

WITNESS

SEE IT **FILM IT**
CHANGE IT

APAC Workshop

**Fortifying the Truth in the Age of
Synthetic Media and Generative AI**

Bangkok, Thailand — July 31-August 1, 2024

Report written by Vasin Boonpattanaporn. Editing by Jacobo Castellanos.



Between 31 July and 1 August 2024, WITNESS convened journalists, human rights defenders, technologists, and civil society activists from various parts of Asia-Pacific in Bangkok, Thailand, to identify, discuss and prioritize threats, solutions and opportunities from synthetic media.

This workshop was part of WITNESS's broader efforts, continuing from past workshops in South East Asia, as well as previous sessions held in Kenya, South Africa, Brazil, Colombia and other global consultations. The aim was to deepen understanding, foster regional connections, and align both global and local perspectives in tackling the challenges posed by these evolving technologies.

WITNESS's 'Prepare, Don't Panic' initiative remains central to these efforts, with a focus on intervening in the early stages of the synthetic media ecosystem. The initiative addresses the need to develop tools, policies, and legislative frameworks that not only identify and counteract the threats posed by synthetic media but also ensure the representation of marginalized communities who are most at risk. These communities, often excluded from the spaces that shape emerging technologies, face heightened dangers from the misuse of AI and synthetic media.

This report summarizes the key discussions, challenges, and proposed solutions from the Bangkok workshop, laying a foundation for continued efforts to safeguard human rights in the age of synthetic media and generative AI.

CONTENTS

Summary and Analysis	4
Identified Threats from Synthetic Media	4
What Has Changed in the Threat Landscape	5
Addressing the Bigger Issues	6
Specific Actions That We Need to Take	7
REPORT OF DAY 1	9
<u>Session 1</u> : Welcome & Introduction	9
<u>Session 2</u> : Introduction to WITNESS	9
<u>Session 3</u> : Fortifying Truth in the Age of Synthetic Media	10
<u>Session 4</u> : Generative AI and Synthetic Media in Asia-Pacific	13
<u>Session 5</u> : Experimenting with Generative AI	15
<u>Session 5, Part 2</u> : Hands-on workshop	17
<u>Session 5, Part 3</u> : Session Recap	19
<u>Session 6</u> : Key Themes in Human Rights & Journalism in APAC	26
<u>Session 7</u> : Risk ‘Spectrogram’	29
<u>Session 8</u> : Impact of GenAI and Synthetic Media on Human Rights and Journalism	33
REPORT OF DAY 2	35
<u>Session 1</u> : Kick-off and Announcements	35
<u>Session 2</u> : Recap of Day 1	35
<u>Session 3</u> : AI Regulation and Policy in Asia-Pacific	37
<u>Session 4</u> : Protecting What is True, Detecting What is Fake	39
<u>Session 5</u> : Synthetic Media Detection & Equity	40
<u>Session 6</u> : Range of Solutions and Personas Workshop	42
<u>Session 7</u> : CSO Agenda for AI Resilience and Future of Synthetic Media in APAC	44
<u>Session 8</u> : Wrap up: Collaboration, Next Steps and Thank You	47

Summary and Analysis

Identified Threats from Synthetic Media

1. Misinformation and Disinformation:

- **Spreading False Narratives:** Synthetic media, such as deepfakes and AI-generated content, are increasingly used to disseminate false narratives, particularly in political and social contexts. This misuse can manipulate public opinion, create chaos, and undermine trust in institutions.
 - **Example:** During the Baloch national gathering, state agencies shut down the internet and circulated AI-generated videos to deny the occurrence of protests, complicating civil society's efforts to counter the false narrative.
 - **Challenge:** The rapid spread of AI-generated disinformation in regions with low digital literacy demonstrates the insufficiency of current media literacy efforts. The lack of accessible and effective detection tools for the public exacerbates this issue.
 - 'In the recent incident during the Baloch national gathering, the state agencies ensured the internet was shut down and then showcased AI-doctored videos on social media, denying that any protests were happening.'

2. Undermining Trust and Truth:

- **Erosion of Trust:** The prevalence of AI-generated content makes it difficult to discern real from fake, leading to an erosion of trust in media and digital content. This contributes to broader cynicism and skepticism, even towards genuine information.
 - **Solution efforts:** WITNESS is helping develop provenance and authenticity tools to trace content origins. However, these tools remain inaccessible to many in the Global South and raise privacy concerns.
 - **Challenge:** Despite advancements, a significant gap in accessibility remains, particularly in resource-limited regions, further deepening mistrust in digital media.
 - 'The erosion of trust in our information ecosystem is being aggravated by everything else we're discussing, whether it's deepfakes or AI generation.'

3. Targeting Vulnerable Groups:

- **Exacerbating Inequalities:** AI-generated content can be weaponized against vulnerable communities, such as women, religious minorities, and political dissidents. This content often takes the form of targeted misinformation or gender-based violence, intending to silence or discredit these groups.

- **Example:** Vulnerable communities face disproportionately heightened risks as they lack the digital literacy to distinguish between deepfakes and real videos, and have limited means to have harmful content removed from platforms.

4. Legal and Ethical Concerns:

- **Manipulation of Legal Systems:** Synthetic media poses risks to legal systems by enabling the creation of fabricated evidence, potentially leading to wrongful accusations or imprisonment, especially in regions with limited resources to verify content authenticity.
 - **Example:** AI-generated content has been used to fabricate evidence of attacks, with state backing for these narratives, further complicating civil society's response.
 - **Challenge:** The commercialisation of AI tools without corresponding legal frameworks and safeguards poses a significant risk to the integrity of legal systems globally.
 - 'In some cases, AI-generated content has been used to fabricate evidence of attacks, with state backing for these narratives further complicating civil society's response.'

What Has Changed in the Threat Landscape

1. Rapid Advancements in AI Technology:

- **Improved Realism of Synthetic Media:** Generative AI capabilities have evolved rapidly, making synthetic media more realistic and harder to detect, increasing its potential for misuse.
 - **Challenge:** Detection tools for synthetic media have not kept pace with the advancements in AI, particularly in their availability and effectiveness for civil society and grassroots activists.

2. Widening AI Gap:

- **Disparity in Access and Knowledge:** There is a growing divide in AI literacy and access to detection tools, particularly in the Global South, leaving many communities vulnerable to the harmful effects of synthetic media.
 - **Initiatives:** Some efforts focus on creating locally adapted AI literacy programmes to bridge this gap, though these remain in early stages.
 - 'We face a significant challenge... the people who most need detection tools are the ones with the least access and the ones who are the least prioritized for access.'

3. Commercialisation and Commoditization of AI:

- **Increased Accessibility of AI Tools:** The commercialisation of AI has made powerful tools more accessible, raising concerns about their misuse at scale.
 - **Example:** The lowered threshold for creating sophisticated misinformation allows even those with limited technical skills to generate harmful content.
 - **Regulatory Challenge:** The absence of effective regulatory frameworks has not kept pace with the rapid commercialisation of AI tools.
 - 'Even those with limited education can easily create content with political leanings for disinformation campaigns.'

4. Evolution of Misinformation Tactics:

- **Shift to AI-Driven Disinformation:** The traditional use of human-operated misinformation campaigns is evolving into more sophisticated AI-driven tactics, making it harder to trace the origins and intentions behind disinformation efforts.
 - **Challenge:** The transition to AI-driven disinformation campaigns requires international collaboration to develop and implement effective detection and response strategies.

Addressing the Bigger Issues

1. Media Literacy and Public Education:

- **Enhancing Digital Literacy:** Traditional media literacy must evolve to include critical thinking and source evaluation, focusing on teaching the public how to navigate and critically assess information in the age of synthetic media.
 - **Proposals:** Develop and disseminate media literacy curricula targeting regions with low digital literacy, with a focus on recognising and debunking synthetic media.

2. Developing Robust Detection Tools:

- **Investing in Detection Technology:** There is an urgent need to develop and improve tools for detecting AI-generated content. This includes both company-based classifiers and post-hoc detection methods that can analyze content across platforms.
 - **Challenge:** Detection equity remains a significant issue, with those most in need of these tools—such as grassroots activists and civil society members in the Global South—having the least access to them.
 - 'Detection equity remains a significant challenge... detection tools, even when well-used, are about 85 to 90% effective.'

3. Ethical AI Governance:

- **Establishing Ethical Standards:** Developing comprehensive ethical guidelines and legal frameworks is crucial to govern the use of AI, particularly in high-stakes areas like human rights and journalism.
 - **Challenge:** The lack of transparency in AI development, especially concerning the datasets used for training models, raises significant ethical concerns that must be addressed through rigorous governance.

4. Collaboration Across Sectors:

- **Fostering Multi-Stakeholder Collaboration:** Continued collaboration between civil society, tech companies, governments, and academia is needed to address the challenges posed by synthetic media.
 - **Proposals:** Establish cross-sector task forces to enhance coordination and effectiveness in combating the misuse of synthetic media.
 - 'Recent collaborations between tech companies and civil society organizations have shown promise, but there remains a significant gap in bringing these efforts to scale.'

Specific Actions That We Need to Take

1. Next steps with Deepfakes Rapid Response Force:

- **Focus on High-Risk Areas:** Develop next steps for the dedicated task force to focus on detecting and responding to deepfakes, particularly in high-risk areas such as elections and human rights violations.

2. Develop a Media Literacy Curriculum:

- **Incorporate AI Detection Training:** Develop a curriculum aimed at journalists, human rights defenders, and civil society members, focusing on media literacy and the detection of AI-generated content. This curriculum should be localized and adapted to different cultural and linguistic contexts.

3. Promote AI Literacy and Training:

- **Localized AI Education:** Provide AI literacy and training tailored to the specific needs and contexts of different regions, translating and adapting resources to ensure they are accessible to those most at risk from the misuse of AI technologies.

4. Strengthen Legal Frameworks:



- **Implement Comprehensive AI Regulation:** Advocate for the development and implementation of enforceable legal frameworks to regulate the use of AI, particularly in the creation and dissemination of synthetic media.

