WITNESS

SEE IT FILM IT CHANGE IT

AUDIOVISUAL GENERATIVE AI AND CONFLICT RESOLUTION: TRENDS, THREATS AND MITIGATION STRATEGIES

SEPTEMBER 2024

Cite as:
Vazquez Llorente, R., Gildea R., & anlen, s. (2024) Audiovisual Generative AI and Conflict Resolution: Trends, Threats and Mitigation Strategies. WITNESS.

CONTENTS

EXECUTIVE SUMMARY

This forward-looking report investigates the evolving relationship between synthetic media and the information landscape in situations of armed conflict and widespread violence, with a particular focus on implications for conflict resolution and peace processes. Our analysis anticipates that advancements in audiovisual generative AI over the next 2 to 3 years could have notable implications for global security and stability. Yet, the intersection of generative AI and conflict dynamics remains underexplored, highlighting an urgent need for comprehensive exploration and understanding that prioritizes evidence-based analysis and does not succumb to hyperbolic rhetoric.

The adoption of generative AI techniques by malicious actors remains limited. However, we expect a number of stakeholders, with differing motivations, to increasingly leverage generative AI as awareness and capabilities develop. In particular, we foresee that malicious actors will continue to exploit the general public's difficulty in differentiating between authentic and fake content, aided by complex forms of deception, such as the nesting of synthetic lies and compositional deepfakes, whereby AI-generated outputs are set within layers of content that help to construct a persuasive but fabricated narrative. While it is true that generative AI has the potential to significantly enhance the capabilities of malicious actors, allowing them to produce a higher volume of realistic content efficiently, much of the existing analysis remains limited to issues around the volume or realism of synthetic content online. However, even a low volume of extremely realistic AI content, particularly if targeted effectively during critical moments, may sow confusion and elevate the likelihood of harm.

The distinctive contribution of this report lies in its detailed mapping of how generative AI capabilities link to emerging threats not only in the information landscape, but more specifically to conflict dynamics. We expect key developments in generative AI to include easy realism, enhanced video multimodality, intuitive user interaction, and unprecedented levels of personalization. First, easy realism will challenge technical trend lines by simplifying the process of creation and lowering barriers to entry, including tailoring content and higher fidelity. Second, increasing multimodality, that leverages multiple forms of data and machine learning models, will facilitate similar advances in AI-generated video as we have seen in the production of audio and images. Third, users will be able to interact with bots, deepfakes and avatars in an intuitive and spontaneous way. And fourth, it will be easier to produce deceptive information that credibly impersonates non-public figures, sometimes circulated out of public scrutiny to target specific individuals or population groups.

With a time horizon of 2-3 years, we foresee the potential for increased personalized and automated propaganda, decentralized disinformation networks, distortion of historical narratives, sophisticated psychological operations, misrepresentations in immersive environments, amplified gender-based violence online and offline, and heightened disinformation dynamics. While the risks posed by synthetic media in conflict settings will be influenced by a wide set of domestic and international factors, the impact of generative AI in the information ecosystem could lead key stakeholders to question the authenticity of crucial information; pose a direct threat to

the credibility, legitimacy, and physical safety of diplomats and groups engaged in conflict resolution; and place an additional strain on actors operating in conflict settings.

This report aims to provide a preliminary roadmap to governments, international institutions, the United Nations and non-governmental organizations (NGOs) to address the challenges posed by synthetic media to global security, ensuring that stakeholders are equipped to address and counteract the disruptive potential of AI in conflict resolution and peace processes.

For governments, international institutions and the United Nations, the recommendations include investing in the research, development and deployment of rights-respecting, independently audited detection technologies; as well as provenance and transparency standards that can trace the digital origin of content without compromising user privacy and safety. Governments are also advised to establish international standards for synthetic media creation, dissemination, and detection; criminalize the production and dissemination of non-consensual sexual imagery; and engage in diplomatic efforts to create binding agreements on the responsible use of generative AI in conflicts. Additionally, fostering international collaboration and information sharing, providing specialized training for key stakeholders, and establishing metrics to assess the effectiveness of implemented measures are crucial steps. NGOs should focus on building AI media literacy and disinformation awareness through public education initiatives; organizing workshops and training sessions; and establishing community feedback mechanisms. Facilitating cross-sectoral dialogues; engaging in international norm-setting and policy-making; monitoring and reporting synthetic media usage; and updating internal knowledge and practices are other essential actions for NGOs.

For all of these recommendations, it is imperative that communities impacted by conflict and local actors are centered, ensuring that their specific challenges and needs are addressed. This approach aims to ensure a coordinated, inclusive and effective response to the risks posed by synthetic media in conflict settings, protecting the rights and safety of affected communities.

INTRODUCTION

This report offers insights into key developments in audiovisual generative AI and their potential impact on the information landscape, with a particular focus on global security implications. Our work in this area is driven by the recognition that the role of deepfakes and synthetic media in influencing conflict dynamics and peace processes has not yet received adequate attention. This content can exacerbate mis- and disinformation, manipulate perceptions, and undermine trust, posing threats that can disrupt fragile negotiations and conflict resolution efforts. This said, while generative AI could empower malicious actors, adoption of the technology remains limited. By concentrating on emerging technology trends and building upon existing dynamics in the information landscape, we aim to shed light on future challenges in a constructive and non-hyperbolic manner, proposing policy approaches to mitigate the threats posed by audiovisual generative AI in conflict settings.

This research was conducted with input from consultations with non-governmental organizations engaged in humanitarian and diplomatic efforts in conflict, and it delves into the critical, yet underexplored, nexus of audiovisual generative AI and its implications for conflict dynamics and peace processes. We start by discussing the main categories of synthetic media and the core techniques behind their production. We then analyze advancements and key shifts in the evolution of generative AI, before examining the impact of audiovisual AI in the information ecosystem. Following this, we carry out a threat analysis of synthetic media in the context of conflict dynamics and peace processes. We finally identify and discuss anticipated global security and stability issues. We conclude by outlining policy options, strategies and actionable steps for governments, international institutions, the United Nations and non-governmental organizations involved in conflict resolution and peace processes.

I. TECHNIQUES FOR PRODUCING SYNTHETIC MEDIA

A generative AI model may be understood as "a machine-learning model that is trained to create new data […] one that learns to generate more objects that look like the data it was trained on."[1] For those working in conflict situations, having some knowledge of the techniques to generate synthetic content can help them increase their AI literacy. This can be key for engaging more meaningfully with stakeholders–including those who may be involved in the creation of deepfakes, and those who may have been targeted by them. Heightened levels of AI knowledge can also help organizations identify mis- and disinformation, and assess the reliability and trustworthiness of audiovisual content, as well as understanding trends in AI and how it may affect the information landscape and conflict dynamics.

Moreover, a basic awareness of generative AI techniques can debunk many misconceptions about AI capabilities, and help organizations design counter-strategies that are cognisant of technology trends, improving organizational resilience and the protection of staff and reputation. The tables in this section provide some illustrative examples to help understand the techniques and the synthetic media landscape in situations of conflict, widespread violence or political instability. This said, it is not always possible to identify the specific technique which has been employed, nor who produced the synthetic content–although there are investigative outlets and cyber-security firms who have started looking at the actors behind AI content generation and manipulation, building on their experience analyzing more traditional disinformation and influence operations.[2]

It is important to first clarify key concepts. "Synthetic media" encompasses any form of media, such as text, image, audio, or video, that has been artificially produced or manipulated, especially through use of AI. In this paper, we use "synthetic media" to refer exclusively to image, audio and video that is AI-generated or manipulated–that is why the title mentions "audiovisual AI". We use expressions such as "synthetic content" and "AI-generated or manipulated media'' interchangeably. While often conflated, "synthetic media" and "deepfakes" are distinct terms. "Deepfake" is a portmanteau of "deep learning", a subfield of machine learning technology, and "fake", typically describing realistic and often deceptive portrayals of individuals saying or doing something they never did. While all deepfakes are synthetic media, not all synthetic media are deepfakes. For example, an AI-generated image from a text prompt using a tool like Dall-E is a form of synthetic media but not a deepfake. Conversely, the video of Ukrainian President Volodymyr Zelenskyy asking troops to surrender is both synthetic media and a deepfake.[3]

The creation of synthetic media often involves various techniques, using different types of models and combinations of them to generate realistic content. The most popular techniques include General Adversarial Networks (GANs) and diffusion models.[4] With GANs two neural networks–termed the generator and the discriminator–are pitted

1 See: Zewe A. (2023) Explained: Generative AI, MIT News.
2 See: https://www.mandiant.com/ and https://www.graphika.com/.
3 Twomey J, Ching D, Aylett MP, Quayle M, Linehan C, et al. (2023) Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis Of Tweets That Discuss Deepfakes In The Russian Invasion Of Ukraine. PLOS ONE 18(10): e0291668. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0291668.
4 On GANs and diffusion models, see: Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y. (2014) Generative Adversarial Nets. Advances In Neural Information Processing Systems, 27. https://arxiv.org/abs/1406.2661; Chan SH. (2024) Tutorial On Diffusion Models For Imaging And Vision. arXiv preprint https://arxiv.org/abs/2403.18103.

against each other. The generator creates synthetic data, after which the discriminator tries to catch the generator and evaluates whether or not it is producing fake output. With each iteration, the generator refines the output. For example, a GAN trained on images of human faces can produce new and realistic versions of faces. Diffusion models involve the introduction of noise, that is randomness added to the data, to training datasets and then reconstructing the original data by reversing this procedure. These models have the capacity to generate high-quality images, for instance realistic patterns and textures, that are progressively outstripping the capabilities of GANs in some areas.[5] Recent research also shows that large language models (LLMs), or AI algorithms that excel at learning patterns, relationships, and contextual information in language, can also be leveraged for image generation.[6]

