WITNESS
SEE IT FILM IT CHANGE IT

# Testimony of Sam Gregory, Executive Director, WITNESS

## Before the U.S. Senate Committee on Commerce, Science and Transportation

## Subcommittee on Consumer Protection, Product Safety and Data Security

## Date of hearing: September 12, 2023

## “The Need for Transparency in Artificial Intelligence”

Chairman Hickenlooper, Ranking Member Blackburn and members of the Senate Commerce Subcommittee on Consumer Protection, Product Safety and Data Security, thank you for the opportunity to testify today about transparency in AI.

I am Sam Gregory, Executive Director of WITNESS, a human rights organization.[1] 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 consumers, citizens and communities in the US and globally, and to prepare accordingly.[2] These efforts have included contribution to technical standards development,[3] engagement on detection and authenticity approaches that can support consumer literacy,[4] analysis and real-time response to contemporary usages,[5] research,[6] and consultative work with rights defenders, journalists, content creators, technologists and other members of civil society.[7] Our experience is further informed by three decades of experience helping communities, citizens, journalists and human rights defenders create trustworthy photos and videos related to critical societal issues and protect themselves against the misuse of their content.

Today, I will have a particular focus on how to optimize the benefits, and minimize the harms and risks from multimodal audiovisual generative AI. These tools, with their potential to create realistic image, audio and video simulations at scale, as well as personalized content, will have far-reaching implications for consumers, creative production and generally, our trust in the information we see and hear.

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

[2] For our work on generative AI and deepfakes see: https://www.gen-ai.witness.org/

[3] Jacobo Castellanos, *WITNESS and the C2PA Harms and Misuse Assessment Process,* WITNESS, December 2021, *https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/*

[4] WITNESS Media Lab, *How do we work together to detect AI-generated media?* https://lab.witness.org/projects/osint-digital-forensics/

[5] 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/*

[6] Gabriela Ivens and Sam Gregory, *Ticks or It Didn't Happen: Confronting Key Dilemmas in Authenticity Infrastructure for Multimedia*, WITNESS, December 2019, https://lab.witness.org/ticks-or-it-didnt-happen/

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

## Executive Summary

My organization, WITNESS, has for a number of years promoted a perspective of *Prepare, Don't Panic* in relation to deepfakes and generative AI. But the moment to act and prepare society for audiovisual generative AI and its impacts has come.

Transparency around AI's role in the production of information that consumers and citizens engage with is a critical area. In this testimony, I will focus on questions of disclosure and provenance in audiovisual content, and how these relate to the responsibility of actors in the AI pipeline. WITNESS, in consultation with global experts and communities affected by technology development, focuses on three overarching principles to guide the assessment of the opportunities and risks that generative AI brings to society.

1. Place firm responsibility on stakeholders across the AI, technology and information pipeline.
2. Center those who are protecting human rights and democracy, and communities most impacted by AI, domestically and globally, in the development of solutions.
3. Embed human rights standards and a rights-based approach in the response to AI.

With these principles in mind, US policy makers and legislators have a range of options to promote transparency in AI and protect consumers and their data:

1. Ensure broad consultations with communities impacted by AI when developing solutions to watermarking, provenance and disclosure, and in broader processes of transparency common to all AI systems–including documentation, third-party auditing, pre-release testing, evaluation, and human rights impact monitoring.
2. Push for a rights-based approach to transparency in AI, that promotes standardized systems for disclosing when content has been made with AI, while supporting opt-in solutions that track the provenance of non-synthetic media.
3. Prohibit personal data from being included by default in any approaches to provenance, disclosure and watermarking, for all types of content.
4. Enact comprehensive data privacy legislation, as well as integrate data privacy rights into broader AI legislation that also includes solutions for opting out of models' datasets.
5. Support research and investment in technologies that can detect AI manipulation and generation and are accessible domestically and globally, as well as consumer-facing tools that provide information on content provenance.

