckDuckgo.

The easy availability of the technology and such sites encourages the commercialization of the production of GBV-targeted deepfakes.

In conclusion, participants were asked to think about the manifestations of existing online GBV, what is driving it, who are making money off of it, and what policies and tools can help the most vulnerable.

A distinction was made between the terms "revenge porn", which implies initial consent that was later taken out of context, and the politically correct term, "non-consensual visual images".

Interlude: Group exercises

Exercise A: Prioritization of threat models & vulnerabilities

Objective: To get a sense of the individual concerns in the room with regard to the threats identified from previous workshops

Slides available here.

Participants were each given 12 red dot stickers to be used to recognize and prioritize threats. This was done by providing examples of threats identified from previous workshops, arranged in broad categories (with a recognition of overlap) and put up on a wall in the room. Broad categories included:

- Individualised persuasion for targeting for access
- Enhancing individual harms
- Social movement attacks
- Expands existing attacks on civic activists, politicians, journalists, public figures, & increases access to some attacks
- Targeting of & usage by public figures
- Utilized versus institutional actors, processes
- In relationship to existing mechanisms for misinformation in elections and other processes
- Alters range, speech or potential manipulation of audio/video raw material in journalism, public sphere
- Targeting news processes
- Pushing towards 'zero trust' basis

Participants could place twelve dots. No limits were placed on the number for each threat.

The most notable results are highlighted in the chart below in colour, based on participant choices – with red, yellow, and green providing easy insights into what was prioritized most and highly. The aggregate prioritization per category is also visible.

INDIVIDUALISED PERSUASION FOR TARGETING FOR ACCESS (25)	ENHANCING INDIVIDUAL HARMS (35)	SOCIAL MOVEMENT ATTACKS (27)	EXPANDS EXISTING ATTACKS ON CIVIC ACTIVISTS, POLITICIANS, JOURNALISTS, PUBLIC FIGURES, & INCREASES ACCESS TO SOME ATTACKS (29)	TARGETING OF & USAGE BY PUBLIC FIGURES (43)	ALTERS RANGE, SPEECH OR POTENTIAL MANIPULATION OF AUDIO/VIDEO RAW MATERIAL IN JOURNALISM, PUBLIC SPHERE (28)	VERSUS INSTITUTIONAL ACTORS, PROCESSES (24)	TARGETING NEWS PROCESSES (44)	RELATIONSHIP TO EXISTING MECHANISM FOR MISINFORMATION IN ELECTIONS (53)	PUSHING TOWARDS 'ZERO TRUST' BASIS (58)
Targeted use to deceive critical gatekeepers or information holders e.g. intelligence or national security/critical infrastructure (15)	Cyberbullying, non-consensual sexual imagery without source material (21)	Attacks on social movement narratives & credibility (20)	Gender-based attacks on credibility of human rights activists and journalists (25)	Credible doppelgangers of real people that enhance the ability to manipulate public or individuals to commit rights abuses or to incite violence or conflict (17)	Integration of faked audio/video into ongoing public health or conspiracy campaigns, e.g. anti-vaxx (14)	Under-resourced courts and legal processes reject video and image evidence (19)	'Poisoning the well' in a leak with a few well-faked videos (9)	Fake public safety alerts shared on social media with credible audio and video (16)	Floods of falsehood as part of computational propaganda, individualized micro-targeting contribute to disrupting remaining public sphere (17)
Targeted private blackmailing with threat on exposure (9)	Non-consensual sexual images or so-called revenge porn (14)	Increasingly sophisticated spoofing of identities—of political opponents, journalists, public figures, persons working with critical infrastructure,	Expand, alter reputation, & safety attacks on key participants in elections & on politicians (3)	Automatically enhance social division with synthetic video/audio of non-famous individuals visibly affiliated with groups—e.g. police officers, soldiers (10)	Deepfake confessions (4)	Expands, introduces new threats of subtle edits to critical videos, compromising their value as evidence, info (5)	Altered documentation of war crimes violations compromises credibility of investigators, journalists (9)	Introduces manipulated info in a conflict/pre-conflict situation (13)	Plausible deniability for the powerful on any image (8)

INDIVIDUALISED PERSUASION FOR TARGETING FOR ACCESS (25)	ENHANCING INDIVIDUAL HARMS (35)	SOCIAL MOVEMENT ATTACKS (27)	EXPANDS EXISTING ATTACKS ON CIVIC ACTIVISTS, POLITICIANS, JOURNALISTS, PUBLIC FIGURES, & INCREASES ACCESS TO SOME ATTACKS (29)	TARGETING OF & USAGE BY PUBLIC FIGURES (43)	ALTERS RANGE, SPEECH OR POTENTIAL MANIPULATION OF AUDIO/VIDEO RAW MATERIAL IN JOURNALISM, PUBLIC SPHERE (28)	VERSUS INSTITUTIONAL ACTORS, PROCESSES (24)	TARGETING NEWS PROCESSES (44)	RELATIONSHIP TO EXISTING MECHANISM FOR MISINFORMATION IN ELECTIONS (53)	PUSHING TOWARDS 'ZERO TRUST' BASIS (58)
		human rights defenders (7)							
Prank-trolling where an audio or video deepfaked celebrity encourages people do a dangerous act live (1)			Extortion & cybercrimes (1)	Doppelganger campaigning (7)			Expand range of media used in breaking news exploitation: critical pre-election, media lockdown fakes (8)	'Wildfire incitement' misrepresents marginalized/ discriminated-against groups to incite violence, e.g. claims of refugee/migrant violence (11)	High-profile false positive or negative supports claims of 'deepfakes' (7)
				Widely circulated manipulated media of a political figure doing something inflammatory or unexpected (6)			Expands range of brand hijacking using existing logos, sets, brands in news industry (7)	Reinforces existing problems of 'digital wildfire' shared locally, primarily in closed messaging apps, often to incite violence or promote distrust (8)	Widespread use to destroy trust in institutions in authoritarian societies (6)
				Faked physical impairment of public figures, similar to			Swamping newsroom operations with	Expands easy integration of faked audio/video into	Simulated, individualized hate speech audio (5)