REPORT OF DAY 1

Session 1: Welcome & Introduction

Opening Remarks and Announcements

The event began with general housekeeping and announcements. The organizers emphasized the importance of adhering to the **code of conduct** to maintain a respectful and safe environment.

Introduction and Objectives

The workshop aims to address the evolving challenges posed by **synthetic media** and **AI-manipulated content**. The primary objectives include:

1. **Comprehending developments** in Asia and the Pacific region since the last deepfakes workshop.
2. **Exploring the impact** of technologies that manipulate media on content created by human rights defenders, civic activists, and marginalized communities.
3. **Identifying and prioritizing pragmatic solutions** for defending against the dangers of AI-manipulated media, with a focus on tools for proving provenance and authenticity.

Key Themes:

- **Misinformation and Disinformation:** Understanding the use of synthetic media in spreading false narratives.
- **AI in Human Rights Documentation:** Balancing the benefits and risks of AI in civic journalism and advocacy.

This workshop is a continuation of efforts that began in 2019, with previous workshops held in Thailand and Kuala Lumpur and other regions of the world. The event seeks to deepen understanding, foster connections, and align global and local perspectives in tackling the challenges posed by these emerging technologies.

Session 2: Introduction to WITNESS

WITNESS Overview: The speaker provided an overview of WITNESS, a global organization dedicated to using video and technology to protect and defend human rights. WITNESS has a 30-year history, originating from the 1991 Rodney King incident, where video footage played a pivotal role in highlighting police violence and sparking national debate. This incident inspired the founding of WITNESS, with the idea that ordinary people could document abuses and drive social change.



Global Reach and Key Areas of Work: WITNESS operates globally, supporting communities in regions like Latin America, Africa, Asia, and the Middle East. The organization focuses on several key areas, including:

- **Supporting Marginalized Communities:** Documenting state violence and human rights abuses in regions such as Brazil, Nigeria, and West Papua.
- **Confronting Misinformation:** Empowering communities to combat misinformation and disinformation.
- **Advocacy on Land Rights and Climate Change:** Assisting indigenous and impoverished communities in protecting land rights and raising awareness about climate change.
- **Documenting War Crimes:** Supporting grassroots efforts to document war crimes in conflict zones like Ukraine, Syria, and Myanmar.

Systemic Change and Emerging Technologies: WITNESS also links grassroots efforts with global advocacy, particularly around technology systems. The organization recognises the importance of developing tools and standards for authenticity and trust in emerging technologies like AI and deepfakes. Their goal is to ensure that marginalized communities have a voice in these developments.

The session concluded with a summary of WITNESS' mission and its impact over the past 30 years. Participants were encouraged to consider how they could apply these lessons in their work, as the session transitioned to discussions on technology, advocacy, and the future of human rights documentation.

Session 3: Fortifying Truth in the Age of Synthetic Media

Overview

This session focused on the intersections between synthetic media, generative AI, and human rights, particularly concerning their impact on truth, trust, and authenticity. It highlighted WITNESS's approach, "Prepare, Don't Panic," in dealing with these challenges by leveraging their extensive experience in human rights documentation. The session was led by a speaker who laid out the foundational concepts, provided examples, and addressed questions from participants, emphasizing the need for a collective understanding and strategy to tackle these issues effectively.

Introduction and Context

The session began with the speaker sharing WITNESS's context in working with synthetic media and deepfakes. The organization has a long history of supporting human rights defenders (HRDs) and journalists, especially in their use of video technology. However, about 15 years ago, it became clear that engaging with the underlying technology infrastructure was crucial for frontline activists, as they were often disadvantaged within the systems of major platforms like Facebook and YouTube.



WITNESS's work in this area started around 2012, growing from their efforts to authenticate media during the Syrian civil war, where mobile phone videos' authenticity was often questioned. The organization has since been involved in developing systems for media authentication and tackling synthetic media challenges, focusing on "Prepare, Don't Panic." This approach involves providing direct guidance on verifying media, developing tools like ProofMode for creating more authentic media, and engaging with the infrastructure being built to ensure transparency in media creation.

Defining Synthetic Media and Generative AI

The speaker then defined synthetic media and generative AI, highlighting their relevance to the session. The focus was primarily on synthetic media, including deepfakes, which involve creating media—audio, video, images—using AI. A significant part of the discussion involved the concept of multimodality, where one form of media can be converted into another (e.g., text to image, video to video).

Examples were provided to illustrate the progress and potential issues in this field. The speaker used a series of examples from Midjourney, demonstrating the rapid advancement in text-to-image generation, and from OpenAI, showcasing the development of text-to-video and text-to-3D capabilities. These examples emphasized the increasing accessibility and sophistication of these technologies, raising concerns about their misuse, particularly regarding bias in AI models.

The Impact of AI in Media and Elections

A significant portion of the session was dedicated to discussing the impact of AI-generated content on elections. The speaker presented various examples to illustrate how AI has been used both positively and negatively in political campaigns:

- **AI to Communicate:** AI-generated content, such as voice cloning and avatars, has been used to communicate messages at scale and cross language barriers. For instance, the AI-generated voice of Imran Khan was used during Pakistan's 2024 election to declare his victory from jail, despite his imprisonment and disqualification from running. This was an example of how AI can be used to disseminate messages that might otherwise be impossible.
- **AI to Campaign:** AI-generated images were used in political campaigns in Argentina and India, depicting candidates in scenarios they never actually experienced. These images, although not directly influencing election outcomes, raised ethical concerns about the representation and authenticity of political figures.
- **AI as Soft Fakes:** The concept of "soft fakes" was introduced, referring to AI-generated content that humanizes or promotes political figures, making them appear more relatable or appealing. Examples included AI-generated audio of Narendra Modi singing a Bollywood song and AI avatars of Indonesian candidates, which were used to soften their public image.
- **AI to Ridicule:** AI has also been used for satirical purposes, creating content that mocks political figures. In Taiwan, a satirical video of Xi Jinping was circulated, while



in the US, a deepfake video of President Biden calling for a military draft for Ukraine was created as a "what if" scenario.

- **AI to Deceive:** AI-generated content has also been used to deceive voters. The speaker cited examples from Bangladesh and Pakistan, where deepfakes were used to falsely depict candidates withdrawing from elections or endorsing other parties. These manipulations, although not always impactful, signal a troubling trend in election-related disinformation.

Challenges in Verification and the Burden of Truth

Participants raised concerns about the burden of truth and the challenges in verifying AI-generated content. The speaker acknowledged that detection tools, even when well-used, are only about 85-90% effective. There is a significant "detection equity gap," where those who need these tools the most—HRDs, journalists—have the least access to them. This issue was flagged as a critical topic for further discussion, especially in terms of how to make these tools more accessible and reliable.

Broader Implications and Future Trends

The session also touched on broader implications, such as the use of AI in disinformation campaigns targeting civil society. Examples were shared from various regions, including Taiwan, Pakistan, and Bangladesh, where AI was used to create or manipulate content that undermines civil society efforts. The session highlighted the need for proactive strategies to combat these challenges, including increased digital literacy and advocacy for better content moderation on social media platforms.

The speaker concluded by discussing upcoming trends in AI, particularly the increasing realism and ease of creating synthetic media. They stressed the importance of preparing for a future where AI-generated content becomes even more sophisticated and pervasive, with significant implications for trust, authenticity, and the broader information ecosystem.

Participant Inputs and Key Discussions

Throughout the session, participants shared their experiences and raised important questions, contributing to a richer discussion:

- **Verification and Fact-Checking:** Participants emphasized the need for more accessible and reliable verification tools. There was a consensus that existing tools are not sufficiently available or effective, particularly for those without advanced technical skills.
- **State-Driven Disinformation:** Concerns were raised about the role of state actors in spreading AI-generated disinformation. Examples from recent events in Balochistan and other regions illustrated how governments use AI to manipulate public perception and suppress dissent.
- **Positive Use Cases of AI:** One participant highlighted the lack of research on positive use cases of AI in human rights and civil society. The speaker acknowledged this gap and suggested that future efforts should focus more on exploring how AI can be leveraged for positive outcomes.

- **Commercialisation and Small Agencies:** Another participant pointed out the growing role of small agencies and commercial actors in creating and distributing AI-generated content, often for unethical purposes. This raised concerns about the need for better regulation and transparency in the AI industry.

The session provided a comprehensive overview of the challenges and opportunities presented by synthetic media and generative AI, particularly in the context of elections and human rights. It underscored the importance of collective action, continued research, and the development of tools and strategies to ensure that AI is used ethically and responsibly. The discussions also highlighted the need for greater accessibility to verification tools and a proactive approach to addressing the challenges posed by AI-generated content.

Session 4: Generative AI and Synthetic Media in Asia-Pacific

Overview

The speaker opened the session by discussing the trajectory of generative AI and synthetic media in the Asia Pacific region. His presentation reflected on his experiences over the past five years, particularly since his first engagement with AI at the Internet Governance Forum (IGF) in 2019. The speaker emphasized the significance of the Asia Pacific region in the global AI landscape, outlining its unique challenges and opportunities.

The Early Stages of AI in the Global South:

The speaker's initial encounter with AI at the IGF 2019 highlighted critical discussions about AI development and its control. The forum sparked debates over who benefits from AI, particularly focusing on the disparity between the Global North and South. A representative from the Catholic Church notably questioned the benefits of AI for the Global South, raising concerns about the lack of human rights assessments in AI development. The event also saw the release of the GIS Watch report by the Association for Progressive Communications (APC), which addressed AI's impact on human rights and social justice.

Asia Pacific: A Rising Power in AI:

Five years after these initial discussions, the speaker noted a significant shift: the Asia Pacific region has emerged as a global contender in AI and synthetic media. Unlike the concerns initially raised about AI's invasion from the Global North, countries such as China, South Korea, Singapore, Japan, and India have become leaders in AI development. This transformation underscores the region's growing influence and its ability to compete head-to-head with the Global North in AI and synthetic media.

Key Players and Specialisations:

- **China:** Dominates in AI research, particularly in computer vision, natural language processing (NLP), and synthetic media.
- **South Korea:** Known for its advancements in AI-based language models and digital avatars, with companies like Naver and Kakao at the forefront.