While methods are always evolving, we can identify a variety of AI techniques that are employed in the generation of synthetic media.[7] As seen in Table 1, different types of output such as text, image, video, 3D models, audio, and speech, can be generated from text prompts at speed. For instance, it only takes five minutes to generate 100 fake news articles, and less than a minute to create 50 social media comments. Audio capabilities have advanced rapidly in the last year and are becoming increasingly realistic in their outputs, allowing users, for instance, to clone someone's voice, mimicking their tone, pitch and inflexions, from a very short sample. Text-to-video is still ripe for development, as we have seen with Open AI's SORA (yet to be released to the public at the moment of writing).[8] Objects and context can be added or removed in images and video through functions such as "inpainting" and "outpainting".[9] Facial features can be edited, and motion and expression can be transferred across individuals.[10] Lip-syncing can be used to animate a person's mouth to fit a specific voice or audio, and face-swapping can be used to blend someone's face into existing content. Deepfakes may employ a single data "modality" (the type or form of data used), or combine multiple modalities that, for instance, blend visuals and audio to create particularly convincing outputs (e.g., AI avatars). Research has also shown how AI models can generate and alter satellite imagery.[11]

Table 1: Audiovisual generative AI techniques

Current popular techniques	These give us the ability to
Text-to- • Text • Image • Video • Audio • 3D model **Image-to-** • Image • Video **Audio-to-** • Image • Audio **Video-to-** • Video **Other**	• Style content • Create interactive content • Dub video with lip-syncing • Clone a voice • Edit facial features • Create motion • Transfer movements and facial expressions • Remove objects in video and image • Add objects in video and image • Add context to video and image • Make AI avatars • Swap faces

5 For work comparing outputs from both models, see: Dhariwal P, and Nichol A. (2021) Diffusion Models Beat GANs On Image Synthesis. *Advances in neural information processing systems* 34: 8780-8794. https://arxiv.org/abs/2105.05233; Stypułkowski M, Vougioukas K, He S, Zięba M, Petridis S, and Pantic M. (2023) Diffused Heads: Diffusion Models Beat GANs On Talking-Face Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5091-5100. https://arxiv.org/abs/2301.03396.

6 Koh JY, Fried D, and Salakhutdinov RR. (2024) Generating Images With Multimodal Language Models. *Advances in Neural Information Processing Systems*, 36. https://arxiv.org/abs/2305.17216.

7 For a general introduction to AI developments, including a useful glossary and information on the rise of multimodal systems, see this five-part series: Metz C. (2023) What's The Future For AI? The New York Times, March 31. https://www.nytimes.com/2023/03/31/technology/ai-chatbots-benefits-dangers.html.

8 https://openai.com/index/sora.

9 While inpainting is a technique that allows the user to enhance images and video by filling in missing or flawed areas, outpainting can extend an image or video beyond its original borders by adding visual elements.

10 An illustration of this technique can be viewed here: https://www.youtube.com/watch?v=F4G6GNFz0O8 (Last accessed June 21 2024).

11 Khanna S, Liu P, Zhou L, Meng C, Rombach R, Burke M, Lobell D, and Ermon S. (2023) Diffusionsat: A Generative Foundation Model For Satellite Imagery. arXiv preprint arXiv:2312.03606. https://arxiv.org/abs/2312.03606; Abady L, Horváth J, Tondi B, Delp EJ, and Barni M. (2022) Manipulation And Generation Of Synthetic Satellite Images Using Deep Learning Models. *Journal of Applied Remote Sensing, 16*(4): 046504-046504. https://usiena-air.unisi.it/retrieve/aeec603c-499f-4090-ab4a-5a53803206c7/046504_1.pdf; Knight W. (2021) Deepfake Maps Could Really Mess With Your Sense Of The World, Wired, May 20. https://www.wired.com/story/deepfake-maps-mess-sense-world/.

Table 2: Audiovisual generative AI techniques with examples related to situations of conflict, widespread violence or political instability

Technique	Explanation	Examples of tools	Examples in the wild
Text-to- • Text • Image • Video • Audio • 3D model	Text prompts can be used to generate any type of content from text, image, video, audio and 3D models.	ChatGPT text-to-text and text-to-image[12] Midjourney text-to-image[13] ImageFX[14] Runway ML text-to-video[15] Stable Diffusion text-to-video[16] Eleven Labs text-to-speech[17] Alpha Dragon text-to-voice[18] MusicFX text-to-audio[19]	In September 2023 AI-generated images were shared by al-Qaeda supporters, as part of a "celebration campaign" marking the anniversary of the 9/11 attacks.[20] Since late 2023, synthetic media have spread in Israel and Palestine, including fake images of victims, mobilized crowds, and areas which have been destroyed.[21] In post-coup Mali, numerous videos, seemingly including deepfakes, have circulated widely. These include "computer-generated voiceover talking about domestic politics or France's presence in the country."[22]
Image-to- • Image • Video	Images can be used to generate and style images and video.	Runway ML image-to-image and image-to-video[23]	In 2023, an Instagram account was created to circulate photos envisioning how children abducted during the Argentine dictatorship might look like today. Using pictures of parents whose children had been abducted, the artist prompted MidJourney to combine the images and create a portrait of the child likeness as an adult.[24] Since late 2023, TikTok users have been uploading fake videos of Hamas' victims created from real images of the victims.[25]
Audio-to- • Image • Audio	Audio inputs can be used to generate both images and audio.	Eleven Labs speech-to-speech Siri (new version)	In late August 2023, an anonymous TikTok account started posting dozens of clips of what it said were "leaked recordings" of Omar al-Bashir, the former leader of Sudan. It received hundreds of thousands of views. The campaign had used AI to impersonate al-Bashir.[26]
Video-to- • Video	Video can be used to generate and style video outputs.	Real-time ControlNet video stream-to-video stream[27]	The technology is still in the experimental phase. We have yet to identify any real-world examples in a conflict setting.

12 https://chat.openai.com/.
13 https://www.midjourney.com/home.
14 https://aitestkitchen.withgoogle.com/tools/image-fx.
15 https://runwayml.com/ai-tools/gen-2-text-to-video/.
16 https://stability.ai/stable-video.
17 https://elevenlabs.io/text-to-speech.
18 https://huggingface.co/spaces/AlphaDragon/Voice-Clone.
19 https://aitestkitchen.withgoogle.com/tools/music-fx.
20 Katz R. (2024) SITE Special Report: Extremist Movements Are Thriving As AI Tech Proliferates, SITE, May 17. https://ent.siteintelgroup.com/Articles-and-Analysis/extremist-movements-are-thriving-as-ai-tech-proliferates.html.
21 Klepper D. (2023) Fake Babies, Real Horror: Deepfakes From The Gaza War Increase Fears About AI's Power To Mislead. AP News, November 28. https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47.
22 Bennett C. (2022) Fake Videos Using Robotic Voices And Deepfakes Circulate In Mali, January 10. https://observers.france24.com/en/tv-shows/truth-or-fake/20220110-truth-or-fake-debunked-mali-robot-voices-deepfakes.
23 https://runwayml.com/ai-tools/image-to-image/.
24 Cholakian Herrera L. (2023) Argentina's Dictatorship Abducted Babies In The 1970s. An AI Project Imagines Them Today, July 20. https://restofworld.org/2023/argentina-stolen-babies-ai-art/.
25 Barinka A. (2023) TikTok Struggles To Take Down Deepfake Videos of Hamas' Victims, December 4. https://www.bloomberg.com/news/articles/2023-12-04/tiktok-videos-reanimating-hamas-victims-spur-deepfake-challenge?embedded-checkout=true.
26 Goodman J, and Hashim M. (2023) AI: Voice Cloning Tech Emerges In Sudan Civil War, BBC, October 5. https://www.bbc.com/news/world-africa-66987869.
27 https://huggingface.co/spaces/latent-consistency/Real-Time-LCM-ControlNet-Lora-SD1.5.