INDIVIDUALISED PERSUASION FOR TARGETING FOR ACCESS (25)	ENHANCING INDIVIDUAL HARMS (35)	SOCIAL MOVEMENT ATTACKS (27)	EXPANDS EXISTING ATTACKS ON CIVIC ACTIVISTS, POLITICIANS, JOURNALISTS, PUBLIC FIGURES, & INCREASES ACCESS TO SOME ATTACKS (29)	TARGETING OF & USAGE BY PUBLIC FIGURES (43)	ALTERS RANGE, SPEECH OR POTENTIAL MANIPULATION OF AUDIO/VIDEO RAW MATERIAL IN JOURNALISM, PUBLIC SPHERE (28)	VERSUS INSTITUTIONAL ACTORS, PROCESSES (24)	TARGETING NEWS PROCESSES (44)	RELATIONSHIP TO EXISTING MECHANISM FOR MISINFORMATION IN ELECTIONS (53)	PUSHING TOWARDS 'ZERO TRUST' BASIS (58)
				Pelosi, which used shallowfake techniques (1)			unverifiable media (3)	ongoing state, parastatal disinformation campaigns (3)	
				Manipulated public speech of politicians (1)			'Poisoning the well' of a probe (2)	Induced public panic via fake emergency (2)	Plausible deniability for perpetrators to reflexively claim "deepfake" around incriminating footage or, taken further, to dismiss any contested info as another 'fake news' (4)
				Enables use of 'credible doppelgangers': nationally known public figures manipulated,			Introduces news system weakness, e.g. tamper with single-camera remote broadcasts (2)		Personalized micro-targeted political messages going haywire (3)

INDIVIDUALISED PERSUASION FOR TARGETING FOR ACCESS (25)	ENHANCING INDIVIDUAL HARMS (35)	SOCIAL MOVEMENT ATTACKS (27)	EXPANDS EXISTING ATTACKS ON CIVIC ACTIVISTS, POLITICIANS, JOURNALISTS, PUBLIC FIGURES, & INCREASES ACCESS TO SOME ATTACKS (29)	TARGETING OF & USAGE BY PUBLIC FIGURES (43)	ALTERS RANGE, SPEECH OR POTENTIAL MANIPULATION OF AUDIO/VIDEO RAW MATERIAL IN JOURNALISM, PUBLIC SPHERE (28)	VERSUS INSTITUTIONAL ACTORS, PROCESSES (24)	TARGETING NEWS PROCESSES (44)	RELATIONSHIP TO EXISTING MECHANISM FOR MISINFORMATION IN ELECTIONS (53)	PUSHING TOWARDS 'ZERO TRUST' BASIS (58)
				leveraged for national impact (1)					
							Archival content from 20 years ago includes, implicates contemporary figures (2)		Reinforces existing & introduces new forms of astroturfing with convincing content to create fake unified or divisive public opinion (3)
							'Exclusive' leaks to journalists within rapid news cycle (1)		Exploit liar's dividend to claim compromising content is deepfake/faked (3)
							Manipulated 'sting' videos in a short news cycle (1)		Search engine degradation with volume of similar but incorrect images, video (1)
									Targeting of dissidents in

INDIVIDUALISED PERSUASION FOR TARGETING FOR ACCESS (25)	ENHANCING INDIVIDUAL HARMS (35)	SOCIAL MOVEMENT ATTACKS (27)	EXPANDS EXISTING ATTACKS ON CIVIC ACTIVISTS, POLITICIANS, JOURNALISTS, PUBLIC FIGURES, & INCREASES ACCESS TO SOME ATTACKS (29)	TARGETING OF & USAGE BY PUBLIC FIGURES (43)	ALTERS RANGE, SPEECH OR POTENTIAL MANIPULATION OF AUDIO/VIDEO RAW MATERIAL IN JOURNALISM, PUBLIC SPHERE (28)	VERSUS INSTITUTIONAL ACTORS, PROCESSES (24)	TARGETING NEWS PROCESSES (44)	RELATIONSHIP TO EXISTING MECHANISM FOR MISINFORMATION IN ELECTIONS (53)	PUSHING TOWARDS 'ZERO TRUST' BASIS (58)
									authoritarian contexts with ubiquitous, unconstrained surveillance (1)
									Routine use by political elites to claim or counter-claim (1)

According to the results, most participants thought a 'zero trust' world stemming from the proliferation of deepfakes to be a priority threat, followed by deepfakes' reinforcing misinformation mechanisms in elections, the targeting of news processes, and the targeting of and (mis)usage by public figures.

More specifically, gender-based attacks were their greatest concern, followed by cyberbullying and non-consensual sexual imagery without source material, and attacks on social movement narratives and credibility. They were also concerned about under-resourced courts' rejecting video and image evidence.

Having experienced violence sparked by false information on social media, they also fear the spectre of deepfaked doppelgangers' manipulating emotions towards inciting rights abuses or conflict, as well as floods of falsehood and individualized micro-targeting. The growing threat of COVID-19 at the time of the workshop (early March 2020) also likely drove concerns on falsely authoritative social media posts and health-based misinformation and disinformation.