- **Japan:** Excels in AI applications within media and entertainment, including music composition and digital animation.
- **Singapore:** A hub for AI innovation in financial services and smart city applications, aiming to lead in AI governance and sustainable practices.
- **India:** Rising rapidly with a large pool of tech talent, particularly in deepfakes and image generation.

Challenges and Opportunities in the Region:

The speaker outlined the competitive advantages and challenges faced by the Asia Pacific region in AI development:

- **Competitive Advantages:**
 - **Vast Data Availability:** The region's large population provides extensive datasets essential for AI training.
 - **Educational Focus and Rapid Adaptation:** There is a strong emphasis on education and technology, with a culture that quickly adapts to new developments.
 - **Government Support:** Certain countries in the region have created conducive regulatory environments, accelerating AI growth.
- **Challenges:**
 - **Ethical and Regulatory Issues:** The diversity in regulatory environments across countries poses challenges for creating uniform AI policies and ethical standards.
 - **Talent Gap:** Despite progress, there remains a shortage of highly skilled AI professionals to meet the growing demand.
 - **Data Privacy Concerns:** The region struggles with data protection, lacking strong institutions to enforce privacy regulations.
 - **Infrastructure Development:** There is a need for more robust infrastructure, particularly in computing power, to support advanced AI processes.

Popular Uses of AI in Asia Pacific:

The speaker provided examples of how AI is being integrated across various fields in the region:

- **Entertainment:** AI-generated virtual idols and influencers are becoming mainstream, especially in South Korea. These virtual characters are not only used in entertainment but also in commercial activities, blending technology with pop culture.
- **E-Commerce:** AI is enhancing personalized product recommendations, virtual try-ons, and product image generation.
- **Education:** Countries like China are leveraging AI for personalized learning and virtual tutoring systems.

- **Healthcare:** AI-driven diagnostic tools and personalized medicine are improving healthcare delivery in countries like Japan and Singapore.
- **Politics:** AI is increasingly used in political campaigns, producing content and propaganda to influence voter decisions.

Foresight and Ethical Considerations:

The speaker concluded by discussing the potential future implications of AI in the region:

- **Hyper Realistic Synthetic Media:** As AI-generated content becomes harder to distinguish from reality, ethical and legal challenges will intensify.
- **AI Alignment:** Ensuring that AI systems align with human values and ethical principles is becoming increasingly urgent.
- **AI Safety:** The region must prioritize discussions on AI safety and develop strategies to mitigate risks associated with AI deployment.

The speaker stressed the importance of ongoing dialogue and research on AI's impact in the Asia Pacific region. He encouraged participants to engage with these issues critically, particularly as the region continues to rise as a global power in AI and synthetic media.

Session 5: Experimenting with Generative AI

Overview: The session titled focused on hands-on experimentation with generative AI tools. The aim was to explore how these tools can be leveraged creatively for human rights advocacy, journalism, and other related fields. The session was structured into a 25-minute presentation, followed by 40 minutes of experimentation in breakout groups, and concluded with a 25-minute report-back session.

Participants were encouraged to experiment with various tools, such as Midjourney, DreamStudio, ChatGPT, Runway, and Hugging Face, to create content that could be used in advocacy campaigns or to explore the potential for misinformation.

Session Breakdown:

1. Introduction to the Experimentation Session:

- WITNESS began the session by emphasizing the importance of creatively using AI in storytelling. He acknowledged that many participants regularly work with content creation—whether in video, audio, or imagery—and highlighted the potential of generative AI tools to enhance their work.
- The session's purpose was to initiate conversations around the opportunities and ethical considerations of using synthetic media in advocacy. Jacobo mentioned previous experimentation by WITNESS colleagues, who used AI to create content for human rights campaigns.

2. Key Areas of Experimentation:



- **Identity Protection and Anonymity:**
 - Participants explored how AI can be used to protect the identities of vulnerable individuals. Techniques included using AI for face-swapping to replace the face of a person in a vulnerable situation with a synthetic face while maintaining the emotional impact of the original footage. This method aims to protect individuals while ensuring the narrative remains powerful.
- **Reconstructing a Location:**
 - AI was used to reconstruct locations that are either inaccessible or dangerous to visit. An example discussed involved creating a 3D model of a migration center in the United States, where physical access was restricted. The participants used Runway and Blender to create a digital representation based on reference images, offering a new way to visualize and document such sites.
- **Visualizing Testimonies:**
 - This approach involved generating visuals for testimonies that lacked accompanying footage. For example, an interview with a partner in Cambodia was used as a case study, where text-to-video tools like Runway ML were employed to create visual narratives based on the interview transcript. The ethical implications of altering text to fit AI prompts were discussed, emphasizing the need for careful consideration when interpreting and presenting testimonies.
- **Artistic Expression and Satire:**
 - AI tools were also used for creative and artistic expression, such as creating satirical content to raise awareness about social and political issues. One example provided was the creation of a video campaign advocating for land rights, where all visual and audio elements were generated using synthetic media tools.

3. Breakout Session:

- Participants were divided into four breakout groups, each tasked with experimenting with AI tools to create content based on two options:
 - **Option 1:** Explore the creative potential for human rights or journalism by creating a campaign.
 - **Option 2:** Generate or edit content with the intent to misinform about an election-related topic.
- Each group brainstormed ideas, selected tools, and created content, considering the ethical implications of their work. The tools suggested included Midjourney, Stable Diffusion, Canva, Runway ML, and Hugging Face, among others. The session also covered the need for signing in and the potential costs associated with some of these tools.

4. Report-Back and Conclusion:

- o After the breakout session, participants reconvened to share their creations and insights. Each group presented the content they produced, discussing the tools they used, the challenges they faced, and the ethical considerations they addressed.
- o The session concluded with reflections on the potential of generative AI in advocacy work and the importance of ongoing discussions about its ethical use.

Key Takeaways:

- Generative AI offers significant opportunities for creative storytelling in advocacy and journalism.
- Ethical considerations are crucial when using AI, particularly regarding identity protection, representation, and the potential for misinformation.
- Collaboration and experimentation are key to understanding and harnessing the full potential of AI tools in advocacy.

Next Steps:

- Participants were encouraged to continue experimenting with these tools in their work and to stay engaged in discussions about the ethical implications of AI in media and advocacy.

Session 5, Part 2: Hands-on workshop

Introduction and Group Formation:

Participants began the session by organizing themselves into groups, with the freedom to choose their group based on interest. The facilitator provided options for two main activities:

1. **Option 1:** Explore creative potential for human rights or journalism by creating a campaign.
2. **Option 2:** Generate or edit content with the intent to misinform about an election-related topic.

After a brief discussion, the group decided to focus on political misinformation, opting for the second activity. This would involve creating a synthetic image or video content related to a protest scenario.

Tools and Software Discussion:



The facilitator introduced various AI tools and platforms that could be used for generating and manipulating content:

- **ChatGPT:** Capable of generating text but not images.
- **Stable Diffusion (Stability AI):** Free for three days, then paid.
- **MidJourney:** Paid service, generates images from prompts.
- **Canva and Runway:** Free with limited options, capable of generating images and editing content.

The group highlighted the limitations imposed by some platforms, especially regarding generating politically sensitive or violent content. Specific examples included attempts to generate images of known political figures or scenarios involving violence, which were often blocked by the platform due to community standards.

Experimentation Process:

The group decided to experiment with the **DreamStudio AI** platform, starting with the creation of a hyper-realistic image of a street protest in Bangladesh. The initial prompt was:

- "Hyper realistic image of a street protester from Bangladesh, age of 20, male, wearing a black t-shirt and jeans, holding a stick, beating a full green uniform police officer, male in his 30s."

The platform generated an image, but the results were not entirely realistic, resembling more of a cartoon. The group discussed refining the prompt to achieve a more photo-realistic output. Adjustments included specifying that the image should look like a photo and considering the ethical implications of such content.

Challenges and Insights:

As the group continued to experiment, they faced several challenges:

- **Platform Limitations:** Some platforms struggled to generate realistic images or blocked certain types of content due to sensitivity.
- **Ethical Considerations:** The group acknowledged the need to consider the ethical implications of creating and spreading misinformation, particularly in politically sensitive scenarios.

The group also explored the possibility of using AI to manipulate existing images, such as real photos of protests, to create misleading content. This included techniques like inpainting and outpainting, as well as considering how AI could generate synthetic realities that could deceive audiences.

Final Thoughts and Next Steps:

Towards the end of the session, the group reflected on the exercise. They discussed how the generated images could be used in misinformation campaigns and the importance of refining prompts to achieve desired outcomes. The group concluded that while AI tools offer powerful capabilities, they also pose significant ethical risks, particularly in the context of political misinformation.

The session concluded with a plan to share the generated content with the group for further discussion on the ethical implications and potential uses of AI in advocacy and journalism.

Session 5, Part 3: Session Recap

The session began with an announcement about lunch arrangements, with specific instructions regarding meal tickets. Attendees were reminded to collect their lunch slips to access the designated restaurant areas. For those who indicated vegan preferences on their forms, a separate section was reserved.

Group Presentations

Group 1: Misinformation Campaign Targeting the Rohingya

- **Objective:** The group explored a real-world scenario by creating a disinformation campaign targeting the Rohingya, highlighting how misinformation can exacerbate tensions in Southeast Asia.
- **Execution:**
 - **Midjourney** was used to generate images depicting a Rohingya man holding an Indonesian voter ID and interacting with a police officer.
 - **ChatGPT** was employed to create a fictional WhatsApp transcript from a political candidate urging Rohingya individuals to vote for a certain candidate in exchange for financial incentives.
 - **11 Labs** was then used to convert the transcript into an audio message with a Singaporean accent due to limitations in the platform's accent options.
- **Outcome:** Although the scenario was hypothetical, it demonstrated the ease with which AI tools could be used to create convincing and harmful narratives in political contexts.