Technique	Explanation	Examples of tools	Examples in the wild
Other	Deepfake videos may be created using a variety of machine-learning techniques, like face swap, NeRF (that reconstructs complex three-dimensional scenes from a partial set of two-dimensional images), and motion transfers from one individual to another. AI can be used to generate digital representations of humans that mimic facial expressions, body movements, and speech patterns to deliver realistic messages. These are often called "AI avatars".	Deepfakes web[28] Synthesia[29] Reface[30] This Person Does Not Exist (GAN-generated)[31]	Since 2019, Mandiant has identified numerous instances of information operations that use GAN-generated profile photos, including by actors aligned with nation-states such as Russia, Iran, Ethiopia, Mexico, Ecuador, and El Salvador, along with non-state actors. The AI-generated origin of the profile photos is often obfuscated by adding filters or retouching facial features.[32] In the Russia-Ukraine war, deepfakes have falsely depicted Ukrainian President Volodymyr Zelenskyy announcing a surrender to Russian forces, and Russian President Vladimir Putin announcing a peace deal.[33] The government of Ukraine has released a video showing a deepfake Putin walking around Mariupol and describing the war crimes carried out by Russian forces.[34] After the Islamic State attacked a Russian concert venue in March 2024 a 92-second broadcast clip created by Islamic State supporters showed an AI news anchor in a helmet and fatigue, saying the attack was not a terrorist operation but part of "the normal context of the raging war between the Islamic State and countries fighting Islam."[35] In 2023, following the military's seizure of control in Burkina Faso the previous year, AI avatars of fictitious individuals claiming to be "pan-Africanists" and Americans were used to promote support for the new government.[36] In early 2023, it emerged that "news hosts" were spreading pro-government propaganda in Venezuela. The videos got hundreds of thousands of views on YouTube, went viral on social media apps like TikTok and were inserted as paid advertising on that platform. The propaganda videos appeared on the state broadcaster Venezolana de Televisión, which is affiliated to the Maduro regime.[37]

28 https://deepfakesweb.com/.
29 https://www.synthesia.io/.
30 https://reface.ai/.
31 https://www.thispersondoesnotexist.com
32 Cantos M, Riddell S, and Revelli A. (2023) Threat Actors Are Interested in Generative AI, But Use Remains Limited, Google Cloud Threat Intelligence, August 17. https://cloud.google.com/blog/topics/threat-intelligence/threat-actors-generative-ai-limited.
33 Notably, the Putin video was originally produced as a satirical deepfake and recirculated with a new framing after the Zelenskyy deepfake, illustrating the importance of context for how content is interpreted. On the use of deepfakes in the Ukraine-Russia war, see: Twomey et al. (2023).
34 Ukraine / Україна (2022) [Twitter] April 21. Available from: https://twitter.com/Ukraine/status/1517119052904374272 (Last accessed June 21 2024).
35 Verma P. (2024) These Isis News Anchors Are AI Fakes. Their Propaganda Is Real. Washington Post, May 17. https://www.washingtonpost.com/technology/2024/05/17/ai-isis-propaganda/.
36 Satariano A, and Mozur P. (2023) The People Onscreen Are Fake. The Disinformation Is Real. The New York Times, Feb 7. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html; Bahl V. (2023) Deepfakes Circulate Of AI 'pan-Africans' Backing Burkina Faso's Military Junta. France24, January 27. https://www.france24.com/en/tv-shows/truth-or-fake/20230127-deepfakes-circulate-of-ai-pan-africans-backing-burkina-faso-s-military-junta.
37 Singer F. (2023) They're Not TV Anchors, They're Avatars: How Venezuela Is Using AI-Generated Propaganda, February 22. https://english.elpais.com/international/2023-02-22/theyre-not-tv-anchors-theyre-avatars-how-venezuela-is-using-ai-generated-propaganda.html.

II. ADVANCES AND KEY SHIFTS IN AUDIOVISUAL AI

We now exist within an information environment where AI and non-AI media not only coexist but increasingly intermingle, creating a hybrid media ecosystem. For instance, a video's imagery may not have been edited using AI, but the audio may be synthetic, created from a text prompt or from audio input. In many ways, AI generation and manipulation techniques aside, we are already producing and consuming hybrid media in our everyday lives, often unknowingly. Many modern cameras already integrate AI functionalities to direct light and frame objects. iPhone features such as Portrait Mode, Smart HDR, Deep Fusion, and Night mode use AI to enhance photo quality. Android incorporates similar features and both have options for in-camera AI-editing.

We may classify digital manipulation that uses generative AI within two primary categories: identity and context-driven. Identity-driven synthetic media pertains to the depiction of specific individuals by using, for example, voice cloning or face swap technology, thereby making a person do or say things they never did (often referred to as "deepfakes").[38] Context-driven synthetic media portrays a setting or environment by using, for instance, text-to-image techniques to generate an event that did not occur (see Table 2).

The rapid advancement of generative AI technology in the last two years has resulted in substantial improvements in the realism of both identity and context-driven outputs, making it more difficult to distinguish between machine and human-made media. Compounding this challenge is the growing accessibility of audiovisual generative AI technology, with tools to create audio, image and video, as well as other deepfake techniques, available for monthly membership fees and without requiring exceptional computing power.

Against this background, we can identify several key shifts in audiovisual generative AI:

1. **Accessibility, commercialization and commoditization.** The proliferation of synthetic media has been facilitated by the growing accessibility and commercialization of the technology to generate it. Tools and software are increasingly available from commercial enterprises as well as through open-source libraries. This has removed hurdles to the creation of synthetic content and its commission for malicious and other purposes. Commoditizing AI through integration into smartphone features may also normalize manipulation, blurring lines in our perceptions as to what are accepted standards of truth.

2. **Ease of use.** Creating and tailoring outputs now requires decreasing resources and technical knowledge (e.g., use of text prompts to generate images, videos, and audio; drag and drop audio files for voice cloning). This not only accelerates the production of deceptive content, but also broadens the range of potential actors who may be able to create synthetic media.

3. **Speed.** AI tools can now produce a high volume of content at speed. While video capabilities for the general consumer are still at the early stages, large amounts of image and audio can be generated in a matter of minutes. As a result, there has been a significant growth in the volume of synthetic media circulating online, especially on social media. Platforms do not yet have robust systems to distinguish AI generation or

38 See Table 2: Ukraine President Zelenskyy was falsely depicted calling for the Armed Forces of Ukraine to surrender to Russian forces. For more on deepfakes in the Russia-Ukraine war, also see: Twomey et al. (2023).

manipulation at scale, and the increased volume may contribute to a more general skepticism among the public in their consumption of audiovisual media.[39]

4. **Variation.** AI techniques can be leveraged to create multiple depictions of the same event, for instance by showing different angles of a building, or diverse snapshots of a sequence of events. Variations of this kind can add further depth and credence to fake media, making it more difficult to decipher if the content is human-made.

5. **Multimodality.** Multimodal AI models can process a wide variety of inputs (including text, images, and audio) as prompts, and convert those prompts into various outputs, not just the source type. The proliferation of multimodal applications, along with their increasing accessibility and ease of use, means that even individuals with minimal technical skills can leverage multiple data types to flexibly create and disseminate sophisticated and persuasive deceptive content. This lowers the barriers for entry into the field of disinformation, potentially leading to a flood of complex fake and realistic content that could overwhelm the information ecosystem and sway public opinion or influence events.

6. **Personalization.** Synthetic content can be tailored to the personal appearance, behaviors, and vulnerabilities of specific individuals. Audio content can now be produced with such precision that impersonations become indistinguishably close to the original voice. This capability extends beyond just mimicking well-known public figures; it can also convincingly replicate the voices of less famous individuals, broadening the scope for misuse. These personalized deepfakes not only pose risks to individuals' privacy, security, and psychological well-being, but may be used to target and undermine critical social moments, such as elections or conflict dynamics. The creation and distribution of non-consensual sexual deepfakes has been steadily increasing over the years.[40] This content mostly targets women and girls, and builds on existing patterns of misogyny and sexism to degrade, humiliate and, in some contexts, potentially incriminate them. This not only damages the reputations of those depicted but can also deter participation in public or online spaces, with wider societal consequences in democratic representation, among others.

7. **Open source software.** Many of the harms brought by generative AI are compounded by the fact much of the software is open-source, which allows for anyone to modify and adapt the software.[41] This means that safeguards and disclosure mechanisms can be easily removed, leaving few barriers to the creation of harmful content by malicious actors. However, even with non-open source software that incorporates safeguards downstream, such as filters that prohibit the creation of certain content, WITNESS has been able to bypass some of these content moderation protections without involving much technical knowledge.[42]

Taken together, these trends make it easier for malicious actors to create and spread high-quality deceptive or manipulative information, at a new scale and volume, and with more personalization.

39 On the "liar's dividend" see: Chesney B, and Citron D. (2019) Deep fakes: A Looming Challenge For Privacy, Democracy, And National Security. Calif. L. Rev. 107: 1753. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954; Schiff KJ, Schiff DS, and Bueno NS (2023) The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?. *American Political Science Review*, 1-20. On deepfakes and public skepticism, see: Ternovski J, Kalla J, and Aronow P. (2022) The Negative Consequences Of Informing Voters About Deepfakes: Evidence From Two Survey Experiments. *Journal of Online Trust and Safety*, 1(2); Weikmann T, Greber H, and Nikolaou A. (2024) After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. The International Journal of Press/Politics, 0(0). https://doi.org/10.1177/19401612241233539.

40 Hoover, A. (2024) If Taylor Swift Can't Defeat Deepfake Porn, No One Can, Wired, Jan 26. https://www.wired.com/story/taylor-swift-deepfake-porn-artificial-intelligence-pushback/; Burgess M. (2020) Porn Sites Still Won't Take Down Nonconsensual Deepfakes, Wired, August 30. https://www.wired.com/story/porn-sites-still-wont-take-down-non-consensual-deepfakes/.

41 It is worth noting that there are different levels of openness for these models. For more, see: Solaiman I. (2023) The Gradient Of Generative AI Release: Methods And Considerations. In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 111-122. https://arxiv.org/pdf/2302.04844.

42 To understand the still limited effectiveness of content moderation in audiovisual generation tools, recent images of Taylor Swift of a sexual nature were created with Microsoft's Designer AI image generation software.

III. IMPACT OF SYNTHETIC MEDIA IN THE INFORMATION ECOSYSTEM

In considering negative effects on the information ecosystem, miscontextualized content and "shallowfakes", which do not require the use of AI techniques, remain the primary concern in terms of volume.[43] This said, AI-generated content is already generating two different types of dynamics that are particularly worrying.

The first is "plausible deniability". Already in 2018, Danielle Citron and Robert Chesney wrote of the danger of the "liar's dividend":[44] as the public becomes more aware of the existence of synthetic content, there is a growing skepticism towards all media, including genuine videos, images and audio recordings. This cynicism can be exploited by those implicated in real footage, who can now denounce that content as a "deepfake" to question the authenticity of any compromising material.

