WITNESS SUBMISSION TO USA PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY WORKING GROUP ON GENERATIVE AI

Submitted: 31 July 2023

Contact: For further questions and comments contact Sam Gregory, Executive Director, sam@witness.org or Raquel Vazquez Llorente, Head of Law and Policy - Technology Threats and Opportunities, raquel@witness.org

About us: WITNESS is an international human rights organization that helps people use video and technology to protect and defend their rights.[1] 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.[2] Building upon years of WITNESS' foundational research and global advocacy on synthetic media, we've been preparing for the impact of artificial intelligence (AI) on our ability to discern the truth. In consultation with human rights defenders, journalists, content creators, fact-checkers and technologists on four continents, we've identified the most pressing concerns about how deepfakes, synthetic media and generative AI are impacting the information ecosystem and society at large. As part of this process, we have also developed guidelines for principled action and recommendations to policy makers, technology companies, regulators and other stakeholders.

Summary: This submission puts forward a set of recommendations on how to identify and promote the beneficial deployment of generative AI, as well as how to best mitigate risks. Our submission is informed by three decades of experience helping communities create trustworthy photo and video for human rights advocacy, protect themselves against the misuse of their content, and challenge misinformation that targets at-risk groups and individuals.

OVERARCHING PRINCIPLES FOR THE EQUITABLE, RESPONSIBLE AND SAFE DESIGN AND DEPLOYMENT OF GENERATIVE AI

Since 2018, WITNESS has been leading the first global effort to understand how deepfake technology is impacting communities at the frontlines of mis- and disinformation. Over the past year, building upon these consultations and foundational research, we have also incorporated an analysis of the threats and opportunities of generative AI.

In deep collaboration with leading human rights defenders, journalists, content creators, fact-checkers, technologists and other members of civil society across Africa, Brazil, Europe, South East Asia and the United States, we have identified three overarching principles that should guide the assessment of the opportunities and risks that generative AI brings to society.[3]

[1] WITNESS https://www.witness.org/

[2] Technology, Threats and Opportunities, WITNESS https://www.gen-ai.witness.org/

[3] For example see: WITNESS, *Deepfakes: Prepare Now (Perspectives from South and Southeast Asia)*. (2020) https://lab.witness.org/asia-deepfakes-prepare-now/ ; Corin Faife, *What We Learned from the Pretoria Deepfakes Workshop (Full Report)*. (2020) *https://blog.witness.org/2020/02/report-pretoria-deepfakes-workshop/ ;* Corin Faife, *How Can U.S. Activists Confront Deepfakes and Disinformation?*. (2020) *https://blog.witness.org/2020/12/usa-activists-disinformation-deepfakes/* ; WITNESS, *Deepfakes: Prepare Now (Perspectives from Brazil)*. (2019) *https://lab.witness.org/brazil-deepfakes-prepare-now/*

1. Center people who are protecting human rights and democracy at the frontlines in the development of solutions

With hyperbolic rhetoric undermining trust in visual media, human rights defenders, journalists and civil society actors collecting trustworthy information or debunking falsehoods will be the most impacted by generative AI. Yet, emerging technologies are designed and deployed without their input, ignoring the threats and risks these technologies bring to communities already at a disadvantage or most affected by mis- and -disinformation. Most importantly, many proposals fail to acknowledge the solutions that those who bear the burden of fighting mis- and disinformation prioritize.[4] While these technologies originate primarily in the United States, their deployment affects people globally. When they are developed, deployed, or regulated without an in-depth understanding of other local and national contexts, the people at the frontlines will face harm. This is why it is crucial that the input of these communities should drive the development and inform the deployment of such technologies.

2. Place firm responsibility on stakeholders across the AI, technology and information pipeline

All actors across the AI pipeline have a duty to insert safeguards and proactively address the harms their work can bring. These include:

- those researching and building foundation models;
- those commercializing synthetic media tools (such as text-to-image or text-to-video tools that allow users to describe an image or video they would like produced and have the AI system generate it);
- those creating synthetic media; and
- those publishing, disseminating or distributing synthetic media (such as media outlets and platforms).