Key threat concerns highlighted included:

- Gender-based attacks on credibility of human rights activists and journalists (25)
- Cyberbullying, non-consensual sexual imagery without source material (21); and non-consensual sexual images or so-called revenge porn (14)
- Attacks on social movement narratives & credibility (20)
- Under-resourced courts and legal processes reject video and image evidence (19)
- Floods of falsehood as part of computational propaganda, individualized micro-targeting contribute to disrupting remaining public sphere (17)
- Credible doppelgangers of real people that enhance the ability to manipulate public or individuals to commit rights abuses or to incite violence or conflict (17)
- Fake public safety alerts shared on social media with credible audio and video (16)
- Targeted use to deceive critical gatekeepers or information holders e.g. intelligence or national security/critical infrastructure (15)
- Integration of faked audio/video into ongoing public health or conspiracy campaigns, e.g. anti-vaxx (14)
- Introduces manipulated info in a conflict/pre-conflict situation (13)
- 'Wildfire incitement' misrepresents marginalized/ discriminated-against groups to incite violence, e.g. claims of refugee/migrant violence (11)
- Automatically enhance social division with synthetic video/audio of non-famous individuals visibly affiliated with groups—e.g. police officers, soldiers (10)
- Altered documentation of war crimes violations compromises credibility of investigators, journalists (9)
- 'Poisoning the well' in a leak with a few well-faked videos (9)

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

Objective: To get a sense of the different stakeholders' concerns from a Southeast Asian perspective with regard to the threats identified from previous workshops, the results of which will complement the existing threats map from previous workshops.

Participants had a discussion on the threats posed by deepfakes according to stakeholder groups/areas of interest. The groups were: **disinformation/fact-checking**; **media**; and **human rights and social movements**. Participants opted into their preferred group and each group was facilitated by a WITNESS staff member and had a documenter. The guiding points for the discussion were:

- The impact of new forms of manipulation in terms of expanding or altering existing challenges, introducing new challenges, and reinforcing other threats
- Priority threats
- Missing threats

Participants in the **media** group discussed these challenges and current issues and also moved into initial discussion on proposed solutions:

- The '**zero trust'** basis will spread with the prevalence of fake unified or divisive public opinion from astroturfing. This could lead to video/audio losing credibility as evidence, and the burden of proof falling on the victim/media.
- Fact-checking should not be the sole responsibility of one media or party, as the bias, perceived or not, would not inspire public confidence.
- In Myanmar, the huge influence wielded by figures of authority such as Aung San Suu Kyi has resulted in a blind acceptance of what they say, even in the face of credibility-challenging evidence; this gives her possible deniability and justification for problematic policies. This year's elections, or possible use in the context of either Aung San Suu Kyi or monks are significantly concerning.
- In West Papua, the state itself is disseminating inaccurate news, criminalizing journalists and activists and is not held accountable with nit does this (unlike how journalists are attacked and criticized when they share mis/disinfo); similarly in southern Thailand, the military is suspected of infiltrating social media anonymously to promote disinformation, and in Sri Lanka, most false news attacks target Muslims. Southern Thai participants noted the concerns about how confessions might be forged or faked.
- In Malaysia, the mass media often use user-generated videos and audios as source materials, hence the ability to ascertain their truthfulness is much needed.

As newsrooms lack resources to check deepfakes and are unprepared, it was recommended that:

- More collaboration with other stakeholders, especially tech companies and mobile telephone companies, be initiated;
- More collaboration between media on media forensics, and for technically challenging content such as deepfakes
- There are "simply not enough resources": need more capacity and funding to combat these threats
- Clear informational guidelines be updated; and
- Access to detection and verification tools be improved and widened.

It was also asserted that fact-checking is not a priority to media owners, who see little incentive or financial rewards in expanding that endeavour and are even more likely to feel this around more complex fakes that require more resourcing to spot, hence such collaborations can

help the media do more with less. The long-term solution is to raise the people's awareness through digital/information/media literacy.

Participants in the **disinformation/fact-checking** group raised the following concerns:

- 'Truthpocalypse' arising from exploitation of liar's dividend and the long-term effect of weaponization of information, which also leads to the loss of moderate voices.
- That deepfakes are being created to worsen the situation for vulnerable groups, including via gender-based violence.
- The new threat of deepfake of a 'source', in which real-time face swaps make people think they are chatting online with someone real but who actually does not exist, opening them to the threat of private blackmailing.
- The possible use of deepfakes for financial scams and health misinformation.
- Echo chambers and confirmation bias that deepen social division—it is more difficult to change minds when deepfakes reinforce existing beliefs. For example, the shallowfake of US House Speaker Nancy Pelosi who was made to appear drunk, is still believed by many to be real despite proof of alteration to the video.
- The vast majority of misinformation is visual – for example in India – and there are not enough good tools to catch this type of manipulation rapidly
- Media, civil society investigators and fact-checkers are not prepared to face potential threats—as detection technology is still exclusive to academics and companies, nor do many practitioners have a sophisticated understanding of these forms of manipulation. The suggested solution was to have a database of experts that can be a source of reference for fact-checkers globally. A possible problem with making detection tools accessible is reverse engineering or other misuses; hence there needs to be a 'Goldilocks' solution that provides a range of levels of functionalities and access.
- The group also looked one level further at the question of data collected by deepfake apps and the transparency and accountability of these apps, such as the Chinese face swap Zao, with regard to the data obtained.

While the debate continues on whether governments can treat deepfake technology as a weapon and restrict access, **media literacy** is the key underlying factor, which points to the role of journalists in preparing for the threats posed by deepfakes.