The second dynamic is "plausible believability". Research has shown that people display increased skepticism when presented with material they do not want to see, compared with information that aligns to their preferences.[45] By producing realistic-looking manipulated media, generative AI provides a way for supporters of a cause to cling to their existing beliefs and perpetuate entrenched narratives. This is similar to existing patterns of shallowfakes, which exploit cognitive biases by reinforcing pre-existing beliefs and confirming prejudices. One technique that could feed into this tendency is "nesting" and "compositional deepfakes", where AI-generated content is embedded within multiple media layers, making the detection of fakes significantly more challenging.[46] This method enhances the believability of the content, as the deepfake elements are obscured by the surrounding genuine material, thus reinforcing existing patterns of belief and increasing the overall persuasive power of the disinformation.

More generally, this may lead to what Aviv Ovadya has termed "reality apathy"; faced with an increasing stream of misinformation, people simply give up trying to distinguish what is real from what is false.[47] Consultations carried out by WITNESS with grassroots organizations in Latin-America, Africa, the USA, and South-East Asia found that this erosion of trust in the information ecosystem impacts, by extension, the work of civil society organizations (CSOs) on the ground documenting human rights abuses and international crimes. These CSOs now face the added challenge of proving that their genuine content is not the product of a sophisticated fabrication. This requires additional resources for verification and corroboration, which can be time-consuming, risky, costly, or out of reach.[48]

43 Johnson B. (2019) Deepfakes Are Solvable - But Don't Forget That "Shallowfakes" Are Already Pervasive. MIT Technology Review, March 25. https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/ ; see also: Helmus TC, and Chandra B. (2024) Generative Artificial Intelligence Threats to Information Integrity and Potential Policy Responses. April. https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3000/PEA3089-1/RAND_PEA3089-1.pdf.

44 Chesney B, and Citron D. (2019).

45 An extensive body of literature indicates that individuals may filter and interpret information differently depending on their pre-existing beliefs, including perceptions of fairness and accuracy of media related to conflict and political violence. For example, see: Vallone RP, Ross L, and Lepper MR. (1985) The Hostile Media Phenomenon: Biased Perception and Perceptions of Media Bias in Coverage of the Beirut Massacre. Journal of Personality and Social Psychology, 49(3), 577–585; Hirshberg MS. (1993) The Self-Perpetuating National Self-Image: Cognitive Biases in Perceptions of International Interventions. Political Psychology, 77-98. https://www.jstor.org/stable/3791394; Johnson, DP. *Overconfidence and War*. Harvard University Press, 2004; Jerit J, and Barabas J. 2012. Partisan Perceptual Bias and the Information Environment. The Journal of Politics, 74(3), 672-684. https://www.journals.uchicago.edu/doi/abs/10.1017/S0022381612000187.

46 Horvitz E. (2022) On the Horizon: Interactive and Compositional Deepfakes. Proceedings of the 2022 International Conference on Multimodal Interaction, https://arxiv.org/abs/2209.01714.

47 Ovadya A. (2019) Deepfake Myths: Common Misconceptions About Synthetic Media. Alliance for Securing Democracy, June 14. https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/.

48 Vazquez Llorente R, and McDermott Y. (2024) Truth, Trust, and AI: Justice and Accountability for International Crimes in the Era of Digital Deception. Just Security, June 17. https://www.justsecurity.org/96731/truth-trust-and-ai-justice-and-accountability/.

The above dynamics can build upon existing information operations strategies, such as the "firehose of falsehood" model of propaganda, which refers to an information tactic where multiple channels are used to transmit as much deceptive information to as many places as possible. The speed and volume of realistic content that can be generated using AI could facilitate and amplify the effects of this kind of model. Similarly, synthetic media may also contribute to the spread of "digital wildfires", whereby fake content rapidly spreads and causes harm to individuals, groups, and the wider information environment. As we will analyze later, in conflict and peace scenarios, this may not only mean creating inauspicious conditions for conflict resolution, but it also raises the possibility of concentrated attacks on peace processes or the actors involved in them.

Generative AI empowers those who wish to deny or dismiss real events, and creates a protective layer for mis- and disinformation. WITNESS has worked extensively on policy, legislative and technical responses to protect real content and identify fake material. For up-to-date resources and policy positions, please visit **www.gen-ai.witness.org**

CASE STUDY

Deepfakes Rapid Response Force: Testing Computational Detection in Conflict and Critical Human Rights Contexts

WITNESS' Deepfakes Rapid Response Force connects journalists and fact-checkers with experts in media forensics, AI synthesis and deepfakes to provide in-depth and fast analysis of content that may have been AI-generated or manipulated, when this material may contribute to violence or impact democracy and human rights. Our goal is to provide access to reliable detection tools and expertise, identify AI content (and conversely, content that may be real but is claimed as AI-manipulated), and promote responsible reporting on synthetic media.

ABOUT THE TECHNOLOGY

Computational detection uses AI technology to identify and assess manipulated or synthetic content. These tools use advanced algorithms to analyze various aspects of digital media, such as facial features, voice patterns, and audiovisual inconsistencies. Despite notable advancements, AI detection generally lags behind generative capabilities, and struggles with non-English languages, noisy environments, grainy-footage, low-resolution, and complex scenarios like conflict settings.

WITNESS has tested the effectiveness of publicly available detection tools, read more in our research for the Reuters Institute for the Study of Journalism.

BACKGROUND

In 2021, a video of Phyo Min Thein, the former chief minister of Yangon in Myanmar, was posted online.[49] The video showed him confessing to government corruption, but it was low resolution, the audio seemed out of sync, and his voice sounded unnatural. This led journalists and the public to suspect it was a fake. Initial analysis using online detection tools confirmed these suspicions.[50] However, an in-depth investigation by media forensics experts and organizations like WITNESS revealed that the video was real but staged as a forced confession.[51]

This incident highlights three critical issues in computational detection: access to specialized forensics expertise is out of reach to most actors at the digital frontlines, including in sensitive conflict settings,[52] publicly available detection tools lack accuracy and reliability, and their results are difficult to interpret.[53] These insights prompted WITNESS to begin work on how to address these socio-technical priorities.

DEVELOPMENT AND LAUNCH

After a year of global consultations and workshops involving over 170 participants from five continents, the Deepfakes Rapid Response Force ("the Force") was launched in March 2023. The Force connects journalists and media forensics experts to assess, analyze, and detect suspected deepfakes content in real-time to mitigate crises. The first 12 months served as a pilot, handling a few cases in collaboration with the International Fact-Checking Network. During this period, we received 35 cases, including AI-generated images, shallowfake videos, miscontextualized content, AI audio clips, and deepfakes. We escalated six cases, mainly from Nigeria, Sudan, India, and Venezuela, due to their potential impact in existing conflict, violence and disinformation dynamics.[54] Of these, five involved suspected AI audio clips and one was a suspected deepfake video.

EXPANSION

Facing an election year and anticipated trends of media disinformation and AI manipulation, we started expanding access to the Force in March 2024 to a number of countries preparing for elections. Between February and May 2024, we received an additional 16 cases, with 10 being escalated due to their impact on the local political landscape.

OPERATIONAL MODEL

The Deepfakes Rapid Response Force operates on an escalation model, receiving cases from partner fact-checking networks through a secure form. The fact-checking organization performs initial verification before submitting cases involving suspected AI manipulation impacting human rights or democracy. WITNESS conducts an internal assessment before escalating the case to the Force.

49 https://www.facebook.com/watch/?v=279901916856963 (Last accessed June 21 2024).
50 https://x.com/Milktea_Myanmar/status/1374372690128035851?s=20 (Last accessed June 21 2024).
51 Gregory S. (2021) The World Needs Deepfake Experts to Stem This Chaos, Wired, June 24. https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/.
52 Gregory S. (2021) Pre-Empting a Crisis: Deepfake Detection Skills + Global Access to Media Forensics Tools. WITNESS. https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/.
53 anlen s, and Vazquez Llorente R. (2024) Spotting the Deepfakes in this Year of Elections: How AI Detection Tools Work and Where They Fail, April 15. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail.
54 Christopher N. (2023) An Indian Politician Says Scandalous Audio Clips Are AI Deepfakes. We had them Tested. Rest of World, July 5. https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/.

The members then provide a transparent and detailed analysis within 24 hours, giving details about their analytical process and conclusions, and more importantly, the limitations of their models and interpretation. This evidence-based response serves two purposes: (1) to share back with the individual who submitted the content an answer they can use publicly; and (2) enhance media literacy and promote knowledge dissemination in a technical area difficult to penetrate. In addition to access to the Force, WITNESS has recently started to support the networks with training sessions about generative AI techniques and detection approaches.

KEY TAKEAWAYS

Given its initial success at helping frontline actors addressing content that can have high impact for communities-at-risk, we now possess a sample of cases from which we can draw a number of important insights. However it is important to note that the work of the Force continues to develop.

1. Response time: **Ensuring a rapid response is challenging due to operational, not technical, factors. Despite these challenges, expert teams provided analyses within 4-6 hours on average.**
   When we set up the Force we aimed for a detailed analysis within 24 hours, but we were uncertain if this goal was attainable. While occasionally our response suffered delays, they were unrelated to the detection process per se–for example, the Force team members work in different time zones and it could take some time for them to see a request to work on a case depending on the time of the day. In practice, expert teams managed to provide analyses within 4-6 hours on average.