Group 1

Group 1's Result

Group 2: Protest Misinformation

- **Objective:** This group aimed to create a false narrative by generating images of protesters violently clashing with police forces, reversing the typical protester-police dynamic.
- **Execution:**
 - **DreamStudio** was used to generate a hyper-realistic image of a protest scene, focusing on portraying protesters attacking police officers.
 - **Challenges:** The group struggled with achieving the desired level of realism using the free version of DreamStudio and had to refine prompts to generate more convincing images.
- **Reflection:** The exercise revealed the limitations of free AI tools and the importance of refining prompts to achieve the desired outcome. Despite this, the group recognised the potential for misuse in spreading disinformation.



Tried first to create images with named individuals



Group 2

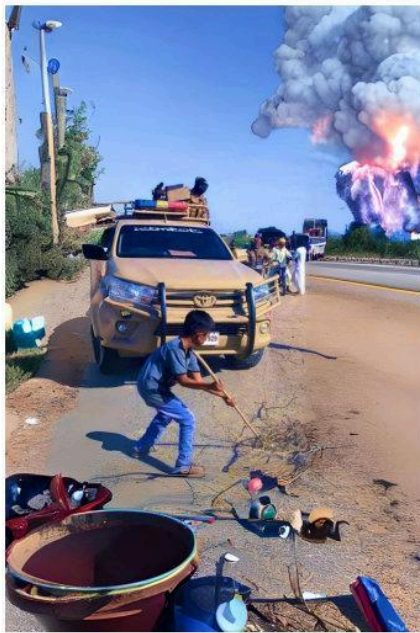
Group 2's Result

Group 3: Misinterpreting Protest Imagery

- **Objective:** The group focused on altering real protest images to misrepresent the context and intention behind them.
- **Execution:**
 - They began with a real image from Balochistan, showing a young boy in front of a military vehicle.
 - **Runway ML** was used for outpainting and inpainting, modifying the image to include an explosion and replacing the boy with a man holding flowers, thereby changing the narrative of the scene.
 - They further experimented with video creation using **Runway Video**, animating the altered image.
 - **11 Labs** was used to create a deceptive audio message that could accompany the manipulated visuals.
- **Outcome:** The resulting content showcased how AI can be used to alter the meaning of real events, making it challenging to discern the truth.



Protest image to be misinterpreted



Outpainting on Runway
ML



Inpainting to make it a man throwing flowers



Make this into a video

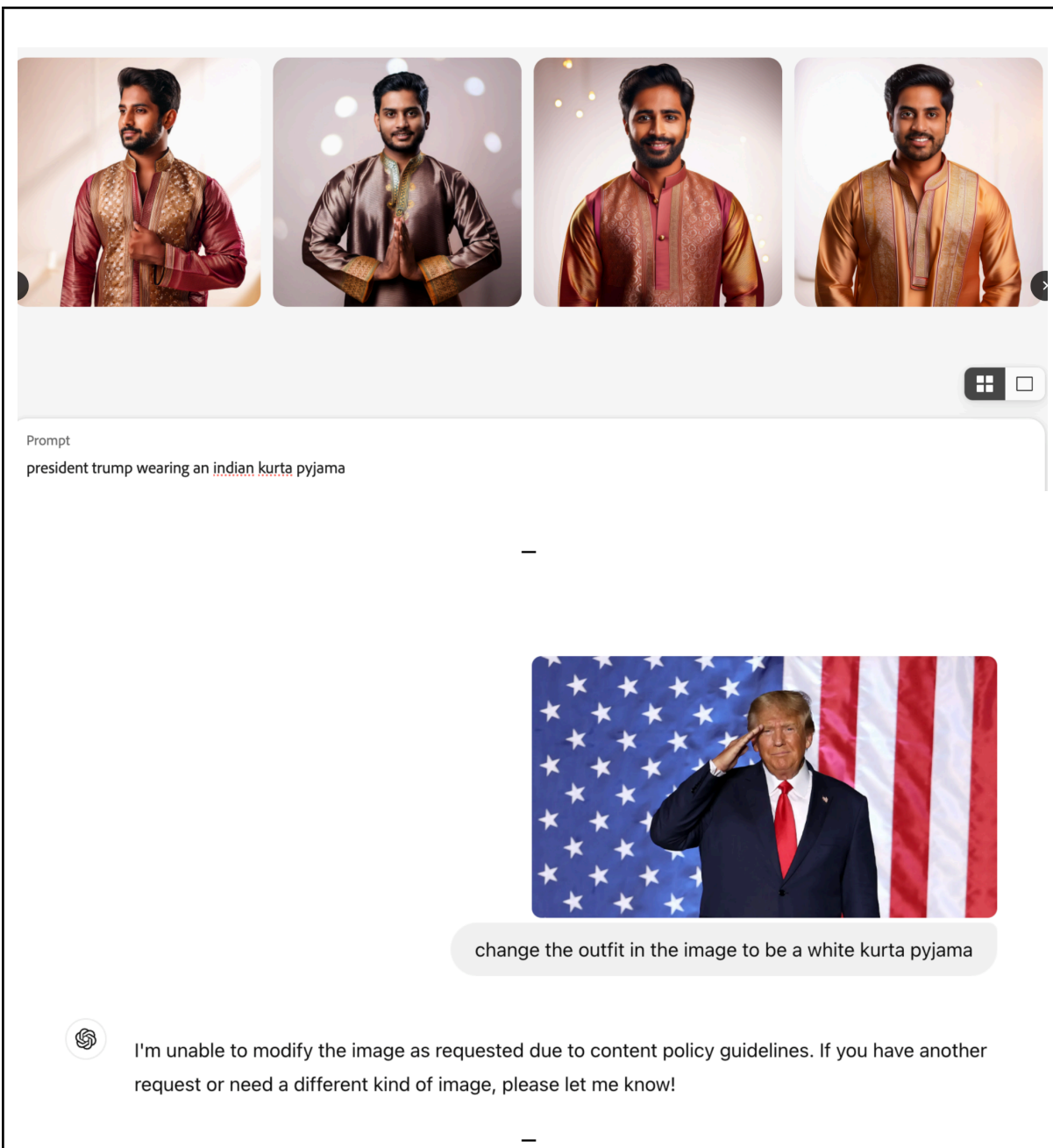
Group 3's Result (cont.)

Group 4: Campaign for Trump

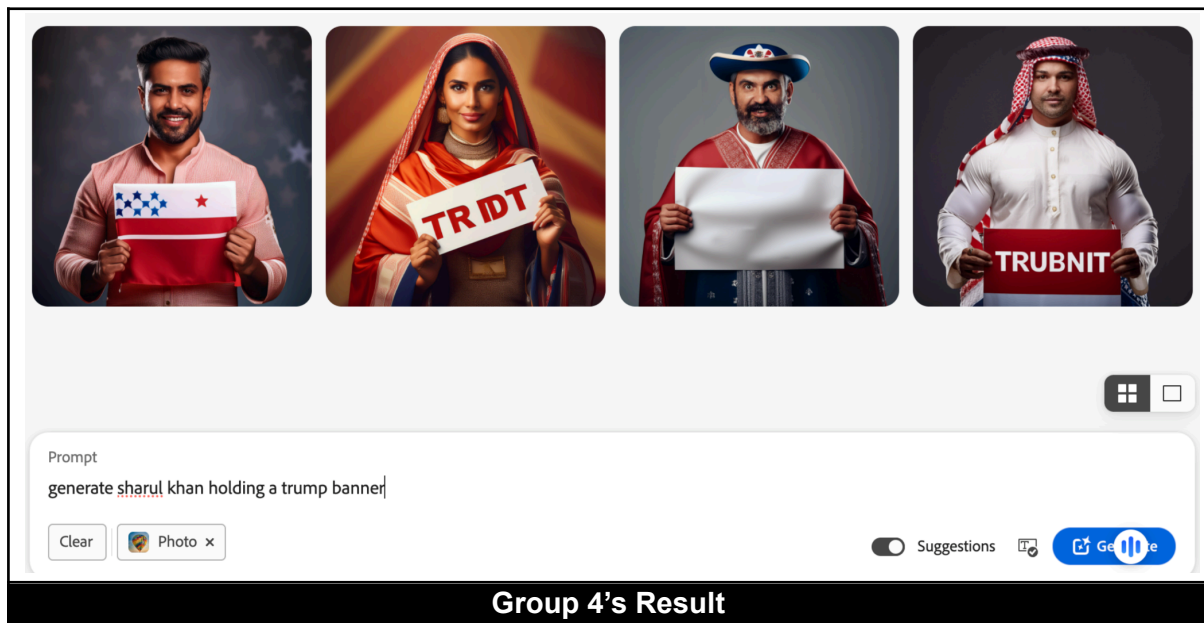
- **Objective:** The group intended to create a disinformation campaign targeting the Indian-American community, encouraging them to support Donald Trump in the US elections.

- **Execution:**

- **ChatGPT, Adobe Firefly, 11 Labs, and Google Gemini** were used to create images and messages showing prominent Indian figures endorsing Trump.
- **Challenges:** Content moderation barriers within the AI platforms prevented the group from generating images featuring real political figures. Even attempts to use an actor impersonating Trump were blocked.
- **Outcome:** The group found it increasingly difficult to use mainstream AI tools for disinformation due to content moderation, highlighting the ethical considerations embedded in some platforms.



The screenshot displays an AI image generation interface. At the top, there are four generated images of a man in a brown and gold Indian kurta pyjama. Below these images is a text input field containing the prompt: "president trump wearing an indian kurta pyjama". A horizontal line separates the prompt from the generated image. The generated image shows President Donald Trump in a dark suit and red tie, saluting in front of an American flag. Below the image is a text input field with the prompt: "change the outfit in the image to be a white kurta pyjama". At the bottom, a circular icon with a spiral pattern is followed by the text: "I'm unable to modify the image as requested due to content policy guidelines. If you have another request or need a different kind of image, please let me know!".



Group 4's Result

The session concluded with reflections on the challenges and ethical implications of AI in disinformation. The exercise underscored the power of AI in shaping narratives and highlighted the importance of its responsible use.