Like the media group, the **human rights** group were also concerned about the push towards 'zero trust' and discussed the following challenges:

- Technology enabling brand hijacking, and where media companies are concerned, it provides another easy weapon to attack journalism and journalists.
- The reinforcement of existing problems of digital wildfire, as deepfakes can be used to stage events on videos to incite violence.
- Intersection with issues like internet shutdowns, where government is able to control the narrative and continue to share information (or mis/disinfo) while civil society cannot respond
- The poor policy and legal framework due to a lack of credible legislators, as most have a conflict of interest in the matter; specifically, that they often pursue controversial counter-measures to 'fake news' that disproportionately impact human rights, and miss the root problem.
- Inadequate fact-checking due to a lack of publicly accessible mainstream tools or competencies. A suggestion for tools to be centralised raised more questions of who controls or manages them and how. The problem is also compounded by the lack of access to detection technology, which is not even available commercially.
- The need to ramp up media and digital literacy even among activists due to the lack of awareness about the dangerous difference between deepfakes and the usual manipulation of audio-visual content.
- A growing apathy about truth, hence the need for awareness-raising campaigns highlighting its importance while recognizing that literacy is challenging – it's hard to implement and there is an apparent erosion of caring about truth
- Given the easy availability of training data for deepfakes due to people's lackadaisical attitude regarding the terms of use of internet platforms, it was suggested that gathering images as training data for the purpose of training AI models be banned. However, technology companies have AI models that are not deepfake-related. Also, preventing

deepfakes when the technology is already out there would be akin to closing the stable door after the horse has bolted.

In summing up, Sam noted that the discussion had also flowed into solutions. The various stakeholders have a role and need to be encouraged to fulfil them: the state can lay the foundation and framework, though its powers and intentions need to be checked and the misuse of fake news laws to suppress dissent must be foregrounded; technology companies must respond with policies on content, and provide the right tools and digital literacy support; and the media and civil society need capacity building and tools to verify such content, paired with literacies for the broader public.

Sam pointed out the results of the earlier exercise in which participants had indicated their priority concerns as regards the threats that had been identified in previous workshops in the other regions. The Southeast Asian participants are greatly concerned about:

- Attacks on vulnerable individuals;
- The detrimental impact on access to justice;
- The ability to verify audio-visual content;
- The systemic impact, as explained by Hannah Arendt, in that the danger with the proliferation of falsehoods is not that they are believed, but rather that cynicism pervades, which then emboldens the incumbent and status quo; and
- Plausible deniability or liar's dividend, wherein people can plausibly dismiss compromising material as false

Sam shared the commonalities among the discussions across the regions, which identified that deepfakes and other new forms of media manipulation:

- Increase the risk to the most threatened activists and community leaders including from the government and state actors;
- Undermine the credibility and use of video as evidence;
- Overload the capacity of journalists in reporting the truth and discerning falsehood;
- Have a negative impact on public trust, as it becomes easy to attack anything as deepfake; and
- Intersect with existing patterns of rapid digital wildfire of misinformation, hate speech and communal violence.

4. Afternoon Session

4.1. Session 5: Technical perspectives on deepfake detection—Francesco Marra, PhD, University Federico II

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 speaker had to conduct this session via Skype due to the new travel restrictions in Italy over concerns regarding the novel coronavirus.

Participants were first tested on their ability to detect videos that were not deepfaked but had been modified by superimposing the facial expressions of an actor. No one guessed correctly that none of the four muted clips of former US President Barack Obama speaking were real. The average accuracy of the human ability to detect fake content is less than 80%, while the quality of such content increases by the day.

The speaker touched on the early history of manipulation of visual content since the beginning of photography itself, comparing the original and doctored photographs of US President Abraham Lincoln and Italian dictator Benito Mussolini. Such alterations were intended to bolster the image of the subjects. When the media does it, as shown by the examples of the cover pictures of *The National Review* and *Al Ahram,* it is intended to spread disinformation.

The presenter showed how the technology has advanced from manipulation of photographs—by splicing, copy-pasting and in-painting—to AI-based image manipulation that can be done by a non-expert user, compared with how video editing tools used to be 'reserved' for expert users. Further, these deepfake videos can be created within a few hours.

How do we protect ourselves from deepfakes?

There are multimedia forensics detection tools that focus on either forgery detection or localisation. They cover the following areas:

- Physical integrity—which is revealed by shadowing or illumination inconsistencies.

- Visual integrity—e.g. different eye colours, very smooth areas, artefacts on edges or in spliced face boundaries (the face is extremely difficult to be reproduced, and is one of the first clues).
- Semantic integrity—which is revealed by identifying mis-correlation in time or location; or the image reconstruction using content from the web.
- Digital integrity—which is revealed by the footprint of subsequent processing in image acquisition and generation, i.e. a sort of digital fingerprint that is unique for each image. Participants were shown a clip from the 2009 movie *Beyond a Reasonable Doubt* to illustrate this method somewhat. One digital integrity detection technique is to use deep learning (similar to what is used to generate deepfakes), which relies on a database and collection/compilation of deepfake videos with which to train the tool to detect a similar deepfake video. However, this generally performs better if images are high-resolution (which is not always the case online).

Underscoring the urgency to stay ahead of deepfake technology, the US Defence Advanced Research Projects Agency funded the MediFor (an abbreviation for media forensics) program (2016-2020).

Major challenges

Physical- and visual-based techniques rely on traces that will disappear with time. Semantics integrity is also problematic, as it depends on the existence of the original content as a basis for comparison.

While digital integrity is the most robust method, reliable extraction is a problem. Its limitations are:

- Compression of the image (by social networks);
- Generalisation, which refers to the incompatibility in identifying forgeries created by different deep-learning techniques using a technique derived from one technique; and
- Attacks on deep-learning detectors.

Moving forward

The recent projects include the recent Deepfakes Detection Challenge that provides consensually obtained data for researchers to experiment with, and Google's large dataset of visual deepfakes, numbering 3500 videos.

There is no universal deepfake detector for two good reasons: Having multiple detectors makes the process more robust and accurate; and it is more difficult to fool a variety of detectors.

The speaker ended his presentation by thanking his colleagues at GRIP, the image processing research group at his university, for their contribution to the research on which this presentation is based.

Sam thanked the speaker for the virtual presentation and told participants that the GRIP members are among the world's leading scientists and researchers on deepfakes.