2. Response format: **Providing detailed, accessible insights rather than binary results proved crucial. This approach helps journalists and fact-checkers make informed judgments.**
   We discovered the importance of providing detailed, accessible insights rather than simple binary results. Detection tools often offer real or fake results with confidence scores, but these can be difficult for journalists and fact-checkers to interpret. Our approach includes detailed explanations of the detection tools used and the results, translated into accessible language for non-technical audiences. This process underscores the importance of incorporating human expertise within any computational detection process.

3. Response accuracy: **Detection tools perform better when they are trained with data that closely matches the specific case they are used for. However, this can raise privacy and data collection challenges.**
   When the training set consists of the same media type (e.g., image, video, audio), manipulated using the same techniques, in the same language, and if relevant, data about the person, the accuracy of the results improve. Achieving this requires either collecting comprehensive data in advance or fine-tuning the model quickly. Original files, which often contain more metadata than copies shared online, can also enhance detection accuracy. However, access to these original files is often not an option for the journalists and fact-checkers we work with.

EXAMPLES FROM INITIAL CASES

AI audio from Nigeria
Our first escalated case involved a suspected AI audio from Nigeria. The content, appearing to be a recorded phone call, was difficult to analyze due to the poor quality and resolution of the audio, with background noise and overlapping voices. Detection tools, typically trained on clear datasets from controlled environments, struggled with the real-life complexity of this case. One analysis team requested additional time to retrain their models with background noise, which improved the accuracy of the results but delayed the response by 10 days. This case highlighted the gap between lab conditions and real-world scenarios, prompting one of the research teams to develop methods to tackle noisy and overlapping audio signals.

Politician's audio clips from India
We scrutinized two audio clips of an Indian politician who claimed the clips were fabricated. After thorough examination by three different teams, they concluded that while the first clip was unclear, the second clip was authentic. The work of the team helped a journalist from The Rest of the World, who referred the case to the Force, to report the results to the public in a nuanced manner, explaining the complexities of detection technology. This case emphasized the ease with which deepfake technology can be used to undermine authentic evidence, and the critical need for media literacy around AI to navigate such complexities.

Audio content and videos from Sudan
During the current civil war in Sudan, we assisted a Sudanese researcher in analyzing four suspected AI audio clips and one suspected AI video purportedly involving the Rapid Support Force (RSF) leader, Mohamed Hamdan Dagalo, who was rumored to be dead. With very few Sudanese actors conducting fact-checking about the country's political and conflict developments, our collaboration underscored the importance of establishing trustworthy relationships with partners with a specialized contextual knowledge and sophisticated understanding of the local information environment.

The work of everyone involved in the Force, from those referring cases to those analyzing them, helps combat disinformation that could have significant societal consequences. It also contributes to the advancement of WITNESS thinking on computational detection and, we hope, to the broader field of synthetic media detection. By learning from these examples, we continue to refine our approaches to bridging the gap in detection equity, aiming to provide more accurate and timely support to those on the digital frontlines.

IV. THREATS TO CONFLICT DYNAMICS AND PEACE PROCESSES

We define "threat" as any potential cause of an unwanted incident, which may result in harm to an individual, organization, or system. In the context of synthetic media, threats could include the creation and dissemination of deepfakes, disinformation campaigns, or AI-generated propaganda. While the scenarios examined in this section include emerging threats that are still on the horizon, our aim is to provide an initial analysis that can help governments and non-governmental organizations navigate key shifts in the information landscape. The precise impact of new risks posed by synthetic media in conflict settings will also be shaped by a wide set of domestic and international factors, such as patterns of violence, societal cleavages, institutional make-up, access to technology, existing regulations, or public education on generative AI, among others.

We can identify a number of potential challenges specific to individuals and non-governmental organizations operating in settings affected by conflict or widespread violence.

First, the possibility that governments and other actors may be employing generative AI to influence the information environment during a conflict, could lead stakeholders in a conflict resolution process to question the authenticity of audiovisual content. Synthetic media can be used to fabricate "evidence" or to depict false statements or actions, causing erroneous evaluations of conflict dynamics. It may also aid the cover up of crimes. By distorting the views of stakeholders and diplomatic personnel, generative AI may affect engagement between relevant parties, as well as influence the decision-making of diplomatic actors engaged in conflict resolution. Growing skepticism in the information environment could compromise negotiations, prolonging conflict resolution efforts and eroding the trust necessary for meaningful dialogue. Additionally, the potential misuse of synthetic media could severely undermine confidence-building measures, which are essential for successful peace processes. This could involve, for example, actors opposed to a peace agreement targeting joint initiatives and de-escalation protocols among negotiating parties through the spread of disinformation using generative AI.

Second, the proliferation of AI content could pose a direct threat to the credibility, legitimacy, and physical safety of diplomats and groups engaged in conflict resolution, such as armed actors, political parties, social movements, and non-governmental organizations. By misrepresenting the actions or intentions of the parties involved, synthetic media could undermine trust between negotiating parties, and the public too. It may also erode confidence in the integrity and neutrality of organizations involved in peace processes. For instance, high-quality and pervasive AI impersonations of leaders could provide "spoilers"—actors who believe that peace emerging from negotiations threatens their power, worldview, and interests, and undermine attempts to achieve it—with additional means to erode efforts at negotiation and reconciliation.[55] Taken together, these challenges could have serious consequences for diplomatic actors and other stakeholders, such as reducing engagement with them, putting staff at physical and reputational risk, preventing them from carrying out their duties safely and effectively, and undermining the delivery of their core missions.

55 For a discussion of spoilers in conflict and peace efforts, see: Nilsson D, and Söderberg Kovacs M. (2011). Revisiting An Elusive Concept: A Review Of The Debate On Spoilers In Peace Processes. International Studies Review, 13(4): 606-626.

Third, generative AI places additional strain on the resources of diplomatic actors and stakeholders operating in conflict settings. Investigating and responding to potential AI content requires increasingly significant time and resources, and may benefit from access to specialized technology, expertise in media forensics, and dedicated personnel to monitor and verify audiovisual content as well as to craft strategic responses. Many groups involved in this work have already faced challenges in acquiring the necessary resources and putting together the requisite infrastructure and protocols to effectively mitigate the risks posed by traditional disinformation, and synthetic media will only compound this burden.

In sum, generative AI, by reshaping the information ecosystem, can affect various conflict dynamics and peace processes on the ground. These technologies can be used to incite hostility and manipulate public perception, leading to the onset of political violence or escalations in previous patterns of attacks. Compromising deepfakes may not only be used to target vulnerable groups, fueling attacks, for example, against ethno-religious groups, women, and gender and sexual minorities, but can be used to target organizations involved in conflict resolution activities on the ground. AI can also dramatically alter the political landscape non-state organizations must engage with, including the perceptions and behaviors of key stakeholders.

Table 3 provides an overview of threats and vulnerabilities related to the use of audiovisual generative AI, with potential scenarios where synthetic media may have detrimental effects in global security. It is important to note that while threats posed by synthetic media often have distinctive features, several of the identified threats build on preexisting mis- and disinformation dynamics, whereby AI exacerbates or provides a new dimension to extant threats. In fact, it is in part through the lens of preexisting threats that we may ground or more reliably anticipate the impact of AI on future conflict dynamics.

Table 3. Threats, vulnerabilities and potential impact of audiovisual generative AI in conflict dynamics and peace processes

Threat	Vulnerability	Impact	Hypothetical examples
Impersonation of political leaders, or falsifying official or key stakeholder communications.	Lack of robust verification processes for audiovisual content. Shortcomings and lack of accessibility to computational detection tools and expertise. High susceptibility of populations to believable synthetic media. Social media platforms' inability to quickly identify and stop the spread of synthetic content. Difficulty in controlling narrative due to high engagement with sensational content. Challenges in fast and effective fact-checking at scale. Heightened plausible believability (that is, where a view appears sufficiently credible) of entrenched narratives.	Erosion of trust in media and official communications. Undermining the trust and security of the peacebuilding, dispute resolution, or conflict negotiation process. Disruption of genuine dialogue and potential derailment of peace processes. Onset of violence, or escalation in existing patterns. Negotiations and public opinion are influenced in favor of one party to the conflict. Increased polarization and manipulation of public sentiment. Heightening tensions and violence among communities. Affected populations or international stakeholders are misled about the realities on the ground. Disruption of decision-making processes. Misallocation of resources.	A deepfake video of a political leader advocating for violence, or declaring war, or making other, controversial statements could cause panic, escalate tensions, and lead to the onset of violence. A deepfake of a mediator sharing sensitive information could undermine the trust and security of the peacebuilding, dispute resolution, or conflict negotiation process. A voice clone of a politician spreading over close messaging platforms could cause confusion and increase tensions.
Spread of biased or false narratives.			Manipulated footage showing one side violating peace treaties could derail peace processes or negotiations. Credible deepfake news clips purporting false victories or losses in conflict zones could lead to misguided military or political actions. Synthetic interviews or speeches that are crafted to sway public opinion or international perception about a conflict could influence support or condemnation. Spreading false information about the activities of aid organizations, alleging biases or misconduct could discredit their efforts.
Authentic media dismissed as fake by those in power, parties to a conflict, or the public.	All of the above. Inability to push back against elite or entrenched narratives.	Genuine evidence is disregarded in decision-making and negotiation processes. Organizations, journalists and fact-checkers become increasingly under-resourced to verify content at scale.	Authentic media showing rights violations could be dismissed as fake by authorities, parties to a conflict or other actors that may find real content embarrassing or against their interests.
Creation of content that seems to comply with peace agreements or codes of conduct, but is misleading.	Lack of thorough vetting processes to authenticate adherence to agreements.	Erosion of the effectiveness of regulatory frameworks intended to foster peace.	Synthetic endorsements or approvals of community agreements or conduct codes, which suggest compliance but are actually fake, or that are actually not widely supported, could influence the peacebuilding, dispute resolution, or conflict negotiation process.
Gathering of sensitive information through credible-looking fake interactions.	Insufficient security measures to authenticate communications.	Compromise of strategic communications and leakage of sensitive information.	Fake video calls that look and sound like key negotiators could be used to extract critical negotiation strategies, discredit the process, threaten those involved, or retrieve confidential information.
Failure to detect or act against synthetic manipulations quickly.	Inability to limit or control the spread of harmful synthetic content on social media platforms.	Digital wildfires and firehose of falsehoods affect a peace process.	Lack of swift action by social media platforms in removing harmful deepfakes could complicate the debunking of fake content.