### The domestic and global context of audiovisual generative AI and deepfakes

While there are creative and commercial benefits to generative AI and synthetic media, these tools are also already connected to a range of harms to US consumers and global users. Chatbots provide incorrect, factual-appearing information. Audio scams using simulated audio are proliferating. Non-consensual sexual images are used to target private citizens and public figures, particularly women, and AI-generated child sexual abuse images are increasing. Claims of AI-generation are used to dismiss verifiable content. Text-to-image tools perpetuate existing patterns of bias or discriminatory representation present in their training data. Creatives and artists have had their production incorporated into training for AI models without consent, and no-one has access to reliable ways to opt their images out of these training data sets.[8]

In the area that I focus on, audiovisual generative AI and deepfakes, research indicates that humans do trust the realism cues of audio and video,[9] cannot identify machine-generated speech cloning accurately,[10] do not recognize simulated human faces,[11] do not fare well spotting face-swapped faces,[12] and retain false memories of deepfakes.[13] In the direct experience of my own organization in analyzing high-profile suspected deepfakes encountered globally, it is challenging to support rapid, high-quality media forensics analysis; detection resources are not widely available to the media or the public; and the gap between analysis and timely public understanding is wide and can be easily exploited by malicious actors.[14]

[8] Rhiannan Williams, Melissa Heikkilä, You need to talk to your kid about AI. Here are 6 things you should say, MIT Tech Review, September 2023, https://www.technologyreview.com/2023/09/05/1079009/you-need-to-talk-to-your-kid-about-ai-here-are-6-things-you-should-say/; Matt O'Brien, Chatbots sometimes make things up. Is AI's hallucination problem fixable?, AP, August 2023, https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4 ; FTC Consumer Alert, Scammers use AI to enhance their family emergency schemes, March 2023, https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes; Benj Edwards, *AI-generated child sex imagery has every US attorney general calling for action*, Ars Technica, September 2023, https://arstechnica.com/information-technology/2023/09/ai-generated-child-sex-imagery-has-every-us-attorney-general-calling-for-action/; Nilesh Christopher, *ibid*; Rida Qadri, Renee Shelby, Cynthia L. Bennett and Emily Denoton, *AI's Regimes of Representation: A Community-Centered Study of Text-to-Image Models in South Asia*, 2023, https://arxiv.org/abs/2305.11844 ; Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru, *AI Art and its Impact on Artists*, August 2023, https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604681;

[9] Steven J, Frenda, Eric D. Knowles, William Saletan, Elizabeth F Loftus, *False memories of fabricated political events*, Journal of Experimental Social Psychology, 2013,https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2201941

[10] Hibaq Farah, *Humans can detect deepfake speech only 73% of the time, study finds,* The Guardian, August 2023, https://www.theguardian.com/technology/2023/aug/02/humans-can-detect-deepfake-speech-only-73-of-the-time-study-finds

[11] Sophie J. Nightingale and Hany Farid, *AI-synthesized faces are indistinguishable from real faces and more trustworthy,* PNAS, February 2022, https://www.pnas.org/doi/10.1073/pnas.2120481119

[12] Nils C. Köbis, Barbora Doležalová and Ivan Soraperra, *Fooled twice: People cannot detect deepfakes but think they can,* IScience, November 2021, https://www.sciencedirect.com/science/article/pii/S2589004221013353

[13] Nadine Liv, Dov Greenbaum, *Deep Fakes and Memory Malleability: False Memories in the Service of Fake News,* AJOB Neuroscience, March 2020, https://www.tandfonline.com/doi/abs/10.1080/21507740.2020.1740351?journalCode=uabn20

[14] Sam Gregory*, 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/ ;* Nilesh Christopher, *ibid.*

These AI tools create *accidental harms* when they don't work as promised or anticipated including when they 'hallucinate' information. Deceptive information is a feature, not a bug of systems. Consumers also face *misuse harms when* generative AI tools work as intended to, but are exploited deliberately for criminal or deceptive purposes, such as cloning voices for scams.