While strong investment in media literacy is crucial, the responsibility cannot be left solely on end-users to determine if the audiovisual content they are viewing is AI-generated and to understand the larger context of the content they are consuming.[5] There are existing frameworks, such as the Partnership on AI's Responsible Practices for Synthetic Media Framework, which offers guidelines for developing, creating, sharing, and publishing synthetic media ethically and responsibly.[6] From its beginning, WITNESS was a part of the process to develop and shape the Framework, which is directed towards those building technology and infrastructure for synthetic media, those creating synthetic media, and those distributing or publishing synthetic media.

3. Embed human rights standards, laws and practices in the development of technical solutions

Legislation, regulation and other norms, as well as company policies and technical infrastructures should have human rights standards baked in. For example, the Coalition for Content Provenance and Authenticity (C2PA), which is developing technical specifications to make it easier to identify how, where and by whom a piece of media may have been created, and the modifications it may have undergone while disseminated, is an example of how human rights standards can inform the develop of technical specifications. As a co-chair to the C2PA Threats and Harms Taskforce,

[4] Sam Gregory, *Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism*. Journalism (December 2021) https://www.researchgate.net/publication/356976532_Deepfakes_misinformation_and_disinformation_and_authenticity_infrastructure_responses_Impacts_on_frontline_witnessing_distant_witnessing_and_civic_journalism

[5] WITNESS, *Synthetic Media, Generative AI And Deepfakes Witness' Recommendations For Action*. https://www.gen-ai.witness.org/wp-content/uploads/2023/06/Guiding-Principles-and-Recs-WITNESS.pdf

[6] Partnership on AI, R*esponsible Practices for Synthetic Media Framework*. https://syntheticmedia.partnershiponai.org/

WITNESS has successfully advocated for globally-driven human rights perspectives and practical experiences to be reflected in the technical standard.[7] In this capacity, we strongly advocate against the standard including personal data that can link generated content to a person's identity. In March 2023, WITNESS highlighted these points in our response to the Office of the United Nations High Commissioner for Human Rights' call for input on the relationship between human rights and technical standard-setting processes for new and emerging digital technologies.[8]

Human rights-informed standards, policies and regulation can help unleash the potential of generative AI and synthetic media while curtailing their misuse and abuse, especially as satire, art and other forms of creative expression test the boundaries of existing legislation and policies.[9]

RECOMMENDATIONS TO ENSURE ACCESS TO VERIFIABLE AND TRUSTWORTHY INFORMATION AND COUNTER AI-GENERATED DISINFORMATION (QUESTIONS 1 & 3)

With the above guiding principles in mind, WITNESS has identified a number of recommendations that can help the US promote a beneficial deployment of generative AI, mitigating potential risks and centering the needs of society at a global level.

Transparency in the production of AI content

In July 2023, seven leading AI companies agreed to a number of voluntary commitments to help move toward safe, secure, and transparent development of AI technology, including committing to earning people's trust by disclosing when content is AI-generated.[10]

WITNESS understands the term *disclosure* to refer to the process of communicating transparently and effectively about image and video synthesis and manipulation. The Partnership on AI's Responsible Practices for Synthetic Media Framework describes direct forms of disclosure as 'visible to the eye', and include methods such as applying a visible label marking the content as AI-generated, adding disclaimers, and watermarking AI-generated content.[11] Indirect forms of disclosure are essentially embedded text, image, or other information in AI-generated digital content. The Framework provides examples of indirect disclosure methods such as applying cryptographic provenance to generated content (e.g.C2PA standard), embedding traceable elements to training data and the content generated, adding metadata which identifies the content as generated, and adding single-frame disclosure statements in videos.[12]

Using direct forms of disclosure, such as labels or watermarks, to signal explicitly to viewers that they are looking at AI-generated content can be a way of ensuring that people understand what they are consuming. However there are