V. ANTICIPATED GLOBAL SECURITY AND STABILITY ISSUES

In the next 2 to 3 years, the evolution of audiovisual generative AI could pose sophisticated threats to conflict dynamics and peace processes. As these technologies become more advanced and accessible, their potential misuse in sensitive geopolitical contexts could become more significant and complex. Likely technical advances even in the next year include continued improvements in the ease of instructing AI tools with plain prompts; reduction in the quantity of input required; higher ability to tailor outputs to audiences and produce more realistic outputs; improved simulation of voice, emotion and acoustic environment; and comparable advances in video to what we now see in audio and images. More generally, we can expect more widespread production of sophisticated synthetic content within the information ecosystem.

We can categorize these likely advances into four main directions in which synthetic media will evolve: **easy realism, enhanced video multimodality, intuitive user interaction, and unprecedented levels of personalization.**

First, easy realism will challenge technical trend lines by simplifying the process of creation and lowering the barriers of creation, including tailoring and increased realism. Second, increasing multimodality, that leverages multiple forms of data and machine learning models, will facilitate similar advances in AI-generated video as we have seen in the production of audio and images. Third, users will be able to interact with bots, deepfakes and avatars in an intuitive and spontaneous way.[56] And fourth, it will be easier to produce not just personalized feeds, but deceptive information that credibly impersonates non-public figures, sometimes circulated out of public scrutiny to target specific individuals or population groups.[57]

As a consequence of this evolution, we may see:

1. **Personalized and automated propaganda.** Advancements in AI allow for real-time generation of fake video and audio, facilitating live interactions with fabricated personas. This capability could be misused in video calls, virtual meetings, and even interactive broadcasts, where the authenticity of live communication is exploited to deceive and manipulate viewers. Additionally, the use of AI to automate the creation and dissemination of propaganda at an unprecedented scale could exacerbate polarization, sow distrust in democratic institutions, undermine the integrity of electoral processes, and lead to political violence. AI could tailor propaganda to individuals' biases using vast amounts of data, effectively reinforcing false narratives or stereotypes and bringing division on a large scale. This could include AI outputs that overcome limits in dissemination due to linguistic barriers, allowing for the targeting of audiences across languages. This would heighten the risk of manipulation by creating fabricated and personalized audiovisual content that could, for instance, influence conflict dynamics, peace agreements, elections, regime stability, and other events with national security implications.

2. **Decentralized AI-powered disinformation networks.** The variation and volume in content that AI makes possible, could be further leveraged by decentralized networks that can

56 Horvitz E. (2022).
57 On AI and the effects of micro-targeting see: Simchon A, Edwards M, and Lewandowsky S. (2024) The Persuasive Effects Of Political Microtargeting In The Age Of Generative Artificial Intelligence. PNAS nexus, 3(2): 35. https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134.

operate autonomously to generate and spread disinformation without central coordination, making the sources of disinformation campaigns harder to trace and resistant to shutdowns.[58] By operating in a decentralized manner, these networks avoid central points of failure and can resist attempts at censorship or tracking. This makes combating disinformation significantly more difficult, as there is no single server or company to target for takedown requests.

3. **Distortion of historical narratives at scale.** Generative AI could be used to distort historical narratives and memory by fabricating audiovisual content that rewrites past events, conflicts, international crimes or atrocities. While archival and media manipulation is hardly new,[59] an accelerated and more sophisticated form of manipulation enabled by AI could do much to alter the public understanding of a conflict's origins or the current situation, affecting the international community's response and negotiation efforts. Complex forms of deception, such as the nesting of synthetic lies, whereby AI-generated outputs are set within layers of content that help to construct a persuasive but fabricated narrative, could be effectively utilized towards this end. Similarly, we may see content that tricks the human brain by being specifically designed to exploit how our minds process visual and auditory information.

4. **Sophisticated psychological operations (psy-ops).** One possible scenario is that the availability of generative AI tools will hasten the adoption of AI into "information operations and intrusion activity" by stakeholders to a conflict.[60] The integration of audiovisual generative AI technology into military strategies raises significant legal and ethical questions. For example, U.S. Special Operations Command (SOCOM) is reportedly exploring synthetic media technology for psy-ops, and other governments are seeking or already leveraging audiovisual AI.[61] The employment of synthetic media could fundamentally challenge trust and verification processes between parties in a conflict resolution process. As these technologies allow for the creation of highly realistic and convincing content, they increase the capabilities for propaganda, making it possible to influence public opinion and the stance of neutral parties more effectively than traditional methods. Furthermore, strategic disinformation using synthetic media could, for example, falsify orders from military leaders, sow confusion among the public and armed forces, and lend legitimacy to wars and uprisings to perpetuate violence.[62] The use of synthetic media by any branch of government introduces complex new dimensions to modern warfare,[63] diplomacy, international relations, and conflict resolution.

5. **Misrepresentations in immersive environments.** Virtual Reality (VR) and Augmented Reality (AR) technologies can create immersive simulations or augment real-world environments with digital elements. Research shows that immersive technologies that can effectively simulate real-world properties can create a strong sense of presence, making the virtual experiences feel real.[64] Experiences

58 For an instructive discussion of an autonomous disinformation experiment using AI, see: Banias MJ. (2023) Inside CounterCloud: A Fully Autonomous AI Disinformation System. The Debrief, August 16. https://thedebrief.org/countercloud-ai-disinformation/.

59 For example, manipulation of archival images or "photographic falsification" notoriously occurred under Soviet leader Joseph Stalin. For a discussion of this, and broader context, see: Fineman M. (2012) Faking it: Manipulated Photography Before Photoshop. Metropolitan Museum of Art, 2012, 69.

60 Cantos M, Riddell S, and Revelli A. (2023).

61 Biddle S. (2023) U.S. Special Forces Want To Use Deepfakes for Psy-Ops. The Intercept, March 6. https://theintercept.com/2023/03/06/pentagon-socom-deepfake-propaganda/. Generative AI is also alleged to have been used, for instance, in influence operations by groups from Russia, China, Iran, and Israel, see: De Vynck G. (2024) OpenAI Finds Russian And Chinese Groups Used Its Tech For Propaganda Campaigns. The Washington Post, May 30. https://www.washingtonpost.com/technology/2024/05/30/openai-disinfo-influence-operations-china-russia/; The Graphika Team (2023). Deepfake It Till You Make It. Graphika February. https://graphika.com/reports/deepfake-it-till-you-make-it.

62 Byman DL, Gao C, Meserole C, and Subrahmanian VS. (2023) Deepfakes And International Conflict. Brookings Institution. January. https://www.brookings.edu/articles/deepfakes-and-international-conflict/.

63 Beauchamp-Mustafaga N. (2024) Exploring The Implications of Generative AI For Chinese Military Cyber-Enabled Influence Operations: Chinese Military Strategies, Capabilities, and Intent. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/testimonies/CTA3191-1.html.

64 Heller B. (2020) Watching Androids Dream Of Electric Sheep: Immersive Technology, Biometric Psy, Biometric Psychography, And The Law, 3(1). https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1000&context=jetlaw.

in a virtual environment can have profound psychological effects, potentially altering attitudes, beliefs, and behaviors in the real world, which can either support peace initiatives or exacerbate conflicts depending on the integrity of the information presented. AR and VR have great potential as part of innovative solutions that facilitate dialogue, build empathy, and improve strategic planning. For example, multiple creative projects have leveraged VR technology to provide users with immersive experiences of historical atrocity crimes, simulating first-hand encounters and providing educational insight into cases of genocide and other forms of repression.[65] VR can also create neutral virtual spaces where parties to a conflict can engage in dialogue, reducing the need for physical presence and improving their security. However, AI-generated avatars or deepfakes could lead to deceptive engagement, manipulating negotiations and undermining trust.

Additionally, AR and VR could offer immersive training for mediators, negotiators, and peacekeepers by simulating complex conflict scenarios, and helping trainees develop skills for managing negotiations, peacekeeping operations and emergency responses.[66] VR could also be used for conflict analysis and scenario planning, allowing stakeholders to visualize the impact of different peace strategies and make more informed decisions. AR could assist in this process by overlaying real-time data on physical environments, aiding in strategic planning during peace processes. Similarly, AR could provide real-time battlefield information overlays to peacekeeping soldiers, enhancing their situational awareness. However, if real-time information is hacked or manipulated to include alternative scenarios or high-fidelity AI-generated disinformation, it can impact decision-making and lead to disastrous consequences.