There are *supply chain harms* derived from representational biases that are a reflection of both developers' choices and prejudices embedded in the training data, as well as harms that come from the inappropriate incorporation of personal data, creative production or intellectual property into the development processes of AI.

Finally, given the lack of public understanding of AI, the rapidly increasing verisimilitude of outputs, and the absence of robust transparency and accountability, the combination of poorly functioning and misused technology brings *structural harms*—in this case undermining broader trust in information and public processes.[15]

High-risk usages of synthetic media and generative AI may not be easily defined and will depend on context, and the potential for differential impact. Risk-based approaches to regulation transfer a lot of responsibility to the private sector and may result in regulatory gaps impacting those consumers already most at risk. A rights-based approach—not only a risk-based approach—that is grounded in US Constitutional values is key to protecting the interests of consumers and citizens in the US. Existing legal frameworks on data protection and sector specific regulations for health or financial markets, for instance, provide a basis for action, as do the protections in the White House AI Bill of Rights in relation to broader AI issues, Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, and the Organization for Economic Co-operation and Development's (OECD's) 2019 Recommendation on Artificial Intelligence that the US adopted.

As I highlight later, comprehensive data privacy legislation and if needed AI-specific regulation that incorporates strong privacy protections would provide a core bedrock for addressing both strong implementation of safeguards for consumers such as transparency, as well as supporting future developments in AI technologies.

[15] Matt Davies and Michael Birtwistle, *Seizing the 'AI moment': making a success of the AI Safety Summit*, Ada Lovelace Institute, September 2023, https://www.adalovelaceinstitute.org/blog/ai-safety-summit/; for additional discussion of evaluating impacts, Irene Solaiman et al., *Evaluating the Social Impact of Generative AI Systems in Systems and Society*, June 2023, https://arxiv.org/abs/2306.05949

Synthetic media tools are now able to produce images of real-life events and realistic audio of individuals with limited input data, and at scale. Increased volume of easily made, realistic synthetic photos, audio and eventually audio, of specific real individuals and contexts is a paradigm shift. In the future, this will include more accurate, targeted and interactive personalization for a given context, individual consumer, specific user or audience in existing social media contexts, as well as emerging formats for communications such as VR and AR.[16] Generative AI tools are increasingly multimodal, with text, image, video, audio and code functioning interchangeably as input or output.

It is unreasonable to expect consumers and citizens to be able to 'spot' deceptive and realistic imagery and voices. As the Federal Trade Commission (FTC) has already noted,[17] most of the challenges and risks with generative AI cannot be addressed by the consumer acting alone. Similarly, responses to the risks of these tools cannot be adequately addressed by regulatory agencies or laws without a pipeline of responsibility across foundation models, developers and deployers of AI models.

Single technical solutions will not be sufficient either. In the case of audiovisual generative AI, deepfakes and synthetic media, technical approaches to detection will need to be combined with privacy-protecting, accessible watermarking and opt-in provenance approaches, and with mandatory processes of documentation and transparency for foundation models, pre-release testing, third-party auditing, and pre/post-release human rights impact assessments.

## Harms and risks of deepfakes, generative AI and synthetic media identified by WITNESS

Through the past five years of WITNESS consultations, civil society leaders have consistently identified a set of existing harms and potential threats from synthetic media and deepfakes. As tools have become more accessible and personalizable, and easier to use, a higher number of people have had the ability to engage with them. They have been able to imagine—or experience—how these technologies could impact their lives.

