US Office of Science and Technology Policy Request For Information: National Priorities for Artificial Intelligence

Submitted on: 7 July 2023

Submitted by: WITNESS

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

INTRODUCTION

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.

This submission focuses on WITNESS' recommendations to ensure that global human rights laws and standards are baked into the design, development and deployment of generative AI into societies across the globe. Our submission addresses questions 1, 11, and 16 and is informed by WITNESS' 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.

[1] WITNESS https://www.witness.org/
[2] Technology, Threats and Opportunities, *WITNESS* http://witnessgenai.global/

OVERARCHING PRINCIPLES FOR THE DEVELOPMENT OF AI STRATEGY

We have identified three overarching principles that should guide the development of countries' national AI strategies, including the United States' National Artificial Intelligence Strategy, to ensure that countries are able to benefit and mitigate the risks of AI-generated audiovisual content online (questions 1, 11, and 16).

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

Human rights defenders, journalists and civil society actors ensuring information is trustworthy will be the most impacted by synthetic media and generative AI, especially when hyperbolic rhetoric undermines trust in visual media. Since 2018, WITNESS has led regular global consultations with leading human rights defenders, journalists, fact checkers, content creators and others, across Africa, Brazil, Europe, South East Asia, the United States.[3] These consultations have focused on how deepfake technology may impact on different communities.[4] In 2022 and 2023, we have also focused our consultations on understanding the threats and opportunities of generative AI systems.

Emerging technologies including generative AI are being created without the input of those who will be impacted the most as the technologies are deployed. Communities such as human rights defenders and journalists are at the frontlines of democracy and human rights. When powerful technologies are developed without an in-depth understanding of local and national contexts, 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.

[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/*

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

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

People and organizations working across the AI pipeline all have a duty to insert safeguards and address the harms their work can bring. This includes:

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

It is a recipe for failure if the responsibility is left solely on end-users to determine if the audiovisual content they are viewing is AI-generated and the larger context of the content they are consuming.

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

Any proposed solution should ensure that these are designed and deployed with human rights standards baked in. This should range from technical infrastructures, norms, and platform policies, to laws and regulations. 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 – whether it is by media outlets or on social media feeds, assessed their specifications to understand the potential harm that could come from it. They also proposed and developed strategies and actions that could avert and mitigate those harms.[5] 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

[5] The Coalition for Content Provenance and Authenticity, *Harm assessment of the C2PA Technical specifications*. https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html

creative expression test the boundaries of these norms.[6] Norms and standards for generative AI, and in particular for synthetic media, which opens up a range of creative possibilities for satire and art, should be shaped from a global human rights perspective and should include particular attention to satire and other forms of expression.

With these three guiding principles in mind, WITNESS has a number of recommendations that can help to ensure that AI systems are designed, developed, and deployed in a manner that protects the rights of people globally.

Firstly, to ensure that a country's AI strategy advances democratic values and does not perpetuate further harms on global populations that have been historically underserved (question 11), **technology companies, legislators, and policymakers must identify risks and create threat models that include the perspectives and expertise of these communities. Proposed solutions should center these perspectives and draw from the lived, practical, and expert experience of vulnerable and excluded people who have faced similar harms.** The well-documented harms that have come from companies' decisions to allow unequal or negligent policy enforcement in certain countries demonstrates the importance of including the expertise of a broader range of context-specific stakeholders in the development and deployment of emerging technologies.[7]

There are a range of existing solutions that have been developed which WITNESS would like to highlight in our submission. As tools that allow people to better understand the source and context of the content they consume continue to be developed and deployed, they **must be shaped by human rights and accessibility concerns**.

Below are some examples of these tools, some of the identified human rights and accessibility concerns, and further recommendations.