[7] Jacobo Castellanos, *WITNESS and the C2PA Harms and Misuse Assessment Process.* WITNESS (2021) https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/ ; The Coalition for Content Provenance and Authenticity, *C2PA Harms Modelling.* https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html

[8] WITNESS, *Submission to call for input: The relationship between human rights and technical standard-setting processes for new and emerging digital technologies* (2023) https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-csos-witness-4-input.pdf ; The Office of the United Nations High Commissioner for Human Rights, *Human rights and technical standard-setting processes for new and emerging digital technologies.* (June 2023) https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/sessions-regular/session53/advance-versions/A_HRC_53_42_AdvanceUneditedVersion.docx

[9] WITNESS, *Report: Just Joking! Deepfakes, Satire and the Politics of Synthetic Media.* (2022) https://cocreationstudio.mit.edu/just-joking/ ; Also see: shirin anlen and Raquel Vazquez Llorente, Using Generative AI for Human Rights Advocacy. (June 2023) https://blog.witness.org/2023/06/using-generative-ai-for-human-rights-advocacy/

[10] The White House, *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.* (July 2023) https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[11] Partnership on AI, *Responsible Practices for Synthetic Media Framework.* https://syntheticmedia.partnershiponai.org/; Also see: Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/

[12] Partnership on AI, *PAI's Responsible Practices for Synthetic Media.* https://syntheticmedia.partnershiponai.org/#read_the_framework

limitations to this approach.[13] For example, visible labels or watermarks tend to be small and easily missed, and there is not necessarily always space to provide meaningful context on how the media was created or why the piece of media was generated. Further, it has been shown that when any piece of media, even labeled and watermarked, is distributed across politicized and closed social media groups, its creators lose control of how it is framed, interpreted, and shared.[14] It is also important that watermarks do not become equated with the idea that the content is untrustworthy - or that the absence of watermarks means the content is trustworthy. Watermarks, as well as provenance infrastructures, should be understood as signals of content's source and how it has changed over time, but not as an absolute confirmation of this information.

Crucially, disclosure technologies and tools should not include information that could reveal a person's identity. Since 2019, WITNESS has been raising concerns that technologies such as watermarks should not include information about the identity of the person who created it.[15] People using these tools to create imagery should not need to forfeit their right to privacy to adopt these technologies. As such, these solutions can be powerful approaches to effectively communicate when an image or video has been AI-generated, but they can also result in harm, especially for those people and communities that are already at risk. For instance, governments could require provenance schemes that capture personally identifiable information to augment surveillance and stifle freedom of expression.[16]

Tracking media provenance with authentication technologies and standards

Initiatives that track the origin and alterations of a piece of media can help journalists, activists, human rights defenders and others to ensure societies are able to ascertain how a piece of content was created and if and how it has been modified. However, like watermarks, these technologies can also lead to potential harms to a broad range of individuals and communities, especially those at the frontlines of defending human rights.[17]

To ensure that authenticity and provenance frameworks are developed in line with global human rights laws and standards, they should not include the collection of information that could reveal a person's identity.[18] In addition, the technologies should allow people to choose to opt-in or out of using them. The tools should also be built with accessibility in mind and in a way that allows all levels of technical knowhow to opt-in or out and have their identities protected.[19]

Detection of AI-generated or manipulated content

Detection tools may allow people to run a piece of content through it and receive information about how likely it is that the content had been generated or edited by an AI system. As such, these tools can play an important role in the beneficial deployment of generative AI, and in mitigating risks.

[13] Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/ ; Sue Halpern, *Will Biden's Meetings with A.I. Companies Make Any Difference?* The New Yorker (July 2023) https://www.newyorker.com/news/daily-comment/will-bidens-meetings-with-ai-companies-make-any-difference

[14] Sam Gregory, *Ticks Or It Didn't Happen*. WITNESS (December 2019) https://lab.witness.org/ticks-or-it-didnt-happen/

[15] Ibid.