Participants' questions and comments

The first question asked was whether we should panic. Although the speaker cautioned against overdoing it, he added that a little *is* needed because technology is constantly competing to create better deepfakes that cannot be spotted with the naked eye, so the work on providing tools for detection has to keep up.

As for the possibility of vested stakeholders such as the workshop participants' supporting GRIP by providing video materials as locally based training data, the speaker said they are focused on extending the DARPA project for another four years and do not have many opportunities to collaborate with journalists and commercial actors. He recognized the demand for deepfake detection. Sam added that scientific researchers have shown a lot of interest in factoring in real-world concerns in their work, and WITNESS has organized meetings for deepfake researchers to engage with journalists.

A participant asked about the infrastructural approach to detecting deepfakes. Sam clarified that many projects publish the codes behind deepfake detection tools, but these are not available to journalists as ready tools. There are some recently launched tools, such as Assembler, produced by Alphabet's subsidiary Jigsaw, that incorporate some deep learning-based techniques but do not work for actual deepfakes. There are no commercially available tools, but things might change in six months.

To a question about the possibility of a repository of deepfake expertise that can serve as a helpline for journalists, Sam replied that there is none at the moment. The few experts, i.e. the academics, would not have the time, and their job is research. He acknowledged this to be an oft-raised unmet need.

4.2. Session 6: Solutions & 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, as well as obtain from participants their proposed interventions and solutions. Slide deck available here.

The nine questions, or areas of response, that point to solutions are:

1. Can we **teach people to spot deepfakes**?
2. How do we **build on existing journalistic capacity** and coordination?
3. Are there **tools for detection**—and who has access?
4. Are there **tools for authentication and provenance**—and who is excluded?
5. Are there tools for **hiding our images** from being used as training data?
6. What do we want from **commercial companies producing synthesis tools**?
7. What should **platforms** do?
8. What do we want from **politicians and civil society groups using deepfakes**?
9. What should **lawmakers** do?

Teaching people to spot: As regards the first question about learning how to spot deepfakes, this example underscores the problem: A researcher once publicised his finding that deepfake faces don't blink, thus inadvertently issuing a challenge to be proved wrong—which came a few weeks later in a deepfake sent to him which did blink and whose training data had incorporated blinking. Hence, current weaknesses of deepfakes should not be trusted or promoted as detection tips or solutions – for fear people will just remember the current algorithmic 'Achilles Heel'.

Preferable is to support media literacy that is informed by technical signals (for example to indicate invisible-to-the-eye manipulation) so that people can make informed judgements. Some

guidelines on spotting fake content are the SHEEP framework (which stands for **S**ource, **H**istory, **E**vidence, **E**motion, **P**ictures) and SIFT (**S**top, **I**nvestigate the source, **F**ind better coverage, **T**race claims), though the former is a better option as it includes emotion and pictures, both of which are powerful pulls in social media.

Building on existing journalistic experience and capacity: As for building on existing journalistic capacity, there are currently not many resources for that. Illustrating this point, none of the participants admitted to understanding media forensics when asked. This has been observed in mixed meetings such as this, let alone journalist-exclusive trainings. There is a gap between the availability of forensics tools and the need for them to cope with the reality of fast-churning compressed visuals. Hence, better collaborations are needed to find cross-disciplinary solutions building on existing practice and actual needs. Sam shared a link to a resource on detection needs and tools for journalists— 'How do we work together to detect AI-manipulated media?' which is based on the outputs from previous meetings.

Tools for detection: Tools for detection were addressed in Session 5. The takeaway is that ingenuity plays a large part in figuring out human characteristics that are most difficult to deepfake, such as pulses, which can only be detected on visuals via infrared sensors, and distinctive mannerisms of a person. The problem with the latter, however, is that it would only work for famous personalities who already have loads of data in the public. An important question that participants had raised was—who has access to the tools? Those most vulnerable, in diverse platforms, using methods relevant to real-world harm scenarios, and frontline defenders (journalists, fact-checkers etc.) must be able to use the tools, which must be as explainable as possible.

Tools for authentication and provenance: A later session focused on this topic.

Hiding our images from being used as training data: There is ongoing research on how an adversarial perturbation, or an invisible change to the image, can confuse an AI algorithm into misidentifying objects, thus preventing detection of that particular object. However, a participant pointed out that this is a double-edged sword as it can hamper open source investigative reporting against criminals. It reduces the possibility of people's images from being deepfaked but does not eliminate the threat. It is also difficult to implement on an ongoing basis and at scale.

From **commercial companies producing synthesis tools**, it would help if creators incorporate a misuse-prevention feature in their tools such as hidden watermarks and ensure they are detection-friendly.

The question of **what platforms should do** will be discussed further in Session 7. Basically, it concerns their present and in-the-works policy regarding deepfakes and detection resources, and their transparency about the actions they are taking, as well as the tools and services they provide to consumers, fact-checkers and investigators, civil society and government.

As regards our **expectations of politicians and civil society groups using deepfakes**, which is an ethical issue, a Pledge for Election Integrity circulated in Europe and the US that includes not using such manipulated content but take-up by the Democrat Presidential candidates has been pathetic, though this lack of political will was by no means limited to the US. The earlier cited case of the Indian politician's deepfaked video is another example of usage that may confuse people.

So, **what should lawmakers do**? There is a significant amount of laws being made on this, often election-driven and related to non-consensual sexual or intimate images, but also in the bigger context of fake news and governmental repression where the concept of 'fake news' or deepfakes is being used to justify attacks on free speech and media.

4.3. Session 7: Discussion & prioritization of solutions in a Southeast Asian context

The objective of this group exercise 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 present.

