AI Accountability Policy Request for Comment: National Telecommunications and Information Administration

Submitted: 12 June 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 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 how AI accountability mechanisms can inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI (question 5)–in particular in relation to synthetic videos, image and audio. This document 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] See: WITNESS https://www.witness.org/
[2] See: Technology, Threats and Opportunities, WITNESS http://witnessgenai.global/

In response to question 5:

Guiding principles for developing AI accountability mechanisms

1. **Center people who are protecting human rights and democracy at the frontlines in the development of AI accountability mechanisms**

   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. Vulnerable communities and others facing similar harms already have experience confronting such harms, and should be included in the solutions.

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

   People and organizations working on foundational models, tech builders, companies commercializing tools, tech deployers, content creators and content distributors (both traditional media and social media), all have a duty to insert safeguards and address the harms their work can bring.

3. **Embed human rights standards, laws and practices in the development of AI accountability mechanisms**

   Proposed AI accountability mechanisms should ensure that they are designed and deployed with human rights standards baked in. This should include particular attention to satire and other forms of expression. Synthetic media opens up creative possibilities for satire and art, but there are no clear cut lines for discerning genuinely satirical or artistic content from harmful or malicious content.

Recommendations to ensure that AI accountability mechanisms inform people about how such tools are operating and whether the tools comply with standards for trustworthy AI

1. **Support further research into how to communicate transparently and effectively image and video synthesis and manipulation, as well as how to**

standardize, regulate and implement transparency

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', such as labeling and some watermarks,[3] and indirect forms of disclosure such as fingerprints, single-frame warnings or provenance data.[4]

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.[5] [6] While these tools were intended for specialized and niche uses, more recent and ambitious projects are underway by the private sector. One of these initiatives is 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.[7]

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.[8]

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.[9]

In addition to solutions that track the provenance of content, human rights and

---

[3] See for example: Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/
[4] See for example: Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all* (2020) https://blog.witness.org/2020/05/authenticity-infrastructure/
[5] Proofmode is an application for signed visual evidence. See more here: https://guardianproject.info/apps/org.witness.proofmode/
[6] See: Eyewitness to Atrocities: https://www.eyewitness.global/
[7] See Coalition for Content Provenance and Authenticity https://c2pa.org/
[8] See: 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/
[9] See: List of potential harms of the C2PA specifications: https://c2pa.org/specifications/specifications/1.0/security/_attachments/Due_Diligence_Actions.pdf

accessibility concerns should also shape the design and implementation of visible indicators of synthetic media, such as watermarks or labels.

Signaling explicitly to viewers that they are looking at synthetic media can be a way of ensuring that people understand when they are consuming AI generated or manipulated content. However there are limitations to this approach.[10] For example, visible labels tend to be small and easily missed, and there isn't 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 can create the additional misconception that anything that doesn't 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 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]

These solutions outlined in the paragraphs above can be powerful approaches to effectively communicate when an image or video has been synthetically created or manipulated and how, but they can also result in harm, especially for those people and communities that are already at risk.

As such, technology companies, legislators, and policy makers should put at the center the perspectives and experience of vulnerable and excluded people who are facing similar harms across the globe when working on identifying risks, creating threat models, and prioritizing solutions As image and video synthesis and manipulation technologies continue to evolve, so too should disclosure solutions.

---

[10] 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/

[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/

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

2. **Mandate proactive human rights assessments and ground accountability mechanisms in real world needs**

   As a baseline, accountability mechanisms should mandate that companies and others developing AI systems must undertake comprehensive human rights assessments prior to deploying AI models and tools, particularly those involved in the creation or manipulation of synthetic images, videos, and audio ("synthetic media"). This requirement has already gained recognition from leading companies and individuals in this field, as well as in emerging regulations and soft law. However, it is important to emphasize that the establishment of threat-mitigating mechanisms such as proactive human rights assessments are a piece of the puzzle and not a lone solution.