Session 6: Identifying and Prioritising Key Themes/Concerns in Human Rights & Journalism in APAC

Objective: The session focused on identifying key threats and concerns impacting human rights and journalism in the Asia-Pacific (APAC) region. Participants were divided into three groups, each addressing specific themes: Freedom of Expression, Trust and Truth, and Equity and Inclusion.

Group 1: Freedom of Expression, Assembly, and Association

Discussion Focus:

- **Key Threats:**
 - **Intimidation and Censorship:** Participants shared insights on how freedom of expression, assembly, and association are being attacked in their respective regions, especially in countries like Bangladesh, Thailand, and Indonesia.
 - **Government Suppression:** Participants discussed how governments in the region, particularly in authoritarian states, use intimidation tactics, mass arrests, enforced disappearances, and legal tools like blasphemy and criminal defamation laws to suppress dissent.
 - **Online Harassment:** The conversation highlighted how online harassment, particularly targeting women and journalists, is used as a tool to silence voices, with deepfakes and sexualised images being prevalent.

- **Data Security:** The lack of robust data protection mechanisms was also noted as a significant threat, particularly for marginalized communities, where breaches can lead to severe consequences.

Key Examples/Countries:

- **Bangladesh:** Extensive use of mass arrests and enforced disappearances, particularly targeting political opposition and student movements.
- **Pakistan:** The use of blasphemy laws and online harassment, particularly against women journalists, as a form of censorship.
- **Thailand:** Nationalism and the denial of rights to indigenous people, alongside legal and extrajudicial suppression of dissent.

Group 2: Trust and Truth in the Information Ecosystem

Discussion Focus:

- **Disinformation:** Participants explored how disinformation campaigns, often targeting vulnerable groups like religious minorities and political opponents, erode trust in the information ecosystem. The hyper-localisation of disinformation, especially via social media, was noted for increasing polarization and conflicts. Historical distortion, such as in the Philippines with historical revisionism, was also highlighted.
- **Censorship, Content Moderation, and Media Bias:** The session examined the role of technology companies and traditional media in either enabling or curbing disinformation. Issues like inconsistent content moderation, biased media coverage, and the blackout of protests were discussed. Examples included media biases in Sri Lanka and the lack of coverage of protests in Pakistan.
- **Challenges for Independent Media:** The struggle for survival faced by independent media was discussed, with emphasis on the financial instability and political pressure they face. Politicians discrediting traditional media further diminishes public trust, while the rise of online news platforms was seen as both a challenge and an opportunity.
- **Digital Literacy:** A significant challenge identified was the widespread lack of digital literacy, making communities more susceptible to disinformation. The absence of platforms for marginalized voices, including indigenous communities, was also a concern.
- **Role of Social Media Platforms:** Social media platforms were criticized for reinforcing filter bubbles and failing to curb misinformation, contributing to real-world violence and deepening divisions.
- **Emergence of AI:** The session highlighted concerns about the growing role of AI in journalism and elections, such as AI tools influencing voter decisions in Indonesia. Risks include deepfake threats, the burden on journalists to prove content authenticity, and the need for better tools and skills to manage AI's impact.

Key Examples/Countries:

- **Philippines:** Disinformation in political campaigns and historical revisionism.
- **India:** Disinformation targeting Muslim communities to incite violence.
- **Pakistan:** Media blackout of Balochistan and Pashtun protests.
- **Sri Lanka:** Media dominance by the Sinhala majority, suppressing opposing voices.
- **Indonesia:** AI's role in elections, raising concerns about artificial trust.

Group 3: Equity and Inclusion

Discussion Focus:

- **Intersectionality and Marginalisation:** The group discussed how certain communities, particularly women, LGBTQI+ individuals, and religious and ethnic minorities, face disproportionate threats in the digital space.
- **Economic Exclusion:** The session also touched on the economic disparities in access to digital tools and AI, which exacerbate existing inequalities. The discussion emphasized how AI and other digital technologies are often inaccessible to marginalized communities, increasing their vulnerability.
- **Policy and Legal Frameworks:** The group raised concerns about the adequacy of existing international human rights frameworks, particularly their application in non-Western contexts. The need for more culturally contextualized policies was emphasized.

Key Examples/Countries:

- **West Papua:** The ongoing suffering of indigenous people due to historical and current exploitation and the manipulation of their identity and history by the Indonesian government.
- **Thailand:** The challenges in advocating for transgender rights, particularly due to the lack of data and support systems.
- **APAC Region:** Disparities in digital infrastructure access across countries, leading to further marginalization of already vulnerable communities.

Conclusions and Recommendations:

- **Database Creation:** A critical need for comprehensive databases to document and support advocacy against online harassment and digital threats, particularly for women and other marginalized groups.
- **Intersectional Policy Approaches:** Advocacy for policies that recognise and address the intersectional nature of threats faced by different communities.
- **Equitable Access to Technology:** Efforts to make digital tools and AI more accessible to marginalized communities to reduce the digital divide and prevent further marginalization.

This session laid the groundwork for understanding the diverse and complex threats facing human rights and journalism in APAC, emphasizing the need for collaborative efforts to address these issues effectively.

Session 7: Risk ‘Spectrogram’

The session began with a continuation from the previous discussions, focusing on connecting the topics discussed earlier to the specific risks associated with generative AI (GenAI) and synthetic media. Participants were informed about a presentation on key risk factors and vectors related to GenAI, followed by a participatory exercise where they would evaluate these risks.

Presentation on Risk Vectors: WITNESS initiated the session by presenting the primary risk vectors identified in consultations related to GenAI and synthetic media. The categories discussed were:

1. Harms from Misuse and Abuse of GenAI and Synthetic Media

This includes deliberate misuse of AI to manipulate, deceive, or cause harm, often for personal, political, or economic gain.

- **Undermining Activists, Civil Society, or Journalists:** AI-generated content is used to discredit or silence these groups.
- **AI-Facilitated Gender-Based Violence (GBV):** Deepfakes and AI-generated content target women, LGBTQ+ communities, or other vulnerable groups.
- **Disinformation and Manipulation:** AI is used to spread false information, such as propaganda or manipulated media.
- **Fraud and Live Interaction:** AI mimics voices or creates deepfakes to perpetrate fraud in real-time.

2. Unintentional Harm from the Use of GenAI and Synthetic Media

This covers unintended consequences where AI, though not used maliciously, still causes harm.

- **AI Content During Elections:** AI-generated media can unintentionally mislead or confuse voters.
- **AI Content in Advocacy:** Similar risks apply outside elections, where AI-generated content can polarize public opinion.
- **Hallucinations:** AI-generated errors that produce convincing but incorrect information can spread misinformation.

3. Erosion of Trust

This highlights how AI can undermine public trust in media, institutions, and information.



- **Plausible Deniability & Believability:** Realistic AI fakes make it harder to trust true information.
- **Higher Burden to Verify Content:** The volume of AI-generated content overwhelms the capacity to verify its authenticity.
- **Higher ‘Burden of Truth’ on Journalists and Activists:** AI fakes increase the challenge of proving the authenticity of legitimate work.

4. Suppression of Rights

This involves using AI to suppress fundamental rights like freedom of expression.




- **Suppression in Satire, Parody, or Dissent:** AI is used to curb critical speech and creative expression, often through legal or extralegal means.

Exercise: Spectrogram Walk-Around: Participants were instructed to engage in a walk-around exercise where they would use coloured stickers to indicate the level of risk they perceived in various categories. The colors represented different levels of threat:

- **Pink:** High threat. (assigned value of 3)
- **Green:** Moderate threat. (assigned value of 2)
- **Yellow:** Minimal threat. (assigned value of 1)

Participants were encouraged to also add comments or examples using post-it notes to provide more context or specific concerns related to the risks discussed.

Threat category	Threat and comments				T
Information Manipulation and Deception	Personalized Messaging and Targeting <ul style="list-style-type: none"> ‘Amplify hate speech to taunt/target groups (LGBTI, women, etc.).’ ‘Used in India from 2021, where people are targeted within WhatsApp by their names & villages.’ 	17	5	3	64
	Creation of Synthetic Histories <ul style="list-style-type: none"> ‘Misinformation in contexts with high information vacuums.’ ‘Clear the history, evidence of HR violations.’ ‘Can be used as a tool for historical revisionism (i.e. political).’ 	20	4	2	70
	Disinformation and Manipulation <ul style="list-style-type: none"> N/A 	13	5	1	50
Threats to Democratic Processes and Governance	AI Content for Communication During Elections <ul style="list-style-type: none"> ‘Can be countered by AI/digital literacy.’ ‘AI in coms [communications] haven't really impacted election outcomes, but that (except Turkey, Bangladesh) doesn't mean it won't.’ 	9	8	5	48
	AI Content for Communication/Advocacy Outside Elections <ul style="list-style-type: none"> ‘Reduce trust in media as sources of un-altered/original media (photo, videos, illustrations, documents).’ ‘Photorealism concerns me – fake image could become more powerful/persuasive than real image. In advocacy outlets, does that really create justice for victims?’ ‘Govts, platforms should come out regulations, rules to regulate AI usage & labelling of news.’ 	0	9	13	31
	Suppression of Rights (in Satire, Parody, Art, Science, or Other Forms of Communication) <ul style="list-style-type: none"> ‘Expand censorship in many forms.’ 	6	7	10	42
Undermining Trust and Erosion of Truth	Plausible Deniability & Plausible Believability <ul style="list-style-type: none"> ‘Without good methods of proving digital authenticity, especially reduced sources/evidence due to higher burden of proof/authenticity.’ 	11	5	3	46
	Higher Burden to Verify (Individual Content and at Volume) <ul style="list-style-type: none"> ‘Can be countered by AI/digital literacy.’ 	5	11	4	41
	Hallucinations <ul style="list-style-type: none"> N/A 	0	5	11	21
	From Chatbots to Agentic AI <ul style="list-style-type: none"> ‘Interesting to see how low this is prioritized – this is the biggest priority that can be further eroded if Gen AI starts creating trustable information + reduces the impact of trust in humans.’ ‘If viable, can help with work of journalists e.g. to find out all personal and org relationships of politicians.’ 	0	1	23	25

	Higher 'Burden of Truth' on Journalists and Activists <ul style="list-style-type: none"> 'Also – interplay with higher burden that defenders/journalists face in understanding how AI systems work/supply chain of AI, etc.' 'For HR documentation in high risk area, AI could help, but can be challenged to "prove" the fact.' 	9	11	3	52
Gender-Based Violence and Targeted Harassment	AI-Facilitated GBV (Gender-Based Violence) <ul style="list-style-type: none"> 'Use AI-generated photo/video to defame LGBT/feminist activist.' 	16	0	0	48
	Undermining of Activists, Civil Society, or Journalists & Media <ul style="list-style-type: none"> 'Self-censorship by marginalized communities and women due to abuse using AI.' 'Trans-women targeting become more vulnerable.' 'Women, tribal & minority (race, color, caste) targeted.' 	20	3	0	66
	Leveraging Multimodality <ul style="list-style-type: none"> N/A 	3	11	14	45
Threat category	Threat and comments				T
Fraud and Criminal Activity	Fraud and Live Interaction <ul style="list-style-type: none"> N/A 	2	11	7	35
Ethical and Societal Implications	Commercialisation, Commoditization, and Accessibility of GenAI <ul style="list-style-type: none"> N/A 	5	12	8	47

T is a total threat value.