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

[7] *Rohingya sue Facebook for £150bn over Myanmar genocide*, The Guardian. (December 2021) https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence

DISCLOSURE TECHNIQUES AND TECHNOLOGIES

We use the term *disclosure* to refer to the process of communicating transparently and effectively about image and video synthesis and manipulation. There are direct forms of disclosure that are 'visible to the eye', which include methods such as applying a visible label marking the content as AI-generated, adding disclaimers, and watermarking AI-generated content.[8] There are also indirect forms of disclosure (which are behind the scenes and invisible to most people) such as applying cryptographic provenance to generated content (such as the Coalition for Content Provenance and Authenticity (C2PA) standard), applying traceable elements to training data and the content AI systems generate, adding metadata which identifies the content as generated, and adding single-frame disclosure statements in videos.[9]

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 limitations to this approach.[10] 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. Further, simple text-based labels or watermarks can create the additional misconception that anything that does not have a label is not manipulated, when in reality, that may not be true, as research on provenance data has also shown.[11]

One way to mediate these limitations is to support research into how artistic creativity can be used to disclose to the public the nature of the content they are consuming For example, in David France's documentary *Welcome to Chechnya*,[12] interviewees at risk of persecution were

[8] Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/

[9] For more on the Coalition for Content Provenance and Authenticity https://c2pa.org/ ; Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all* (2020) https://blog.witness.org/2020/05/authenticity-infrastructure/ ; Also see: Partnership on AI, *PAI's Responsible Practices for Synthetic Media*. https://syntheticmedia.partnershiponai.org/#read_the_framework

[10] Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/

[11] Gabi Ivens, Sam Gregory, *Ticks or it Didn't Happen (2019) https://lab.witness.org/ticks-or-it-didnt-happen/*

[12] David France, *Welcome to Chechny*a (2020) https://www.welcometochechnya.com/

digitally disguised with the help of inventive synthetic media tools like those used to create deepfakes, while signaling to the audience the use of such techniques with the use of subtle visual cues.[13]

Companies are increasingly incentivized to apply indirect labels to generated content because there is emerging evidence that when AI models are trained on AI-generated content (as opposed to human-generated content) the models become increasingly useless.[14]

It is important that in the process of developing a national AI strategy, there is **broad support for further research into how to communicate transparently and effectively to the public both direct and indirect markers on AI generated images and videos, as well as how to standardize, regulate, research and implement these approaches**. Crucially, disclosure technologies and tools should not include information that could reveal a person's identity.

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. This highlights that, as generative AI technologies continue to evolve, so too should disclosure solutions and also highlights how crucial it is to include vulnerable and excluded people in the creation of such solutions.

AUTHENTICITY AND PROVENANCE TECHNOLOGIES

For years, organizations and individuals have been researching and building technology that helps people capture photos and video in a way that renders the content able to be easily used as evidence in litigation.

The development of pioneering technologies like Proofmode and eyeWitness began over ten years ago by the human rights sector, offering options to track the provenance of a piece of media and help prove its integrity.[15] While these tools were intended for specialized and niche uses, more recent and ambitious projects are underway by the private sector. For example the

[13] Carolyn Giardina, *Academy Reveals 2023 SciTech Award Recipients*. The Hollywood Reporter (February 2023) https://www.hollywoodreporter.com/movies/movie-news/2023-scientific-technical-academy-award-winners-1235316245/

[14] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, Ross Anderson, *The Curse of Recursion: Training on Generated Data Makes Models Forget*. arXiv (May 2023) https://arxiv.org/abs/2305.17493

[15] See for example: Proofmode: https://guardianproject.info/apps/org.witness.proofmode/ and Eyewitness to Atrocities: https://www.eyewitness.global/

C2PA has the potential to provide broader transparency into the history of an image or video. This means that it would become easier to see the original source of an image or video, and how it has been shared and changed over time.

WITNESS has successfully advocated for globally-driven human rights perspectives and practical experiences to be reflected in the C2PA, where we co-chair the Threats and Harms Taskforce to assess the technical specifications for their potential to be misused and cause harm.[16]

Initiatives like these can help journalists, activists, human rights defenders and others protect their own work and ensure societies are able to ascertain the source of a particular piece of media and if it was modified or not. However, these technologies can also lead to potential harms to a broad range of individuals and communities, especially those who are already most at risk. For instance, governments could require provenance schemes that capture personally identifiable information to augment surveillance and stifle freedom of expression.[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]

It's also important to ensure that authenticity and provenance architectures are seen as a signal about a piece of content's source and how it has changed over time, but not as an absolute confirmation of this information. Research has shown that when warning labels are added to news stories that have been challenged by independent fact-checkers, it can cause people to believe that when there is no such warning tag, the news stories must be true, even if the story simply hadn't been tagged.[19] This is known as the 'implied truth/implied falsehood' effect and reiterates the need for people to see these solutions in the broader context.