The emphasis in this session was to prioritize and identify input and prioritization from the region on what solutions they wanted to see implemented, resourced or explored. Participants were given the option to choose the groups they would like to work on based on their interests. The groups were:

- **Media literacy**: Discussing the precursors necessary to build media literacy in order that social media users would be less susceptible to deepfakes as well as the continuum of shallowfakes.

- **Collaborations/coordination and journalistic skills and tool, detection and authentication solution**: Discussing the needs of media practitioners and journalists including tools, skills and practice together with detection/authenticity solutions.
- **Platforms**: What should they do?

The media literacy group was the most popular, followed by the collaborations/detection group. The platforms group had fewer participants joining in.

The discussions were guided by the following questions:

- What solutions feel most relevant to you?
- What would you need from those solutions to make them viable? What would be ways to approach this in South and Southeast Asia?
- What worries you about these solutions?
- What would be a concrete next step in this area that you would like to see?

Each group was also given more detailed questions to help them focus.

Participants in the **journalistic skills, detection and coordination** group talked about what worked in their experience. Consortium and collaboration models between media, civil society and broader communities can work. There are working models of informal and formal collaboration. For example, in Indonesia, 40 mostly online news organisations shared fact-checking work among themselves. However, they are limited by the lack of technical tools, especially for detecting deepfakes. Their success is all due to their networking that covers all regions of the country. In Taiwan, outreach is made to senior citizen and other communities to do offline detection of false news, which are sent for verification at the Taiwan Fact Check Center. The collaborations group also discussed what is expected of platforms, with providing tools so that you can do an easy 'lab blood test' on suspected fakes being one of the mooted ideas—though that also raises a new problem of preventing abuse, as opening up tool and data access can also provide opportunity for bad actors. The collaboration group suggested that the ability to combat deepfakes will depend on the capacity of the malicious creators as much as the capacity of civil society actors, and it might be better for civil society actors to 'pick its battles' and focus on easier-to-spot fakes, rather than state actor campaigns.

The **media literacy** group thought that educational entertainment ('edutainment'), using social media, would be a good format for public education. The approach should return to root of the

problem—confirmation bias. The preventive measure should also be as basic—immunisation. Digital literacy, including questioning the basic 5Ws and 1H (who, what, when, where, why and how), needs to be promoted (as has been modelled in Indonesia). The demographic target is young people and housewives, as a study shows they are the main culprits in sharing false messages. Young people, especially, need to be awakened in schools to such threats. Participants from Taiwan also suggested the value of engaging older people and using the right technologies for the right target constituency, for example Pokemon Go in this instance for older users.

For the short-to-medium term, more training for stakeholders of diverse backgrounds on media monitoring and harm reduction, including the provision of tools. CSOs can hold a joint effort with other stakeholders to do media monitoring—e.g. a "Face-to-Facebook" training on ethical and safe ways to use Facebook (as done in Sri Lanka). Technologists should do outreach and do more public education, e.g. by meeting up with media editors. It is crucial for the community to build trust and find sources from other stakeholders so that they could share knowledge and verify information, especially when critical issues emerge. Intra-agency meetings could be held in local and regional settings towards this end.

A representative from Cambodia shared her experience in giving media literacy training to rural communities. Adults are their target, focusing on false news. Given the lower education levels, and the multiple languages to overcome where the indigenous peoples are concerned, the

material has to be as simplified as possible, such as an image of someone's head on another's shoulders, the wrongness of which is immediately understood. Sam concurred that by just showing an obviously doctored image, anyone would understand the impact, and explanations about AI or all the new developments are not necessary for certain needs. This prompted a reflection from another trainer–participant who has done similar work with high school students. A preoccupation with terminology was superfluous and even counterproductive, as an everyday person would have a simpler frame of reference. A "one size fits all" approach is not effective when it comes to complicated terms. For example, it would not do to focus on a particular platform when the target audience use a different one and would have problems applying the principles accordingly.

This discussion was a helpful reminder of an immediate next step in media literacy—i.e. think of descriptions of terms in everyday language; it may not even be strategic to use the word "deepfake".

The **platforms** group thought content moderation was necessary, and all messaging apps should have a report button, as these tools such as WhatsApp were likely to be the most challenging. Their specific recommendations are for the major players to cooperate with and support the smaller platforms on this, combining this with third-party oversight for other platforms as well, not only Facebook.. They also suggested looking at mechanisms to stop the spread of non-consensual sexual and intimate image across multiple venues. However, participants could not come to a consensus on which action is appropriate within platforms to respond to deepfakes—take-down or labelling. A low-hanging fruit would be to focus more on existing miscontextualization 'shallowfake' problems.

To round off this session, **all participants** were given an idea of what their Brazilian and South African cohorts prioritized, which often echoed their concerns as well:

- Media literacy needs to be contextualized in the bigger misinformation and disinformation problem, especially for grassroots communities and the most vulnerable.
- Create detection tools for existing problems of manipulated video and audio 'shallowfakes'—and in closed networks like WhatsApp—otherwise, deepfake tools are a luxury.
- Create deepfake detection tools that are cheap, accessible and explainable for the layperson. However, access also involves compromise, hence a balance needs to be struck, which is another challenge.
- Investment is needed in the capacity of journalists to understand new forms of media forensics.
- Major platforms need to be part of the solution with transparency and support to separate truth from falsehood. This includes providing tools for detection and promoting media literacy with regard to the spread of false news.

4.4. Session 8: Emerging trends in authenticity infrastructure—provenance & image integrity

The objective of this session is to provide input from the region on emerging discussions in which WITNESS is participating that look to build better infrastructure for media authenticity and provenance, such as the Content Authenticity Initiative. Slide deck here.