[16] List of potential harms of the C2PA specifications: https://c2pa.org/specifications/specifications/1.0/security/_attachments/Due_Diligence_Actions.pdf Also see: Gabrielle Lim and Samantha Bradshaw, *Chilling Legislation: Tracking the Impact of "Fake News" Laws on Press Freedom Internationally*. Center for International Media Assistance (July 2023) https://www.cima.ned.org/publication/chilling-legislation/

[17] Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all*. WITNESS (2020) https://blog.witness.org/2020/05/authenticity-infrastructure/

[18] Raquel Vazquez Llorente, *Trusting Video in the Age of Generative* AI. Commonplace (June 2023) https://commonplace.knowledgefutures.org/pub/9q6dd6lg/release/2

[19] Sam Gregory, *Ticks Or It Didn't Happen*. WITNESS (December 2019) https://lab.witness.org/ticks-or-it-didnt-happen/

However, existing detection tools require expert input to assess the results and are not generalizable across multiple synthesis technologies and techniques. Detection tools also need to be trained on data related to the scenarios in which they are deployed. It's important to note that detection tools can also lead to unintentioned confusion. For example, in a number of global cases, the use by the general public of detection tools available online has contributed to increased doubt around real footage rather than contributing to clarity.[20]

In addition to ensuring the tools are available and usable for those who need them most, further research into improving detection capabilities should be supported. WITNESS is currently piloting a Deepfakes Rapid Response Force that allows members from the International Fact-Checking Network to escalate cases of suspected deepfakes, and get a timely assessment on the authenticity or origin of the content.[21]

Access and investment in tools that help verify content online

Most of the cases brought to our Deepfakes Rapid Response Force were not escalated due the content being more simple mis-contextualized or unsophisticated manipulations, rather than examples of technically complex media. This is one of the reasons that WITNESS advocates for platforms and messaging apps to support further research, development, and deployment of more accessible tools that can explain and contextualize 'shallowfakes' – or mis-contextualized, mis-attributed, or edited images and video.

Platforms should also invest in the implementation of platform-level intuitive reverse image search and more investment in similar capabilities for video. There is also a need for the development of cross-platform reverse image and video search approaches, which would unlock the ability to search for audiovisual content across a range of platforms simultaneously. At present, there are no widely available reverse video search tools and existing reverse image search tools are largely oriented towards commercial uses such as online shopping or protecting copyright, rather than curbing disinformation.[22] Accessible reverse video search would allow people to, in effect, simply click a button and conduct a search to see where a video was originally posted and how it has been shared or edited over time. Accessible reverse video search tools would allow researchers to better train detection tools and also allow a less technical audience to benefit from the tools.[23] Although AI-generated media is an emerging technology, the threats it poses are not new - in WITNESS' years of organizing global workshops, a primary concern that has arisen repeatedly is the mis-contextualization, mis-attribution, or editing of video and audio on social media platforms.

Finally, in the process of identifying and promoting the beneficial deployment of generative AI while also mitigating risks, lawmakers have the opportunity to ensure that platform policies, regulations, and laws on content moderation (particularly around satire) incorporate internationally recognized human rights standards for freedom of expression. WITNESS provides specific recommendations for this in our Just Joking report.[24]

[20] Sam Gregory, *The World Needs Deepfake Experts to Stem This Chaos*. WIRED (June 2021) https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/

[21] Nilesh Christopher, *An Indian politician says scandalous audio clips are AI deepfakes. We had them tested.* Rest of World (July 2023) https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/

[22] See for example Google Lens: https://lens.google/

[23] Sam Gregory, *Shallowfakes are rampant: Tools to spot them must be equally accessible*. The Hill (August 2022) https://thehill.com/opinion/technology/3616877-shallowfakes-are-rampant-tools-to-spot-them-must-be-equally-accessible/

[24] WITNESS, *Report: Just Joking! Deepfakes, Satire and the Politics of Synthetic Media.* (2022) https://cocreationstudio.mit.edu/just-joking/