The main overarching concern echoes across countries: threats from synthetic media will disproportionately impact those who are already at risk, because of their ethnicity, gender,

[16] Eric Horvitz, *On the Horizon: Interactive and Compositional Deepfakes*, 2022, https://arxiv.org/abs/2209.01714; Thor Benson, *This Disinformation is Just for You,* WIRED, August 2023, https://www.wired.com/story/generative-ai-custom-disinformation/

[17] FTC Business Blog, *Chatbots, deepfakes, and voice clones: AI deception for sale*, March 2023, https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale

sexual orientation, profession, or belonging to a social group. Women particularly already face widespread threats from non-consensual sexual images that do not require high-quality or complex production to be harmful. Many marginalized and vulnerable populations have already been affected by the existing AI-driven dynamics of the information ecosystem. They have experienced AI and other forms of technology that have brought differential and/or disparate impact to them. This reflects both the biases in these tools (e.g. representational bias), as well as their use and misuse to disproportionately target these populations.

Elections in the coming year are poised to be deeply influenced by the malicious or deceptive use of generative AI. We hear how the fear of synthetic media, combined with the confusion about its capabilities and the lack of knowledge to detect AI-manipulation, are misused to dismiss authentic information with claims it is falsified. This is so-called plausible deniability or the "liar's dividend".[18] In our work analyzing claims of deepfakes, incidents of the liar's dividend are highly prevalent.

Similarly, these tools could be used by foreign governments to close civil society space by, for instance, incorporating them into patterns of criminalization and harassment of journalists and human rights defenders, and disinformation targeting their activities and those of political opponents at home and abroad. The potential threats brought by synthetic media and generative AI have motivated governments to enact laws suppressing free expression and dissent, posing a threat to the principles of free expression, civic debate and information sharing. Proposed rule-making and legislation on generative AI and deepfakes in China is indicative of this trend.[19]

Lastly, pressures to understand complex synthetic content, and claims that content is synthesized, place additional strain on already under-resourced local and national newsrooms and community leaders responsible for verifying digital content. With hyperbolic rhetoric as well as the realities of advances in generative AI undermining trust in content we encounter, human rights defenders, journalists and civil society actors will be among the most impacted by generative AI.

These technologies need to be developed, deployed, or regulated with an in-depth understanding of a range of other local and national contexts. The voices of those impacted by

[18] Robert Chesney and Danielle Keats Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, 107 California Law Review 1753, July 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954

[19] Karen Hao, *China, a Pioneer in Regulating Algorithms, Turns Its Focus to Deepfakes*, Wall Street Journal, January 2023 https://www.wsj.com/articles/china-a-pioneer-in-regulating-algorithms-turns-its-focus-to-deepfakes-11673149283

AI, need to be central to the discussion and prioritization of solutions.[20] Yet, emerging technologies are designed and deployed without the input of those most impacted, ignoring the threats and risks these technologies bring to communities already at a disadvantage.

**Why consumer-facing transparency matters in an information environment with more complex creative and communicative production**

Media, communication and content production are increasingly complex. Increased access to tools for creative generation and knowledge production will bring benefits to society. However, to realize this, one key component is transparency across the pipelines of AI design, content production and information distribution.[21] Transparency approaches can also support better control for individuals and others on how their data is used in AI models.

Frequently in my work, I am asked to provide advice to consumers on how to spot an AI-generated image—for example, to look for 'the distorted hands', or in the case of a deepfake, to see if it does not blink. I discourage this as these heuristics are the current Achilles heel or temporary failings of a process, not long-term durable or scalable guidance. Most audiovisual content we create and consume involves AI. In a world with wider access to tools that simplify the generation or edition of photos, videos, and audio, including photo and audio-realistic content, it is important for the public to be able to understand if and how a piece of media was created or altered using AI. We refer to this as 'content provenance'. Such labeling, watermarking or indications of provenance are not a punitive measure to single out AI content or content infused with AI, and should not be understood as a synonym of deception, misinformation or falsehood. The vast majority of synthetic media is used for personal productivity, creativity or communication without malice. Satirical media made using AI is also a critical and protected form of free speech.[22]

We have heard repeatedly from information consumers around the world that responsibility should not be placed primarily on end-users to determine if the content they are consuming is AI-generated, created by users with another digital technology or, as in most content, a mix of