6. **Amplified gender-based violence (GBV) online and offline.** Building on current harms such as stereotyping and misogynistic attitudes, generative AI can exacerbate the targeting of women and gender minorities, leading to severe offline consequences. The misuse of generative AI to create non-consensual sexual deepfakes of ordinary citizens is already a reality, but its use for targeting women in public positions will become more prevalent, amplifying existing gender biases and potentially inciting violence. Deepfakes depicting female leaders and activists in compromising or false situations could be used to discredit their authority, undermine their influence, and erode public trust. This could have a chilling effect, discouraging women from participating in political, social, and economic arenas, thereby reinforcing gender inequality. The psychological toll of being targeted by such malicious content could also be profound, leading to mental health issues and social isolation for the victims. In conflict zones, the deployment of synthetic media to spread gendered disinformation could intensify GBV by manipulating public perception and inciting hatred. Adversaries may use these tools to portray women and gender minorities as enemies or morally corrupt, justifying acts of violence against them. Such tactics could destabilize communities, disrupt peace processes, and prolong conflicts by sowing division and mistrust. The creation and dissemination of these AI-generated false narratives could also undermine efforts to seek justice and accountability for GBV, as it becomes harder to distinguish between real and fabricated incidents.

65 Examples include immersive VR experiences related to the Anne Frank House and ISIS' genocidal violence against the Yazidis and other victims. See: https://www.annefrank.org/en/about-us/what-we-do/publications/anne-frank-house-virtual-reality/ and https://www.nobodys-listening.com/. VR has also been used in a project to highlight prison conditions in Venezuela, see: Otis J. (2024) Virtual Reality Offers a Chilling 3D Look Inside Venezuela's Spiraling Prison, NPR, March 23. https://www.npr.org/2024/03/23/1239248108/virtual-reality-venezuela-prison-tour.

66 Studies have suggested the promise of AR/VR for enhancing the preparation, skills and knowledge of professionals responding to crisis scenarios, see for example: Sebillo M, Vitiello G, Paolino L, and Ginige A. (2016) Training Emergency Responders Through Augmented Reality Mobile Interfaces. Multimedia Tools and Applications, 75, 9609-9622.; Magi CE, Bambi S, Iovino P, El Aoufy K, Amato C, Balestri C, Rasero L, and Longobucco Y. (2023) Virtual Reality And Augmented Reality Training In Disaster Medicine Courses For Students In Nursing: A Scoping Review Of Adoptable Tools. Behavioral Sciences, 13(7).

7. **Heightened disinformation dynamics.** As generative AI continues to advance, we may see the amplification of other dynamics that were already present in the pre-AI era. For instance, simulated attacks such as fake attacks on politicians, citizens, or critical installations that could cause widespread panic, trigger security alerts, and destabilize regions. Unlike traditional shallowfakes, synthetic media can produce highly realistic and convincing content that is much harder to detect and debunk, making the panic and instability it causes more pronounced and harder to control. Economic turmoil is another grave concern; AI-generated media directed at financial institutions or other economic targets could lead to market manipulation, causing economic instability, rapid inflation, and food insecurity.[67] Synthetic media can also simulate authoritative sources or insider communications, making fraudulent information more believable and market manipulation more impactful. Furthermore, synthetic content could exacerbate social divisions by amplifying existing prejudices and undermining social cohesion. By disseminating falsified media that reinforces stereotypes, incites hatred, or spreads false narratives about marginalized communities, AI could heighten intergroup competition and increase the likelihood of conflict. The realism and personalization capabilities of synthetic media make it far more effective at targeting specific groups and individuals, thereby deepening societal divides more efficiently than traditional disinformation. Moreover, the erosion of trust in news and public institutions could accelerate as generative AI makes audiences increasingly skeptical of the authenticity of audiovisual content. This skepticism could lead to heightened polarization, the worsening of echo chambers, and the corrosion of key domestic and international institutions. Unlike regular disinformation, synthetic media's advanced capabilities in mimicking real human speech and behavior make it difficult for even the most discerning audiences to distinguish between genuine and fake content, thus deepening the crisis of trust.

Collectively, these anticipated scenarios underscore the profound and far-reaching implications of generative AI on global security and stability, compared to traditional methods of disinformation and audiovisual media manipulation.

67 We have already seen indirect effects on markets due to an AI generated explosion near the Pentagon, see: Marcelo P. (2023) Fake Image Of Pentagon Explosion Briefly Sends Jitters Through Stock Market. AP News, May 23. https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4.

VI. POLICY OPTIONS AND ACTIONABLE STEPS

In this section, we provide recommendations for governments, intergovernmental bodies and the UN, as well as NGOs, who seek to address the challenges posed by synthetic media to global security. These recommendations are grounded in extensive research, consultations, and grassroots engagement conducted by WITNESS over several years to understand the needs and priorities of communities most impacted by emerging technologies, especially advancements that facilitate the production and dissemination of deepfakes, synthetic media, and mis- and disinformation.

RECOMMENDATIONS FOR GOVERNMENTS, INTER-GOVERNMENTAL BODIES AND THE UNITED NATIONS

1. Invest in research and development of rights-protecting, independently audited and standardized detection and provenance technologies.

1.1 Allocate dedicated funding for the development and deployment of computational detection technologies that are rights-respecting, independently audited and standardized, ensuring resources are available for continuous improvement and adaptation. This should include regular audits of detection technologies and strategies to ensure they are effective and up-to-date with the latest advancements in AI, investment in training data that supports a diverse range of global languages relevant to conflict zones, access to detection tools by selected civil society organizations to help mitigate the risks of audiovisual generative AI in conflict, and ongoing technical support to frontline users to help them implement and maintain these tools effectively.

1.2 Invest in technical approaches to help identify if content originates from an AI system or a human, while protecting the privacy and security of users at all times and avoiding the collection of sensitive biometric data as a precursor to engage in online activities.

1.3 Invest in decentralized technologies to track the origin and modifications of digital content, ensuring traceability without requiring the default collection of a user's identity.

1.4 Engage directly with conflict-affected communities, journalists, fact-checkers and civil society groups with lived experience of, or that are familiar with, conflict environments to understand their specific challenges regarding synthetic media, ensuring that detection and provenance technologies address their needs and are available to them while effectively protecting their rights, privacy and safety.

2. Strengthen regulatory frameworks and international standards that go beyond risk-approaches and center international human rights and societal issues.

2.1 Advocate for the establishment and adoption of international standards regarding the creation, dissemination, and detection of synthetic media, including media provenance and transparency standards, that protect the privacy, freedom of expression, safety, and security of individuals producing and disseminating content online, whether synthetic, human-made, or hybrid.

2.2 Enact and enforce legislation that prohibits the use of generative AI to target or harass individuals through the production and distribution of non-consensual sexual imagery.

2.3 Engage in diplomatic efforts to create binding agreements on the responsible use of audiovisual generative AI in conflict settings, specifically addressing information campaigns using AI-generated content, documentation requirements for the use of generative AI in conflict settings, and oversight mechanisms to prevent misuse.

2.4 Establish norms prohibiting the use of synthetic media for psychological operations that spread disinformation, harmful content, or deceptively manipulate public opinion during armed conflicts or situations of widespread violence.

2.5 Incorporate input from local leaders, activists, and affected populations in conflict zones to ensure that regulatory frameworks and international standards address their specific experiences and needs, fostering an inclusive approach to policy-making.

3. Advance international collaboration and information sharing.

3.1 Foster collaboration between governments, tech companies, and civil society organizations to share information, research, and best practices for detecting and mitigating the impacts of synthetic media in conflict.

3.2 Provide specialized training for government officials, journalists, diplomats, peace negotiators, and civil society to recognize synthetic media and understand its potential impact on their work.

3.3 Create an international infrastructure to provide redress and support to individuals and organizations directly affected by synthetic media, with a priority focus on assisting individuals targeted by non-consensual sexual synthetic media and other tech-enabled gendered harms.

3.4 Provide financial and logistical support to CSOs working in conflict zones to help them implement the recommended strategies effectively.

3.5 Establish metrics to assess the effectiveness of implemented measures, including tracking the reduction in the spread of synthetic media, the improvement in public awareness, and the success of AI media literacy and other training programs.

3.6 Encourage direct involvement of community representatives from conflict-affected areas in international forums and collaborative efforts, ensuring their voices and experiences are integral to shaping strategies and solutions.

RECOMMENDATIONS FOR NON-GOVERNMENTAL ORGANIZATIONS (NGOS) ENGAGED IN CONFLICT RESOLUTION AND PEACE PROCESSES

1. Develop AI media literacy and disinformation awareness programs.

1.1 Collaborate with local media, community leaders, and educational institutions to build public education initiatives for conflict-affected populations to increase awareness of synthetic media and its potential to exacerbate conflict. These campaigns should focus on teaching people what generative AI can do and its limitations, how to critically evaluate audiovisual content, and how to report suspicious media.

1.2 Organize workshops, seminars, and training sessions for mediators, negotiators, and government officials to recognize and combat AI-driven information dynamics that may impact their work, including how to craft effective communication strategies to counteract mis- and disinformation. These initiatives should involve conflict-affected communities in the design and implementation of AI media literacy programs to ensure the content is relevant and culturally sensitive, fostering trust and effective communication.

1.3 Establish mechanisms for community feedback to quickly identify and address mis- and disinformation spread by synthetic media, including content that is not AI-generated or manipulated, but that governments and other actors are arguing to be "deepfaked". This could include hotlines, suggestion boxes, and community meetings.