Key Discussions:

- Commercialisation and Commoditization of AI:** The accessibility of AI technologies, due to commercialisation, raises concerns about their misuse at scale.
- Synthetic Histories:** The ease with which AI can create synthetic histories was highlighted, emphasizing the risk of altering public perception or historical records.
- Personalized Messaging and Targeting:** AI's ability to personalize content for specific audiences was discussed as a potential tool for manipulation, especially in political contexts.

Conclusion of the Session: The session concluded with an open discussion on the insights gathered from the exercise. The facilitator emphasized the importance of understanding the dual-use nature of AI—while it offers powerful tools for communication and advocacy, it also poses significant risks that need to be carefully managed.

Final Thoughts: Participants were reminded to consider the broader implications of GenAI on freedom of expression, trust, and equity, and how these technologies could both amplify existing threats and create new ones. The session ended with an invitation for further reflection and discussion in upcoming sessions.

Session 8: The Impact of GenAI and Synthetic Media on Human Rights and Journalism

Overview: This session was a guided plenary discussion where participants reflected on the previous sessions' insights, especially connecting the identified harms and risks from the Spectrogram exercise to the broader issues of human rights and journalism. The session began with an invitation for participants to contribute their thoughts, particularly those who had engaged in breakout discussions earlier.

Opening Comments:

- The discussion started by encouraging participants to link specific examples from their work to the issues highlighted during the Spectrogram exercise. She invited them to add further comments using post-it notes on the wall or directly contribute to the ongoing discussion.

Key Themes Discussed:

1. Disparity in Risk Perception:

- **AI's Potential for Harm:** A journalist covering AI expressed surprise at the general underestimation of the risk posed by AI going rogue (like the concerns around "Skynet"). This highlighted a significant gap between the concerns of those working in AI development (focusing on the existential risks of AI) and those dealing with on-the-ground human rights issues.
- **Contextual Differences:** A participant from Vietnam noted that while some technical aspects of AI are heavily debated in tech circles, human rights defenders are more concerned with immediate and tangible threats, reflecting a human rights-based approach rather than a purely technical one.

2. Human Rights and AI Misuse:

- **Troll Armies and Dissent Suppression:** Participants discussed the current use of human-operated troll armies to suppress dissent and how this could evolve with AI's involvement. There was concern about the future use of AI agents to automate and scale such repressive tactics.
- **Smear Campaigns:** The use of smear campaigns, both online and offline, was identified as a growing threat, particularly targeting activists, human rights defenders, and minority communities. These campaigns, often fuelled by AI, are becoming increasingly sophisticated.

3. The Burden of Proof and Trust Issues:

- **Fabricated Evidence:** A case from India was highlighted where activists were imprisoned based on fabricated evidence, a situation exacerbated by AI's ability to create convincing fake content. The difficulty in proving the authenticity of evidence in a "post-truth" world was a major concern,

especially as institutions like courts may not be equipped to handle such challenges.

- **Polarisation and Trust Deficit:** The discussion also touched on how AI-generated content could further deepen societal polarization and erode trust in institutions, making it harder to maintain a shared understanding of the truth.

4. **AI's Role in Communication:**

- **AI for Positive Communication:** While concerns about AI's misuse were prevalent, some participants also highlighted its potential for positive use, such as protecting identities and improving accessibility in hostile environments. However, there was a consensus that such uses must be balanced with ethical considerations.
- **AI's Impact on Journalism:** The use of AI-generated content in journalism was debated, with concerns that it could undermine the authenticity and trust in journalistic work. The ethical implications of using AI to create media when access to real images or sources is limited were also discussed.

5. **Future Challenges and Ethical Considerations:**

- **Intergenerational Divide:** Concerns were raised about how different generations interact with AI and digital content, especially how older generations might be more susceptible to misinformation. This highlighted the need for digital literacy across all age groups.
- **Ethical AI Guidelines:** The discussion concluded with a call for the development of ethical guidelines and regulatory frameworks to govern AI's use, especially in high-stakes areas like human rights and journalism. Participants emphasized the importance of addressing the biases inherent in AI models and ensuring that technology does not exacerbate existing inequalities.

WITNESS wrapped up the session by thanking participants for their contributions and reminding them of the next day's agenda, which would focus on developing practical solutions and strategies to address the challenges discussed. The session concluded with an encouragement for continued reflection and engagement on these critical issues.

REPORT OF DAY 2

Session 1: Kick-off and Announcements

WITNESS initiated the session with logistical announcements and a brief energising activity.

- **Logistics:** Attendees were reminded to check their airport pickup times, submit any reimbursement forms by the tea break, and collect any additional materials available.
- **Morning Ritual:** The session began with a simple dance activity to re-energise participants, which involved a playful routine called the "banana dance." Participants followed along, mimicking actions such as peeling, shaking, and running with bananas.

Agenda Overview

WITNESS provided an overview of the agenda for Day 2, outlining key sessions and changes from the original schedule.

- **Session Sequence:**
 - Recap from Day 1.
 - A session on the technological and policy landscape in the Asia-Pacific region.
 - Discussion on Witness's work regarding synthetic media, focusing on disclosure, transparency, and detection.
 - A group photo session during the break.
 - A workshop on creating personas for fact-checkers, journalists, human rights activists, and consumers to inform discussions on detection, transparency, watermarking, and provenance solutions.
 - Plenary discussions on a civil society agenda for AI resilience and the future of synthetic media in the Asia-Pacific.
 - The day will conclude with action points related to policy, tech, legislation, and civil society initiatives.

Session 2: Recap of Day 1

WITNESS led a recap session, focusing on key themes and risks identified the previous day.

- **Key Themes and Risks:**
 - **Technological and Policy Gaps:** Participants expressed concern that policymakers, particularly in India, are not adequately addressing AI's risks. There is a lack of action plans to regulate AI responsibly.

- **Credibility in Research and Journalism:** A concern was raised about the potential for AI-generated content to undermine the credibility of research and human rights organizations, especially in sensitive contexts like Bangladesh.
- **Understanding and Engagement:** There was an observation that a very limited number of people understand AI's socio-technical implications, especially in regions with less exposure to these technologies.
- **Trust in Journalism:** Journalists expressed concern over maintaining audience trust and ethical standards in the face of AI-generated content.
- **Localization:** The importance of localizing AI discussions and combatting its negative effects, particularly in rural and TikTok-driven communities, was highlighted.
- **Challenges in Fact-checking:** The rapid advancement of AI-generated content, such as deepfakes, poses significant challenges to fact-checking, with concerns about big tech companies like Meta controlling detection technologies.
- **Homophobic Hate Crimes and Media Literacy:** There was a call to monitor AI's role in spreading homophobic hate crimes and to enhance media literacy across the Asia-Pacific region.
- **Narrative Control and Translation:** In Bangladesh, shifting government narratives around AI and digital blackouts were noted, emphasizing the need for real-time translation and transliteration services powered by AI.
- **Sexual Content and Moral Values:** Concerns were raised about the impact of AI-generated sexual content, with a broader discussion on the widening gap between technology, policy, and human behavior.
- **Responsibility to Educate:** Participants discussed the responsibility of journalists and civil society to broaden public understanding of AI's impact beyond just technological and journalistic circles.

Key Observations and Future Directions:

- **Technical Understanding:** It was emphasized that participants should not hesitate to seek clarification on technical terms and concepts. The goal was to ensure that discussions were accessible and inclusive, avoiding the exclusion of non-technical participants.
- **Prioritizing AI Risks:** It was noted that while some within the tech community, particularly those in companies like OpenAI, prioritize concerns about AI going rogue, many participants in this event were more concerned about immediate, on-the-ground risks such as surveillance, misinformation, and the erosion of trust.
- **Intersectionality and Targeting:** The discussion also highlighted the need to consider the intersectional nature of AI risks, particularly how they affect women, activists, and other marginalized groups.
- **Contextualization:** Participants were encouraged to contextualize AI-related challenges within known issues of freedom of expression, surveillance, and misinformation, rather than viewing AI as an isolated or purely technical issue.

Session 3: AI Regulation and Policy in Asia-Pacific

Overview: This session delved into the legislative tendencies and challenges surrounding AI regulation and policy in the Asia-Pacific (APAC) region. The discussion highlighted the broader context of AI governance, the influence of global norms, and the regional variations in legal approaches.

Key Points on Regulation in APAC

- **Regulation Beyond Law:** Shahzeb emphasized that regulation is not only about legal frameworks but also involves market forces, policies, strategies, and codes. However, for this discussion, the focus was on legal regulations.
- **Global Contagion Effect:** The "Washington Effect" and "Brussels Effect" were noted as significant influences, with emerging markets in the APAC region adopting regulations from dominant global markets to ensure compatibility and attract foreign investment.
- **Tensions in AI Governance:**
 - **Consumers vs. Producers:** The Global South often plays the role of consumers rather than producers of AI technologies, with significant financial flows moving from the South to the North.
 - **Technical vs. Socio-Technical Framings:** There is a predominant focus on the technical aspects of AI, with insufficient attention to the social implications.
 - **Principle vs. Practice:** The gap between theoretical principles like explainability and their practical application was highlighted.
 - **Binding vs. Non-Binding Frameworks:** Many current AI regulatory frameworks are non-binding, consisting of strategies and ethical guidelines that lack legal enforceability.