When it comes to authentication tools, there are three types of technology: verified-at-capture (at source); verified-at-publishing (tracking edits or at distribution); and reverse video or similarity search (in platforms). WITNESS is venturing into this area due to a trend in mainstreaming these tools. Tech and media giants Adobe, Twitter and the New York Times have a joint project called the Content Authenticity Initiative to develop an industry standard for digital content attribution, and various stakeholders, such as WITNESS, are being engaged in the consultation process.

WITNESS has come up with a report detailing the 14 dilemmas arising from this issue, "Ticks or It Didn't Happen: Confronting Key Dilemmas in Building Authenticity Infrastructure for Multimedia' (see: report here).

Primarily, the mainstreaming of such tools is concerning because, while they are currently more like a niche opt-in, it seems very likely that they would later be driven by government or platforms. WITNESS has been studying related developing issues such as fake news laws and blogger registration, and are concerned about the implications of online users' being compelled to share as much as possible information about their content, i.e. the metadata etc., or their persistent or real identity for authentication purposes. This may inhibit some users whose work is sensitive, which raises the question of whose voices are excluded, be it accidentally or deliberately. Whistle-blowers and citizen journalists, for example documenting police violence in Brazil might be hampered, and they would be limited in their choice of safe platforms to use.

There are also arbitrary, tech-driven exclusions—certain devices like jailbroken phones cannot use authentication tools.

Another problem is the increased burden of proof that comes with such tools—where is it placed and who bears it? The answer is science, instead of humans, because of the 'CSI effect', which refers to expectations of forensic evidence as a fool-proof verifier of truth or falsehood, due to its dramatic portrayal in the American TV crime series the effect is named after. Yet, as this workshop has revealed, there is a severe shortage of people with good forensics skills to confirm the science.

Further, what if certain authentication symbols, such as ticks, do not work or work too well? The presence of ticks has been misinterpreted as trustworthy, giving a false sense of security and leading to user carelessness in doing their own verification.

The other dilemma is, who gets access to the underlying data, both illegally and legally? There is a not unfounded danger of abuse of technology to crush dissent.

Also, what further pressures will this put on media platforms that are caught in between expectations and solutions?

Lastly, it must be reminded to all that the science of media forensics is not fool-proof.

Participants were asked for feedback of their preferences on the following questions in connection with the setting up of large-scale authenticity infrastructure solutions:

1. How closely does identity need to be tracked?

a. A well-known identity for who created and changed media is an essential part of creating trust for consumers of video and images.
b. A well-known 'authorship' identity is a barrier to entry for people working in difficult environments (e.g. reporting on war, human rights violations, or criminal activity).

A participant thought this should not be an either-or solution but seen as a whole, as it depends on the stakeholders' needs, which can defer depending on who and what is at stake. A journalist would want to cite a publicly identified source to establish trust for the story but at the same time also protect their sources where warranted.

In rejecting B, a participant said that an extensive and detailed metadata generated from authorship identity requirements would threaten human rights work in difficult environments. Dia shared that it poses a real threat to human rights activists and their work to expose the conditions in war-torn Syria and other similarly dangerous places.

The problem with A is that the metadata will deteriorate when they go through intermediaries, hence platforms need to create a mechanism that allows for a level of accountability through retention of some of that data.

A poll taken after the five views were aired showed no support for A. **All agreed with B**, but one added that the situation could be improved with more protections like encryption.

2. How much information should be presented on edits to videos or images?

a. Every edit to a photo or video should be displayed. Without that, the presentation lacks credibility.
b. A short summary of the edits should be displayed. Showing every edit is exhaustive and most users would not bother to read through it.
c. Some edits should not be visible. For example, if people's faces are blurred for privacy.

A question was asked on what "every edit" entails—is it a series of files or a log of changes? If it is just the latter, a participant would agree to A.

Another participant chose B because in his experience in collecting videos as trial evidence, the court would need to know the edits that had been done, and a short summary would suffice. A detailed list in the war crimes archiving context, however, could put their sources at risk of multiple victimization should the information fall into the wrong hands.

A participant thought a combination of A and B is possible as a two-step process. However, she was undecided about the impact of exposing innocuous edits, which happens a lot.

Sam pointed out that what constitutes 'inoffensive edits' is not easily determined—a sunset that has been enhanced to look good for tourism purposes may seem acceptable in comparison with the malicious intent to change the context of an incident, until tourists discover the truth for themselves when they arrive to see that the sunset view is not quite like how it was promoted.

3. Do photos and videos have to be generated and tracked throughout their life to be trusted?

a. There is a "golden path" of "controlled capture" when a photo or video is taken on a smartphone or camera, controlled edit, and controlled distribution and sharing. Once a photo or video falls off that path, it is permanently untrusted. This implies all historical media can never participate in these authenticity infrastructure workflows.
b. There needs to be a path for on-boarding of media with unknown history. This implies that there is a mechanism that signals "*unknown things have happened.*"

A participant agreed with B because in his experience, source pictures are usually difficult to obtain. However, the caveat is that there must be additional strong verification.

Another questioned the implication that A would mean a breach of privacy. This is not a certainty, as there are photographic methods to anonymise pathways in theory, though they are not commercially viable yet. Another participant wondered whether a timeline of edits is important and suggested that rather than being a series of humdrum logs, the record should be made intelligible to people, perhaps with icons for different types of edits. It was further suggested by another participant that they should be calibrated to show a prioritization according to importance.

4. Online vs offline operation?

a. Any capture or edit of a video or image must be performed while online so that the relevant information can be captured in a secure, controlled environment. End-user devices can never be trusted.
b. The online requirement poses a barrier to entry for people working in remote or unstable regions. There needs to be a mechanism to work offline and (eventually) upload content in a way that can be accepted into an authenticity infrastructure.

Participants chose B, though one said the problem with it is that in conflict situations, there is a risk of the content being damaged or erased if it was not uploaded as soon as possible. He cited how Myanmar citizen journalists had to throw away about 3,000 smartphones into rivers during the riots for fear of being shot dead by the authorities if found in possession of them. This is related to access to verification tools, a lot of which are only online, so a mobile device with a strong connection is needed.