[20] Sam Gregory, Journalism, December 2021, *ibid.*
[21] Sam Gregory, *Synthetic media forces us to understand how media gets made,* Nieman Lab, December 2022, https://www.niemanlab.org/2022/12/synthetic-media-forces-us-to-understand-how-media-gets-made/
[22] Henry Ajder and Joshua Glick, *Just Joking! Deepfakes, satire, and the politics of synthetic medi*a, WITNESS and MIT, December 2012, https://cocreationstudio.mit.edu/just-joking/

both.[23] To ensure disclosure—and more broadly, to promote transparency and accountability—all actors across the AI and media distribution pipeline need to be engaged. These include:

- Those researching and building foundation or frontier models;
- Those commercializing generative AI tools;
- Those creating synthetic media;
- Those publishing, disseminating or distributing synthetic media (such as media outlets and platforms); and
- Those consuming or using synthetic media in a personal capacity

There is now a significant trend in AI governance towards a pipeline approach and a focus on labeling and disclosure. In July 2023, seven leading AI companies agreed with the White House 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.[24] In the European Union, companies who have signed on to the voluntary EU Code of Practice on Disinformation have agreed to a similar commitment, with the EU's Commissioner Věra Jourová calling on these companies to label AI-generated content.[25] The EU AI Act includes significant requirements for disclosing deepfakes and machine-generated content from foundation models.

Most provenance systems will require methods that explain both AI-based origins or production processes, but also document non-synthetic audio or visual content generated by users or other digital processes–like footage captured from 'old fashioned' mobile devices.[26] As the White House notes in its statement on the voluntary commitments, "companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI." It will be hard to address AI

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

[24] 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/ and The White House, *Ensuring Safe, Secure, and Trustworthy AI,* https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[25] Foo Yun Chee, *AI-generated content should be labelled, EU Commissioner Jourova says*, Reuters, June 2023 https://www.reuters.com/technology/ai-generated-content-should-be-labelled-eu-commissioner-jourova-says-2023-06-05/

[26] Sam Gregory, *To battle deepfakes, our technologies must track their transformations*, The Hill, June 2022, https://thehill.com/opinion/technology/3513054-to-battle-deepfakes-our-technologies-must-lead-us-to-the-truth /

content in isolation from this broader question of media provenance. Implementing this approach to transparency will require standards that focus on how to design durable, machine-readable shared standards that provide useful signals to consumers, as well as other actors in the information pipeline (e.g. content distributors and platforms).

## The opportunity in transparency and disclosure

It is crucial for democracy that people are able to believe what they see and hear when it comes to critical government, business and personal communications, as well as documentation of events on the ground. It is also critical for realizing the creativity and innovation potential that generative AI holds, that consumers are informed about what they see and hear.

WITNESS has actively participated in the Partnership on AI's Responsible Practices for Synthetic Media Framework.[27] This Framework describes *direct* forms of disclosure as those methods that are 'visible to the eye', such as labels marking the content, or adding context disclaimers. *Indirect* forms of disclosure are not perceptible to the human eye and include embedded metadata or other information that is machine readable or presentable, such as cryptographic provenance or embedding durable elements into either or both the training data and the content captured or generated.[28] Importantly, the Framework also offers a useful breakdown of how responsibility for supporting this disclosure should be considered at different stages across the AI pipeline.

There is significant and unhelpful confusion around terms used to show use of AI in content.[29] 'Watermarking' is used as a catch-all term that includes:

a. Visible watermarks, signals or labels (e.g. a 'Made with AI' description on an image);

b. Invisible watermarks and technical signals that are imperceptible to the human eye and can be embedded at pixel level, or as early as the training stage of AI processes; and

c. Cryptographically-signed metadata that shows the production process of content over time, like the C2PA standards.