2. Facilitate dialogue and collaboration across sectors and disciplines.

2.1 Facilitate discussions between AI researchers, tech companies, conflict resolution experts, CSOs, and other relevant stakeholders to help bridge the gap between technological advancements, and practical conflict resolution and peacebuilding needs. For instance, these may include pilot programs to test and refine detection technologies in real-world conflict settings, and conducting or supporting studies to understand the psychological and sociopolitical impacts of synthetic media on conflict dynamics.

2.2 Engage in collaborations with governments, tech companies, and other civil society organizations to share information, research, and best practices for detecting and mitigating synthetic media impacts.

2.3 Promote active participation of local communities in cross-sectoral dialogues to ensure their experiences and insights are considered in developing solutions, and to foster agency, ownership and collaboration.

3. Engage in international norm-setting and policy-making.

3.1 Engage with policymakers, technologists, and international organizations to develop guidelines or policies that mitigate the risks associated with audiovisual generative AI. This could involve setting standards for the ethical use of such technology in sensitive contexts like peace negotiations or conflict zones.

3.2 Advocate for transparency in audiovisual AI, encouraging technology companies, tool providers, and governments to clearly disclose the capabilities and limitations of their AI technologies, incorporate safeguards that prevent the creation of harmful content, and develop accountability measures.

3.3 Incorporate feedback from conflict-affected communities in international norm-setting discussions to ensure that policies and guidelines reflect their needs and protect their rights effectively.

4. Monitor and report synthetic media usage in conflict.

4.1 Facilitate the provision of critical front-line fact-checkers, local NGOs, and community leaders with access to AI detection tools and training on how to use them effectively. This should include engaging local communities in monitoring efforts, and leveraging their local knowledge and networks to enhance the effectiveness and reach of synthetic media detection and reporting.

4.2 Monitor the use of generative AI technologies in conflict zones, reporting on abuses and recommending corrective actions to technology companies, international bodies, or national and local governments and actors. This may include developing and deploying real-time monitoring systems to detect and respond to the dissemination of synthetic media.

4.3 Form rapid response teams equipped to handle mis- and disinformation crises caused by synthetic media. These teams should be trained in media forensics and crisis communication.

5. Update internal knowledge, protocols and practices.

5.1 Promote internal education on generative AI within organizations involved in conflict and peacebuilding to improve preparedness and resilience. This would help proactively anticipate risks associated with synthetic media technology. This might include internal assessments, evaluations, and feedback mechanisms to identify areas of improvement.

5.2 Involve community members in internal education efforts to share their experiences and insights on the impact of synthetic media, enriching the understanding and preparedness of the organization. This consultative process may also include engagement with other relevant parties to understand their approach to the use of generative AI, as well as connecting with experts in computational detection and content provenance and authenticity.

5.3 Conduct a risk assessment to identify areas of organizational activities that are particularly susceptible to the influence of synthetic media, and enact protocols to address these vulnerabilities and mitigate risk. This may include updating internal information verification procedures, incident reporting mechanisms, and crisis communications plans, such as protocols for how to respond to claims that material is faked with AI, and/or used to dismiss credible material.

ACKNOWLEDGEMENTS

This report was researched and written in June 2024 by Ross Gildea (consultant at Technology Threats and Opportunities), shirin anlen (Media Technologist, Technology Threats and Opportunities), and Raquel Vazquez Llorente (Associate Director, Technology Threats and Opportunities).

WITNESS is deeply grateful to all those who have helped shape our thinking in this area, in particular the staff at The Centre for Humanitarian Dialogue (HD) for sharing their time, expertise and experiences with us. We would also like to thank all the journalists, fact-checkers and researchers who have referred cases to our Deepfake Rapid Response Force, as well as the over forty experts who continue to provide WITNESS and their partners with pro bono computational models and expertise.

We wish to thank the following individuals, all of whom provided insightful comments at different stages of the drafting process: Sam Gregory (Executive Director, WITNESS), Maude Morrison (Deputy Director, Digital Conflict Division, Centre for Humanitarian Dialogue), Kat Duffy (Council on Foreign Relations), Natalie Caloca (Council on Foreign Relations), Daragh Murray (Senior Lecturer in Human Rights Law, Queen Mary university of London; UKRI Future Leaders Fellow), Belkis Wille (Associate Director Crisis, Conflict, and Arms Division, Human Rights Watch), Wendy Betts, Shannon Raj Singh (Special Advisor on Social Media and Conflict, Centre for Humanitarian Dialogue), Alexa Koenig (Research Professor of Law and Co-Faculty Director, Human Rights Center, UC Berkeley), and Adi Cohen (Memetica). We are also grateful for the reviewers who provided comments but wished to remain anonymous.

Design: Adam Cohen

OTHER WORKS AND PUBLICATIONS BY WITNESS

WITNESS is an international human rights organization that helps people use video and technology to protect and defend their rights. Our Technology Threats and Opportunities Team engages early on with emerging technologies that have the potential to enhance or undermine society's trust in audiovisual content. Since 2018, WITNESS has led a global effort, "Prepare, Don't Panic", to understand how deepfake and synthetic media technologies and, more recently, large language models (LLMs) and generative AI, are impacting at-risk communities around the globe, and how responding to real world harms and risks can bolster human rights.

These efforts have included contributions to technical standards, pioneering work facilitating real-time analysis of suspected deepfakes that can have important consequences for democracy and human rights, input to technology companies' policies as well as legislative proposals, experimentation with generative AI tools for human rights advocacy, public advocacy, and in-depth consultations with activists, journalists, content creators, technologists and other members of civil society.

GENERATIVE AI AND SYNTHETIC MEDIA

2024. Vazquez Llorente R, and McDermott Y. Truth, Trust, and AI: Justice and Accountability for International Crimes in the Era of Digital Deception. Just Security, June 17.

2024. anlen s, and Vazquez Llorente R. Spotting The Deepfakes In This Year Of Elections: How AI Detection Tools Work And Where They Fail. Reuters Institute, April 15.

2024. Vazquez Llorente R. Deepfakes in the Dock: Preparing International Justice for Generative AI. The SciTech Lawyer, 20:2, Winter.

2023. Vazquez Llorente R. To Protect Democracy in the Deepfake Era, We Need to Bring in the Voices of Those Defending it at the Frontlines. Council on Foreign Relations, December 18.

2023. Gregory S. Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI. Journal of Human Rights Practice, Volume 15, Issue 3, November, 702–714.

2023. Vazquez Llorente R, and Gregory S. Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights. Tech Policy Press, October 18.

2023. Gregory S. Testimony of Sam Gregory, Executive Director, WITNESS Before the United States House Committee on Oversight and Accountability, Subcommittee on Cybersecurity, Information Technology, and Government Innovation. 'Advances in Deepfake Technology.' November.

2023. Castellanos J. Fortifying the Truth in the Age of Synthetic Media and Generative AI: Perspectives from Latin America.

2023. anlen s and Vazquez Llorente R. Using Generative AI for Human Rights Advocacy. WITNESS, June 28.

2023. Vazquez Llorente R. Trusting Video In The Age Of Generative AI: Myths And Realities Of Building Provenance And Authenticity Technology. Commonplace, June 1.

2023. Cicek K, and anlen s. The Thorny Art of Deepfake Labeling. Wired, May 5.

2023. Vazquez Llorente R, Castellanos, J and Agunwa N. Fortifying the Truth in the Age of Synthetic Media and Generative AI, Perspectives from Africa.

2023. Castellanos J. Building Human Rights Oriented Guidelines for Synthetic Media: An Important Step Forward with PAI's Responsible Practices Framework. WITNESS Blog, February 28.

2022. Ajder H, and Glick J. Just Joking! Deepfakes, Satire And The Politics Of Synthetic Media. WITNESS and MIT Open Documentary Lab.

2019. WITNESS. Deepfakes: Prepare Now (Perspectives from South and South-East Asia).

2019. WITNESS. Deepfakes: Prepare Now (Perspectives from Brazil).

2019. WITNESS. What We Learned from the Pretoria Deepfakes Workshop (Full Report). February 19.

PROVENANCE AND AUTHENTICITY OF AUDIOVISUAL MEDIA

2024. Evaluating Digital Open Source Imagery: A Guide For Judges And Fact-Finders (2024), published online at www.trueproject.co.uk/osguide.

2023. WITNESS. The Relationship Between Human Rights And Technical Standard-Setting Processes For New And Emerging Digital Technologies. Submission to call for input for the Report of the High Commissioner for Human Rights, March 3.

2022. Gregory S. Synthetic Media Forces Us To Understand How Media Gets Made. NiemanLab.

2022. Castellanos J. The Rubber Hits The Road: From Niche To Potentially Systematic Use Of Provenance And Authenticity Infrastructure. WITNESS Blog.

2022. Vazquez Llorente R and Betts W. Coding Justice? The Tradeoffs of Using Technology for Documenting Crimes in Ukraine. Opinio Juris, October 22.

2021. Castellanos J and Gregory S. WITNESS and the C2PA Harms and Misuse Assessment Process: Confronting Potential Harms Early in the Process of Developing Authenticity and Provenance Infrastructure. WITNESS Blog, December 2.

2021. Gabriele C, Matheson K, and Vazquez Llorente R. The Role of Mobile Technology in Documenting International Crimes: The Affaire Castro et Kizito in the Democratic Republic of Congo. Journal of International Criminal Justice 19(1), March, 107–130.

2019. Ivens G and Gregory S. Ticks Or It Didn't Happen: Key Dilemmas In Building Authenticity Infrastructure For Multimedia. WITNESS Report, December.

WITNESS

SEE IT FILM IT CHANGE IT

www.gen-ai.witness.org