[16] Jacobo Castellanos, *WITNESS and the C2PA Harms and Misuse Assessment Process* (2021) https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/
[17] List of potential harms of the C2PA specifications: https://c2pa.org/specifications/specifications/1.0/security/_attachments/Due_Diligence_Actions.pdf
[18] Raquel Vazquez Llorente, *Trusting Video in the Age of Generative* AI. Commonplace (June 2023) https://commonplace.knowledgefutures.org/pub/9q6dd6lg/release/2
[19] Gordon Pennycook, Adam Bear, Evan Collins, *The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings*. Management Science (August 2019) https://www.researchgate.net/publication/321887941_The_Implied_Truth_Effect_Attaching_Warnings_to_a_Subset_of_Fake_News_Headlines_Increases_Perceived_Accuracy_of_Headlines_Without_Warnings

AI-GENERATED CONTENT DETECTION TOOLS

Another way that people can understand if they are viewing AI-generated content is by using detection tools. In the long-term, these tools could allow people to run a piece of content through the tool and receive information about how likely it is that the content had been generated or edited by an AI system. As such, these tools could play an important role in the broader solution and help to ensure that people are able to understand the context of the content they are consuming. However, existing detection tools tend not to be reliable at scale, and require expert input to assess the results. For example, in a number of global cases, the use by the general public of detection tools available online has contributed to confusion and increased doubt around real footage rather than contributing to clarity.[20]

Even so, **access to current detection tools should be given to those who need them most, such as journalists and fact-checkers**, as they look to debunk realistic forgeries or dismiss claims that genuine journalistic audiovisual content is fake. Equity in access to detection tools and capacities is critical to ensure that civil society and media can have tools designed with their needs in mind, as well as the relevant skills needed to use them.

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 Deepfake Rapid Response Force that allows International Fact-Checking Network members to escalate cases of suspected deepfakes, and get a timely assessment on the authenticity or origin of the content.[21]

TOOLS THAT HELP VERIFY CONTENT ONLINE

However, in our experience many of the cases brought to the Force were not escalated due the content being mis-contextualized or unsophisticated manipulations, rather than sophisticated deepfakes. This is one of the reasons that WITNESS advocates for companies and other stakeholders to invest in media forensics and detection capacity, for instance by pushing for

[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] See the International Fact-Checking Network here: https://www.poynter.org/ifcn/

more accessible reverse video search capabilities.

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.[22]

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.[23]

Platforms should therefore **invest in the creation of platform-level intuitive reverse image and video search and other forms of easily providing context to visual content**. 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.

Platforms and messaging apps **should also 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**.

POLICIES THAT RESPECT FREEDOM OF EXPRESSION

In shaping a national-level AI strategy, 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**

[22] 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/

[23] 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/*

freedom of expression. WITNESS provides specific recommendations for this in our Just Joking report and these are briefly summarized below.[24]

- **Legality**: Any proposed legislation or policies that restrict or burden expression must not, among other things, be vague or overly broad. In the case of deepfakes and other forms of AI-generated content, this includes defining clearly what specific forms of content and their uses are being restricted. This practice gives people clear guidelines, limits the discretion of those implementing the policy (which helps to avoid selective and discriminatory enforcement), and avoids restricting practices that do not pose risks of harm.
- **Legitimacy**: The reason for limiting expression must be a legitimate public interest objective, as set forth in International Covenant on Civil and Political Rights Article 19, such as protecting the rights of others, national security, or public health. Protecting a regime, head of state, or government official would not constitute legitimate grounds for limiting expression.
- **Necessity and proportionality**: This test should be applied using interdisciplinary, multi-stakeholder input to determine when it is truly necessary to limit the use of AI-generated content. Specific questions to ask that are helpful in determining if limiting expression is necessary and proportionate are available in WITNESS' report Just Joking.[25]