Country-Specific Trends in AI Regulation

- **Australia:** The country is characterized by extensive public consultations, particularly in areas like privacy and copyright law. However, despite regulatory provisions, there has been little enforcement action against major tech companies like Meta and Alphabet.
- **Japan:** Japan has taken a leadership role at the G7, promoting AI principles focused on consumer protection and market competition. The country has also responded to global trends, as seen in its legislative reaction to the Epic Games lawsuits against Apple and Google.
- **Singapore:** The country has recently amended its Broadcasting Act to regulate online communication services, especially social media, to protect against harmful

content. Singapore also has guidelines to curb the abuse of market power by dominant companies.

- **India:** India's recent legislative efforts include the controversial Information Technology Rules and the Digital Personal Data Protection Act. These laws, along with others, have inspired similar regulations in neighboring countries like Bangladesh and Sri Lanka, showcasing a regional contagion effect.
- **Bangladesh:** The country has mirrored India's regulatory approach, adopting similar laws in the digital and telecommunications sectors. Bangladesh's draft regulations reflect this influence, with some provisions directly copied from Indian law.
- **Sri Lanka:** Following mass protests in 2022, the Sri Lankan government swiftly enacted the Online Safety Act in early 2024. This act, passed with minimal public consultation, is a reactionary measure to regulate online spaces and curb dissent.
- **Pakistan:** The country has seen constitutional challenges to its cyber laws, particularly those related to online content and defamation. The judiciary's role in shaping these regulations has been significant, although the capacity of the courts to fully grasp AI's implications remains questionable.

Discussion Points

- **Judicial Capacity:** The judiciary in many APAC countries is still grappling with the complexities of AI, often lacking the necessary understanding to effectively regulate AI-driven technologies. The use of AI tools like ChatGPT by judges for decision-making was noted as problematic, particularly in countries like India and Pakistan.
- **Liability Issues:** A major challenge in AI regulation is determining liability—whether it should lie with developers, sellers, or users of AI technologies. The session underscored the need for innovative legal frameworks to address this issue, as traditional legal principles may not be sufficient.

This session highlighted the complex and evolving landscape of AI regulation in the Asia-Pacific region, underscoring the need for localized, nuanced, and enforceable frameworks to ensure that AI development and deployment are aligned with regional needs and rights.

Session 4: Protecting What is True, Detecting What is Fake

Overview: This session explored strategies for mitigating the risks associated with synthetic media and generative AI. The focus was on three primary areas: **media literacy**, **transparency**, and **detection**. These areas were identified as crucial in fortifying the truth and averting and mitigating the harms posed by AI-generated content. The session provided a framework for understanding these strategies and engaged participants in discussions about their potential effectiveness.

Key Themes

1. Media Literacy

- **Limitations of Teaching Recognition of Deepfakes:** The session emphasized that it is increasingly challenging, and perhaps unrealistic, to teach people to consistently recognise deepfakes. As generative AI evolves, distinguishing between real and fake content becomes harder. The consensus was that traditional media literacy, focused on critical thinking and source evaluation (such as the SIFT framework), remains essential, but expecting users to spot synthetic media is not a viable long-term strategy.

2. Transparency

- **Direct Disclosure:** This involves visible watermarks on AI-generated content to signal to users that the content may not be authentic. Examples discussed included watermarks from tools like OpenAI and Runway. However, concerns were raised about the durability of these watermarks, as they can be easily cropped or edited out.
- **Indirect Disclosure:** This includes invisible watermarks and fingerprinting:
 - **Invisible Watermarks:** These are embedded within the content and are not visible to the human eye but can be detected using specific tools. These watermarks are added during the content creation process, and their detection requires specialized software.
 - **Fingerprinting:** This method involves creating a unique code (hash) for each piece of content that can be compared with other content to detect similarities or alterations. Fingerprinting is a robust method for identifying and tracking content even after some modifications.
- **Verifiable Metadata:** Metadata provides information about the content's origin, including details like the time of creation, the type of device used, and any modifications made. Verifiable metadata is crucial for ensuring the integrity of the content throughout its lifecycle. However, the session highlighted that metadata can be stripped away during file transfer, especially on social media platforms.

3. Detection

- **Company-Based Classifiers:** Companies like OpenAI and Meta are developing proprietary tools to detect synthetic content generated by their platforms. These tools are specific to the content created by their systems.
- **Post-Hoc Detection:** This refers to generic detection tools that can analyse any content, regardless of the tool used to create it, to identify synthetic elements. The discussion stressed the need for widespread adoption and the development of standards to ensure these tools are effective across different platforms and content types.

Challenges and Considerations

- **Technical Challenges:** Despite the technical sophistication of watermarks, fingerprinting, and metadata, there are significant challenges in ensuring these methods are effective and resilient. For example, visible watermarks can be easily removed, and the durability of invisible watermarks is still in question. Moreover, the effectiveness of these tools depends heavily on the adoption of standards and the development of robust detection systems.
- **Social and Ethical Concerns:** The session also highlighted the ethical considerations surrounding the use of these tools. For instance, there is a balance to be struck between enhancing content authenticity and protecting user privacy. Additionally, there is a need to ensure that these tools do not inadvertently exacerbate issues like surveillance or censorship.
- **Legislation and Standards:** The discussion underscored the importance of developing and implementing standards and legislation that can guide the use of these technologies. Participants recognised the need for a comprehensive framework that addresses both the technical and social aspects of synthetic media.

The session concluded that while media literacy, transparency, and detection are crucial in addressing the risks of synthetic media, these strategies are not without their challenges. A multi-faceted approach that includes technical solutions, social awareness, and robust legislative frameworks is necessary to fortify the truth and protect against the harms of AI-generated content. The conversation around these tools and methods will continue as technology evolves and as society grapples with the ethical implications of their use.

Session 5: Synthetic Media Detection & Equity

Overview: The session focused on the detection of synthetic media, particularly in the context of elections and the work of online detectors. Participants were guided through practical exercises and discussions, highlighting the limitations and challenges of current detection tools and the broader implications for equity in media literacy and human rights. The concept of "detection equity" was introduced, emphasizing the challenges faced by those on the frontlines of democracy and human rights who lack access to these critical detection tools.

Key Discussion Points

1. Detection Tools and Their Limitations:

- The facilitator introduced practical exercises to demonstrate how detection tools like "AI or Not" and the 11Labs Classifier work. He highlighted that these tools are often unreliable, especially when dealing with re-recorded or altered versions of synthetic media. For example, a high-fidelity version of an audio clip might be correctly identified as synthetic, while a slightly altered version (e.g., re-recorded with background noise) could yield a misleading result.
- The discussion emphasized that publicly available detection tools often struggle with distorted or re-recorded media, background noise, and

non-standard manipulation. This discrepancy makes it difficult for these tools to keep up with the rapid advancements in generative technologies.

2. Practical Verification Exercises:

- o Participants engaged in exercises such as verifying a purported audio clip of President Biden and examining manipulated images of public figures like Donald Trump. The exercises showed that traditional journalistic practices—like background checks and reverse image searches—are still crucial, as current AI detection tools can be easily tricked by simple alterations such as cropping an image or lowering the resolution of a video.
- o Participants stressed that detection tools should not be the first line of defense; instead, common sense and traditional verification methods should be prioritized.

3. Ethics and the Complexity of Detection:

- o The session touched on the ethical considerations surrounding synthetic media. A participant raised questions about the acceptable threshold for fake content and the implications of hybrid media, where content is a mix of real and synthetic elements. This complexity challenges the notion of authenticity and makes the task of detection even more difficult.
- o It was mentioned that the lack of tools and expertise in the global south further exacerbates the problem, as these regions are often left behind in the fight against misinformation. The concept of "detection equity" was introduced, highlighting the need for accessible and reliable detection tools, especially for those working on the front lines of democracy and human rights.

4. Insights from Collaborations and Studies:

- o The facilitators shared insights from their collaboration with 19 institutions, including academic labs and tech companies. This pilot study, which involved around 40 experts, demonstrated that while traditional verification techniques remain relevant, AI detection tools are still far from being reliable and scalable.
- o It was discussed how specific training data is required for these tools to function effectively, but the models often struggle with real-world content that is noisy or includes non-English languages. This limitation is a significant challenge in regions where English is not the primary language, making it harder to detect and verify manipulated content.

5. Challenges and Future Directions:

- o The session concluded with a discussion on the future of detection tools. The facilitators stressed the importance of developing tools that are not only reliable and accessible but also transparent in their operations. They also pointed out the need for ongoing education and media literacy to help the general public understand the limitations of these tools and the importance of verifying content through multiple methods.



Group Activity: Participants were then divided into breakout groups to create personas (e.g., activist, journalist, fact-checker, consumer) and discuss the specific challenges these personas face in their work. They engaged in group activities to map out the practical steps involved in detection, considering the limitations of current tools. The aim was to connect these personas' experiences with the solutions and responses discussed during the session.

This session highlighted the complexities of detecting synthetic media, the limitations of current tools, and the critical need for equity in access to reliable detection methods, particularly in regions most vulnerable to misinformation.

Session 6: Range of Solutions and Personas Workshop

Breakout Group: Journalist Persona

Objective:

This breakout group focused on the persona of a journalist working to report on critical issues. The session aimed to explore how AI tools could support journalists in verifying and reporting on sensitive information, especially when physical access is restricted.