A question was asked whether A was a contradiction. Sam clarified that the device used would have to be controlled and trusted, and it would have to be done in an online environment. Queried further on the end-user part, he said there are efforts to shift the security burden to the hardware rather than the software and develop better ways to trust at the camera level. Another participant clarified that the cost-effective technology is already there but there is no agreement as yet among manufacturers regarding certification trustworthiness.

5. Who should have access to authenticity information?

a. Authenticity and provenance information should be publicly visible on all media.
b. Detailed information should be visible only to platforms, which can provide summary information.
c. Users should have control on who can access the information.

Participants chose C. Upon being challenged to argue for A or B in the interest of transparency or combating fake media, a participant said that A can expose bad actors who spread misinformation as a strategy. To a question on whether even trivial edits such as adding filters will be revealed to all and sundry, Sam replied probably not, but the scope of this type of work is increasing.

Wrapping up, Sam said WITNESS' position on the optimal principles for authenticity are based on classic human rights values:

- Such technologies are merely a signal of authenticity, not the sole signal; they are a signal towards trust, not a confirmation of trustworthiness.
- Such technologies are an opt-in for creators, not a legal obligation.
- There must be an open ecosystem of tools for independent verification, with user control on privacy at multiple levels.
- Such technologies must be grounded in the needs of people who may benefit most and be harmed most by these technologies.

4.5. Session 9: Feedback on platform policies

The objective of this session is to review the approaches being developed by major internet platforms for handling deepfakes and other manipulated media, and identify feedback and concerns. Slide deck here.

As internet platforms are developing policies regarding deepfakes and synthetic media, WITNESS is playing its part in ensuring that human rights concerns are included by providing its input to the stakeholder consultation processes, a continuing effort that all relevant actors should be involved in. Sam voiced appreciation for the interest shown by Facebook, Google and Twitter for this meeting, which had representation from Facebook (Google and Twitter staff faced travel bans due to COVID-19 concerns and had to withdraw at the last minute).

An extract of **Facebook**'s policy was shown to participants, whereupon it was pointed out that while deepfakes and maliciously manipulated content will be removed, body and scene manipulations are not mentioned. The focus is narrowed on deepfakes, not shallowfakes. The community standards regarding nudity, graphic violence, voter suppression and hate speech also apply, as well as the vetting by the non-partisan International Fact-Checking Network.

In comparison, **Twitter**'s policy is broader. It includes material that is significantly and deceptively altered or fabricated, such as the manipulated videos of Philippine senator Leila de Lima's confession of being a drug lord coddler and US House Speaker Nancy Pelosi's seemingly drunk appearance (unlike Facebook, which would not deem both merited removal). Twitter's

policy also covers material that is shared in a deceptive way and likely to impact public safety or cause serious harm.

Youtube's policy has been less touted. It covers misleading metadata or thumbnails, manipulated media that misleads users and may pose harm.

The platforms offer four notable options to such content: take-down; labelling; adding context; slowed down in its spread. However, Dia noted that there is no structural engagement with global civil society in developing policy on content moderation. There are pertinent issues that need answers: redress for take-downs and other transparency-related matters, such as whether users have a right to know what happened. Many human rights contents, including documentation and first-person stories, have been taken down in rather puzzling circumstances—after all, who could be negatively affected by them? What are the possible dangers?

Participants' comments

A question was asked about Twitter's possible exemption for public figures, as US President Donald J. Trump and his Republican cohorts had tweeted the Pelosi video. Sam said he is not sure of Twitter's policy, but this is a challenge for all platforms and it needs to be flagged as a loophole to be tracked. Philippine President Rodrigo Duterte had also done a similar thing.

A participant from India thought there are many loopholes in Youtube's policy. Videos of false monologues with no images are 'allowed', since they do not fit any of the stated categories. She can also see a backlash on content moderation in India, where the right-wing faction is already taking issue with the fact checkers who are working with Facebook and think any take-down of their contents has a left-wing bias. Who gets to decide the benchmark that platforms must adhere to? How do we address attempts to weaponize such markers?

A suggestion was made to learn from how platforms solved the problem of spam.

Another participant stressed on individual protection and privacy in videos, as the example of Myanmar has shown that police have tracked dissidents in videos.

A participant noted that different and rather complicated responses are required for the various platforms with different features and functions. For example, WhatsApp is a closed messaging platform, lauded for the privacy protection in its end-to-end encryption that at the same time

limits proactive or reactive content information, provision and moderation. He cautioned that any kind of moderation will result in a surveillance environment.

4.6. Session 10: Identification of relevant next steps

To wrap up the meeting, participants were asked what they thought were the low-hanging fruit and important focus that needed doing now that can push companies, governments, CSOs to move forward. These were their suggestions:

- Simplifying the vocabulary for public education purposes, after which an awareness campaign that includes simple, brief, multi-lingual videos can be held.
- Providing accessibility to detection systems.
- Build capacity for shared media forensics.
- A database of experts who can help journalists identify synthetic media.
- Focus on the right tools for existing miscontextualized and shallowfake media.
- Identify mechanisms for managing spread of non-consensual sexual images.
- Updates on what is being done to counter deepfakes, and sharing of best practices around the world, which will require reporting and translation work.

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

- A questionnaire on coordination and feedback regarding the immediate next steps would be shared with participants.
- Training materials from this workshop will be shared with everyone and look at how they can be developed further, especially on media forensics.
- WITNESS to facilitate a network to share information, strategies and solutions.
- WITNESS to input feedback on threats and solutions into global discussions and prioritizations
- Contact details of participants at this meeting will be shared among themselves unless otherwise stated.

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

witness.org

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