Visible signals or labels can be useful in specific scenarios such as AI-based imagery or production within election advertising, as proposed in the REAL Political Ads Act. However,

[27] Partnership on AI, *Responsible Practices for Synthetic Media Framework*, https://syntheticmedia.partnershiponai.org/ See also, Jacobo Castellanos, *Building Human Rights Oriented Guidelines for Synthetic Media*, WITNESS, February 2023, https://blog.witness.org/2023/02/building-human-rights-oriented-guidelines-for-synthetic-media/

[28] For synthetic content, the most recent example is SynthID, released by Google on August 29, 2023. https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid

[29] Claire Leibowicz, *Why watermarking AI-generated content won't guarantee trust online*, MIT Tech Review, August 2023, https://www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/

visible watermarks are often easily cropped, scaled out, masked or removed, and specialized tools can remove them without leaving a trace. Visible watermarks are hence inadequate for reflecting the 'recipe' for the use of AI in an image or video, and in a more complex media environment fail to reflect how AI is used in a meaningful way for consumers.

Technical interventions and signals at the dataset level can indicate provenance as well as embed 'Do Not Train' restrictions that could give consumers more say in who is allowed to build AI models using people's data and content. However, many datasets are already in use and do not include these marks. Additionally, small companies, independent developers, and open-source libraries and tools may not have the capacity and ability to develop reliable and sustainable invisible watermarks. Without accessible and standardized standards, there is a risk of excluding a significant part of the AI innovation ecosystem from the conversation. This, in turn, could lead to a handful of AI companies' dominance becoming further entrenched. Dataset-level watermarks also require their application across broad datasets, which brings questions around ownership and responsibility regarding the content and the repurposing of that content for training purposes. Since, in most cases, the original content creators are not involved in the decision to add their content to a training dataset, they are unlikely to be involved in the decision to watermark their content as well.

Cryptographic signature and provenance-based standards such as the C2PA are built to make it very hard to tamper with the cryptographic signature without leaving evidence of the attempt, and to enable the reconnection of a piece of content to a set of metadata if that is removed. These methods can allow people to understand the lifecycle of a piece of content, from its creation or capture to its production and distribution. In some cases, they are integrated with capture devices such as cameras, utilizing a process known as 'authenticated capture'. Microsoft has been working on implementing provenance data on AI content using C2PA specs[30], and Adobe has started to provide it via its Content Credentials approach.[31] While I do not speak for the C2PA, WITNESS is a member of the C2PA, has participated in the Technical Working Group and acted as a co-chair of the C2PA Technical Working Group Threats and Harms Taskforce. In this context WITNESS has advocated for globally-driven human rights perspectives and practical experiences to be reflected in the technical standard.[32]

[30] Kyle Wiggers, *Microsoft pledges to watermark AI-generated images and videos,* Techcrunch, May 2023 https://techcrunch.com/2023/05/23/microsoft-pledges-to-watermark-ai-generated-images-and-videos

[31] Adobe Content Credentials, https://helpx.adobe.com/creative-cloud/help/content-credentials.html

[32] The Coalition for Content Provenance and Authenticity, *C2PA Harms Modelling*, https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html ; Jacobo Castellanos, WITNESS, 2021, *Ibid.*

An approach like the C2PA can also allow creators to choose whether their content may be used for training AI models or other data purposes. Agency over their data is a critically needed response to concerns from creators and others about the incorporation of their content, personal images and other data into AI models without their consent.

In reality, any effective shared standard, regulation, or technological solution to provenance, disclosure and transparency is likely to require a combination of cryptographically-signed provenance metadata that reflects how both AI, non-AI and mixed media are created and edited over time, as well as visible watermarking and/or technical signals for synthetic content that confirm the use of AI specifically.

### How to provide rights-respecting disclosure

To safeguard Constitutional and human rights, approaches to provenance and disclosure need to meet at least three core criteria. They need to:

- Protect privacy:
- Be accessible with modular opt-in or out depending on the type of media and metadata, and;
- Avoid configurations that can be easily weaponized by authoritarian governments.