There is ongoing work to assess that existing authenticity and provenance infrastructures are being built and deployed in line with global human rights.[26] Principles that shape these assessments could also be useful in ensuring freedom of expression is maintained as countries formulate their national AI strategies. Examples of these policies include ensuring that solutions are built with a critical eye toward potential abuse and misuse of the framework, as well as reviewed for the ability to be abused and cause unintended harms, threats to human rights, or disproportionate risks to vulnerable groups globally. In March 2023, WITNESS highlighted these points in our response to the Office of the United Nations High Commissioner for Human

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

[25] WITNESS, *Report: Just Joking! Deepfakes, Satire and the Politics of Synthetic Media.* (2022) https://cocreationstudio.mit.edu/just-joking/#part-3

[26] Coalition for Content Provenance and Authenticity, *C2PA Specifications.* https://c2pa.org/specifications/specifications/1.0/specs/C2PA_Specification.html#_information_security

Rights' call for input on the relationship between human rights and technical standard-setting processes for new and emerging digital technologies.[27]

EFFORTS THAT ADDRESS ONLINE GENDER-BASED VIOLENCE

In developing a national AI strategy one of the priorities should be **to support national and transnational efforts to address online gender-based violence and harassment, such as technical and legislative responses to counter the use of non-consensual sexual deepfakes**. During WITNESS' regional workshops another theme that is brought up repeatedly, has been the disproportionate harms that generated AI audiovisual content, including deepfakes, has on women and the LGBTQI community.[28]

According to a UNESCO study, "15% of women have reported experiencing image-based abuse (eg: manipulated photos or video, stolen images, explicit images shared publicly without permission), while 4% said they have been victims of 'deepfakes' (ie manipulated videos, often associated with fake porn, designed to damage reputations), and 4% reported 'shallowfakes' (ie decontextualized videos or images, such as the misrepresentation of a crime scene) as a technique used to target them".[29] In May 2023, the US Government released its National Plan to End Gender-Based Violence: Strategies for Action which includes addressing technology-facilitated gender-based violence.[30] The US' National Artificial Intelligence Strategy should ensure that these strategies overlap in their proposed solutions to preventing, addressing, and remediating the use of generative AI technologies to perpetrate gender-based violence online.

[27] 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

[28] See for example: Raquel Vazquez Llorente, Jacobo Castellanos, and Nkem Agunwa, *Fortifying the Truth in the Age of Synthetic Media and Generative AI*. WITNESS (June 2023) https://blog.witness.org/2023/05/generative-ai-africa/ ; WITNESS, *Deepfakes: Prepare Now (Perspectives from South and Southeast Asia). (*2020) https://lab.witness.org/asia-deepfakes-prepare-now/ ; WITNESS, *Deepfakes: Prepare Now (Perspectives from Brazil).* (2019) *https://lab.witness.org/brazil-deepfakes-prepare-now/*

[29] UNESCO, *Online Violence Against Women Journalists: A Global Snapshot of Incidence and impacts.* (2021) https://unesdoc.unesco.org/ark:/48223/pf0000375136

[30] The White House, *Release of the National Plan to End Gender-Based Violence: Strategies for Action.* (May 2023) https://www.whitehouse.gov/gpc/briefing-room/2023/05/25/release-of-the-national-plan-to-end-gender-based-violence-strategies-for-action/

INVEST IN MEDIA LITERACY

To ensure that people are able to understand the opportunities and potential risks of AI systems (question 16), there is a strong need for media literacy. **Media and digital literacy alone are not enough but they remain as necessary as ever.**

While there should be responsibility throughout the AI development pipeline and identifying AI-generated content and its larger context should not be the sole responsibility of content consumers, media literacy remains critically important. As such, governments need to strongly invest in and support media literacy campaigns that inform the public about what synthetic media is, and what is possible with new forms of multimedia manipulation (and what it is not). These initiatives can help prepare the public to view and consume media more critically while not adding to the rhetoric around generative AI. They should acknowledge the breakneck speed at which generative AI technologies are being developed and deployed, but balance this by providing grounded, transparent, and realistic information about the current state of AI.

Moreover, media literacy should also be a vehicle for empowering individuals and communities to engage with governments, civil societies and companies to develop responses and solutions that reflect their needs, circumstances and aspirations. In this regard, media literacy campaigns acquire a critical importance and are a precursor to effective and inclusive public policy making.