Persona: Journalist

- **Scenario:**

As a journalist, I want to report on critical issues to inform my audience. I receive a tip from a source about a well-known fugitive reportedly hiding in a specific location. The source provides a photo, but I cannot physically verify the information due to restrictions.
- **Challenges:**
 - The photo could be AI-generated or manipulated due to the abundance of public images of the fugitive.
 - The journalist cannot access the location to verify the story firsthand.
 - Time pressure to publish the story before competitors while ensuring accuracy.
- **AI Application:**
 - AI could assist in organizing and verifying large volumes of data.
 - AI tools might help cross-reference the image with existing databases to detect potential manipulation.
 - However, the journalist must be cautious about over-reliance on AI, as these tools can sometimes produce misleading results.
- **Process Considerations:**

The journalist would need to balance speed with accuracy, ensuring that AI tools are used to supplement, not replace, traditional verification methods. Peer review and editorial oversight remain crucial, especially in high-stakes reporting.



Breakout Group: Fact-Checker Persona

- **Scenario:**

As a fact-checker, I want to verify content rapidly and accurately to fortify the truth and tackle disinformation. My role is critical in debunking false information circulating during elections or crises.
- **Challenges:**
 - The fact-checker must navigate conflicting information and prioritize which content to verify based on virality and potential impact.
 - Limited access to advanced detection tools can hamper the ability to verify AI-generated content.
 - The fact-checker needs to educate the public about the reliability of AI tools and the complexities of content verification.
- **AI Application:**
 - AI could help automate parts of the verification process, such as image and video analysis.
 - Tools for detecting AI manipulation, while imperfect, are essential in the fact-checker's toolkit.
 - Education on AI tools' limitations and proper use is necessary to maintain public trust.

Breakout Group: Human Rights Activist Persona

- **Scenario:**

As a human rights activist, I want to document violations anonymously to ensure justice and accountability. The activist operates in a highly repressive environment where government surveillance and internet restrictions are common.
- **Challenges:**
 - Limited access to advanced technology and internet connectivity.
 - The constant threat of arrest or violence during documentation efforts.
 - Difficulty in securely sharing evidence due to government monitoring and censorship.
- **AI Application:**
 - AI tools could assist in anonymising and securely sharing documented evidence.
 - Detection tools could verify the authenticity of evidence before it is shared with trusted journalists or human rights organizations.
 - However, the activist must be wary of digital footprints and ensure that AI tools do not inadvertently expose their identity.

Breakout Group: Consumer Persona

- **Scenario:**

As a consumer, I want to stay informed and entertained by consuming content on social media. This persona represents an adult who navigates the internet for news, information, and leisure.
- **Challenges:**
 - The consumer is vulnerable to misinformation and AI-generated content that appears authentic.
 - Difficulty in distinguishing between real and fake content due to the sophistication of AI tools.
 - Potential exposure to online scams and targeted misinformation campaigns.
- **AI Application:**
 - AI could provide tools for consumers to verify the authenticity of content before sharing.
 - Platforms could integrate AI-driven alerts to flag potentially manipulated content.
 - Media literacy initiatives are essential to educate consumers on recognising and responding to disinformation.

Session 7: A Civil Society Agenda for AI Resilience and The Future of Synthetic Media in Asia Pacific

Overview: The session focused on discussing the concept of AI Resilience, particularly in the context of supporting frontline civil society actors such as journalists and human rights defenders (HRDs). It aimed to address the challenges posed by AI in misinformation, the need for AI literacy, and the development of resources to reinforce the credibility of authentic content.

Key Points Discussed:

1. AI Literacy and Training:

- **Basic AI Literacy:** Participants noted that even though AI is a buzzword, there is a significant need for introductory and basic AI literacy, especially among civil society actors. This literacy should focus on understanding AI's capabilities, limitations, and its implications for their work.
- **Localized Training:** There was a consensus on the importance of localized training that is directly relevant to the specific needs and contexts of different regions. This includes translating and adapting resources to local languages and cultural contexts.
- **Detection and Rapid Response:** The need for training on detection and rapid response mechanisms was highlighted, particularly in identifying

AI-generated misinformation and deepfakes. Participants discussed developing modules or guidance resources around this area.

2. Access to AI Tools:

- **Infrastructure for Open-Source Technologies:** There was an emphasis on the importance of infrastructure to support open-source technologies and tools specifically designed for HRDs and journalists. Participants discussed the potential of developing or improving access to these tools to enhance their work.
- **Collaboration with Big Tech:** Concerns were raised about the dominance of big tech companies in AI development, and the lack of access to AI tools for civil society organizations. It was suggested that collaboration with these companies could be explored to bridge this gap.

3. Innovation in Campaigning:

- **Use of AI for Advocacy:** Participants shared experiences and ideas on using AI to innovate in advocacy and campaigning, particularly in contexts where access to regions is restricted. AI could help visualize issues and support storytelling in new and impactful ways.
- **Creative Capacity Building:** The need for creative and engaging activities was emphasized, not just focusing on traditional training but also on making the learning process enjoyable and accessible for all involved.

4. Reinforcing Credibility:

- **Fortifying Truth Mechanisms:** A key discussion point was how to reinforce the credibility of content created by HRDs and journalists. This included exploring methods for content authentication and how to counter government narratives that label truthful information as hoaxes.
- **Human Rights Principles:** There was a discussion on grounding AI use and content verification in human rights principles. Participants highlighted the importance of not generalizing and ensuring that these efforts are sensitive to local contexts.

5. Collaboration and Networking:

- **Continued Collaboration:** The importance of ongoing collaboration between journalism, civil society, and other stakeholders was underscored. This includes creating spaces for sharing experiences and objectives, such as through regular workshops or salon talks.
- **Bridging Gaps:** Participants identified a gap between academics and civil society, stressing the need for more interaction and collaboration between these groups to leverage research and academic insights in practical applications.

6. Ethical and Long-Term Considerations:

- **Ethical Implications of AI:** Discussions touched on the short-term and long-term ethical implications of AI use, particularly in areas like education.

Participants recognised the need to think beyond immediate applications and consider the broader societal impacts.

- **Future of AI in Human Rights:** The potential future use of AI in human rights work was discussed, with participants projecting ahead to identify what steps need to be taken now to ensure AI supports, rather than undermines, human rights efforts in the coming years.

7. Challenges and Obstacles:

- **Tech Company Accountability:** Concerns were raised about the accountability of tech companies, particularly in regions where civil society has limited influence. Participants discussed the need for stronger frameworks to ensure tech companies take responsibility for the impacts of their AI technologies.
- **Localisation and Language Barriers:** The challenge of localizing content in multilingual countries was acknowledged, with participants sharing examples from countries like Bangladesh and Myanmar where multiple languages are spoken, necessitating a tailored approach.

Outcomes and Recommendations:

- **Establishing a Deepfake Taskforce:** A recommendation was made to create a task force focused on the detection and response to deepfakes, particularly in high-risk areas such as elections and human rights violations.
- **Developing a Curriculum for Media Literacy:** The creation of a curriculum for media literacy that includes AI detection and indicators of authenticity was proposed. This curriculum would be aimed at journalists, HRDs, and civil society members.
- **Long-Term Strategy:** Participants agreed on the need to develop long-term strategies for using AI in a way that supports human rights and civil society work, including ongoing monitoring and adjustment of these strategies as AI evolves.

The session concluded with an emphasis on the importance of continued collaboration and the need to focus on practical, actionable steps that can be taken in the immediate future to build AI resilience within civil society.

Session 8: Wrap up: Collaboration, Next Steps and Thank You

Overview: This session marked the conclusion of the two-day event, focusing on reflections, final thoughts, and next steps. Participants were encouraged to share their insights, experiences, and any collaborative opportunities that emerged during the workshop. The session also included logistical announcements and expressions of gratitude.

The session began with a request for participants to take a few minutes to fill out an evaluation form. The organizers provided a QR code for easy access to the form and mentioned that it would also be shared via Signal and WhatsApp for those unable to scan it.



Participant Reflections

1. Participant 1:

- o Expressed gratitude to WITNESS for organizing the workshop and providing a platform to learn about synthetic media and AI. The participant highlighted that this workshop felt more tangible and regionally relevant, particularly with the inclusion of teams from different countries in the Asia-Pacific region.

2. Participant 2:

- o Expressed shock at the rapid pace of AI development, particularly over the past two days. They voiced concern about the growing gap between the current understanding of AI and its ongoing advancements. The participant emphasized the importance of convening such workshops to share knowledge, collaborate, and keep up with this fast-evolving field.

3. Participant 3:

- o Shared initial apprehensions about participating due to a lack of expertise in AI. However, they appreciated the advice to approach AI from their existing experiences and expertise. This approach allowed them to build on their current knowledge of human rights and digital rights, making the learning process more manageable. The participant noted that while AI is intimidating, the workshop provided valuable insights on how to incorporate emerging technologies into human rights work.

Closing Remarks and Next Steps:

• Acknowledgements:

- o The facilitator thanked the Asia team for their hard work in organizing the event and extended gratitude to all participants for their active involvement and commitment over the two days. The facilitator emphasized that the insights and discussions from the workshop would significantly contribute to WITNESS's ongoing work in understanding and addressing the challenges posed by new technologies.

• Next Steps:

- o **Report and Blog:** A comprehensive report summarizing the workshop discussions will be prepared by Vasin, who diligently took notes throughout the sessions. This report, along with a blog post, will be shared with all participants.
- o **Media Sharing:** Photos and videos taken during the event will also be shared. Although a specific timeline was not provided, the facilitator assured participants that these materials would be distributed soon.

• Social Media and Continued Engagement:

- o Participants were encouraged to follow WITNESS's social media pages for updates on future engagements and resources. The facilitator mentioned the continuation of the WhatsApp group created for the workshop, stressing that it



would be used strictly for sharing relevant resources and updates. Participants who felt overwhelmed by the number of messages could opt to leave the group and follow updates via other channels like Twitter.

- **Logistical Announcements:**

- **Reimbursement:** Participants were reminded to collect their reimbursements immediately after the session.
- **Stationery:** Those working with communities were invited to take any remaining stationery items such as markers and notebooks.
- **Dinner Plans:** Participants were informed that dinner would be at a different restaurant from the previous night. They were requested to be punctual, with the meeting time set for 6:30 PM in the lobby.

The session and the event concluded with expressions of appreciation from the organizers for the participants' dedication and active contributions.