This should include **empowering, connecting, training and resourcing a diverse and global range of key frontline actors, like local media, community leaders,civil liberties groups, and non-partisan election officials to better understand and prepare for the threats and opportunities in generative AI**.

These campaigns should include a particular focus on human rights defenders and journalists. As technology evolves, these communities must be supported to adjust to the changes. There should be attention to improving journalist and human rights reporting and advocacy practices to respond to a changed reality and ensure that these communities are able to continue to meaningfully contribute to the protection of democracy and human rights.

PROMOTE ETHICAL STANDARDS ON USING AI-GENERATED CONTENT

Finally, a national AI strategy should support and promote context-specific ethical standards that relate to the use of AI-generated audiovisual content.

For example, WITNESS has produced a series of recommendations for human rights and civil society organizations to keep in mind when deciding if and how to use AI-generated content in their campaigning.[31]

The use of AI to create or edit media should not undermine the credibility, safety and content of other human rights organizations, journalists, fact checkers and documentation groups.

In the global consultations WITNESS has been leading since 2018, we have consistently heard from communities at the frontlines of human rights defense about how their content and credibility are constantly undermined.[32] The advent of synthetic media has made it easier to dismiss real footage. As tools to make AI-generated or edited images, videos and audio become more accessible, it will only be easier for governments and companies to claim that damaging footage is fake. When generating or modifying visual content with AI, it is important to think about the role of global human rights organizations in terms of setting standards and using tools in a way that doesn't have collateral harms to smaller, local groups who face much more extreme pressures. These groups are already overburdened trying to defend their footage or challenge false information and are targeted repeatedly by their governments to discredit them.

[31] 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/

[32] Eromo Egbejule, *Panel of inquiry finds Nigerian army culpable in Lekki 'massacre'*. Al Jazeera (November 2021) https://www.aljazeera.com/news/2021/11/16/panel-of-inquiry-finds-nigerian-army-culpable-in-lekki-massacre

AI output should be clearly labeled and watermarked, and consider including metadata or invisible fingerprints to track the provenance of media.

When publishing content that is generated or manipulated using AI, its use should always be disclosed. For disclaimers or cues to be effective, they need to be legible, meaning they can be seen, read, heard, or understood by those consuming the information. WITNESS strongly advocates for a more innovative and principled approach to content labeling that can express complex ideas and provide audiences with meaningful context on how the media has been created or manipulated.[33] In the best cases, these transparency approaches should also communicate the intent behind the use of AI.

A careful approach to consent is critical in AI audiovisual content.

There are a few exceptions that can be addressed by asking who the represented subjects are, what groups they belong to, and what the intent of the content is. For instance, political satire may not require consent by the subject as it aims to challenge power dynamics of public figures, or criticize and highlight the absurdities of systems that reinforce inequality. Human rights organizations can draw from existing guidelines[34] about informed consent in visual content, as well as good practices in dealing with situations or populations that require special attention–for instance, footage involving minors, people with mental disabilities, people under coercive contexts, or perpetrators of abuse.[35]

[33] Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/

[34] WITNESS, *Announcing WITNESS' Ethical Guidelines for Using Eyewitness Footage in Human Rights*. https://lab.witness.org/announcing-witness-ethical-guidelines-for-using-eyewitness-footage-in-human-rights/

[35] WITNESS, *Obtaining Informed Consent*. https://library.witness.org/product/obtaining-informed-consent/

Expectations about the veracity and the extent of manipulation in the visual content will depend on the context in which it is produced and the genre of the footage.

For instance, narrative and fiction films on human rights may lend themselves to more artistic expressions that can help people connect with difficult topics. In contrast, factual human rights reporting that aims at exposing abuses has implicit assumptions of accuracy, veracity and realism. Some of these expectations may also evolve over time, particularly when footage is revisited or reclaimed for purposes other than the one for which they were created. This is why transparency in the creation and modification of visual content, as well as a careful approach to consent, are both critical to avoid contributing to mis- and disinformation and harming communities and individuals.

WITNESS has also detailed potential use cases of AI-generated content and the ethical considerations for each use case.[36] These uses include for identity protection, visualizing testimonies of survivors, victims, and witnesses, or reconstructing places and events from statements, and for satire and artistic expression.

[36] 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/