People using generative AI tools to create audiovisual content should not be required to forfeit their right to privacy to adopt these emerging technologies. Personally-identifiable information should not be a prerequisite for identifying either AI-synthesized content or content created using other digital processes. The 'how' of AI-based production elements is key to public understanding; this should not require a correlation to the identity of 'who' made the content or instructed the tool.

Since 2019, WITNESS has been raising concerns about the potential harms that could arise from the inclusion of personal data in solutions that track the provenance of media.[33] The US government has the opportunity to ensure that provenance requirements and standards are developed in-line with global human rights standards, protect civil rights and First Amendment rights, and do not include the automated collection of personal data. While a requirement to include disclosure indicating content was AI-generated could be a legal requirement in certain

[33] Gabriela Ivens and Sam Gregory, *Ibid*; Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all*, WITNESS, 2020, https://blog.witness.org/2020/05/authenticity-infrastructure/

cases, this obligation should not extend to using tools for provenance on content created outside of AI-based tools, which should always be opt-in.

Building trust in content must allow for anonymity and redaction. Immutability and inability to edit do not reflect the realities of people, or how and why media is made–nor that certain redaction may be needed in sensitive content.[34] Flexibility to show how media evolves—and to conduct redaction—is a functional requirement for disclosure particularly as it relates to edited and produced content. Lessons from platform policies around 'real names' tell us that many people—for example, survivors of domestic violence—have anonymity and redaction needs that we should learn from.[35] While specifications like the C2PA focus on protecting privacy and don't mandate identity disclosures, this privacy requirement needs to be protected during widespread adoption. We should be wary of how these authenticity infrastructures could be used by governments to capture personally identifiable information to augment surveillance and stifle freedom of expression, or facilitate abuse and misuse by other individuals.

We must always view these credentials through the lens of who has access and can choose to use them in diverse global and security contexts, and ensure they are accessible and intelligible across a range of technical expertise.[36] Provenance data for both AI and user-generated content provides signals—i.e. additional information about a piece of content—but does not prove truth. An 'implied truth' effect simply derived from the use of a particular technology is not helpful, nor is an 'implied falsehood' effect from the choice or inability to use them.[37] Otherwise we risk discrediting a citizen journalist for not using tools like these to assert the authenticity of their real-life media because of security or access concerns, while we buttress the content of a foreign state-sponsored television channel that does use it. Their journalism can be foundationally unreliable even if their media is well-documented from a provenance point of view.

Any credential on content must be an aid to help make informed decisions, not a simplistic truth signal. They work best as a signal in complement to other processes of digital and media literacy that consumers choose to use, to help them triage questions they may have, and that are available to other parties engaging with the content, including potentially platforms.

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

[35] Jillian York and Dia Kayyali, *Facebook's 'Real Name' Policy Can Cause Real-World Harm for the LGBTQ Community*, EFF, 2014, https://www.eff.org/deeplinks/2014/09/facebooks-real-name-policy-can-cause-real-world-harm-lgbtq-community

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

[37] Sam Gregory, Journalism, December 2021, *ibid.*

### The role of detection alongside provenance

'Seeing' both invisible watermarks and provenance metadata that are imperceptible to the eye will require consumer-facing tools. However, the average citizen shouldn't be required to keep up with watermarking advances and detection tools, and cannot be expected to deploy multiple tools to ascertain if a particular commercial brand, watermarking approach, or mode of synthesis has been used.

Detection tools are also necessary for content believed to be AI-generated that does not have provenance information or that has been manipulated with counter-forensics approaches. There is justifiable skepticism about whether after-the-fact detection tools are useful for consumer transparency and consumer usage to identify generative AI and deepfake outputs.[38] Detection of audiovisual generative AI and deepfakes outputs is flawed. Existing detection models frequently require expert input to assess the results and often they are not generalisable across multiple synthesis technologies and techniques or require personalization to a particular person to be protected from fraudulent voices or imagery. As such, detection tools can lead to unintentional confusion and exclusion. We have seen how thuse by the general public of detection tools has contributed to increased doubt around real footage and enabled the use of the liar's dividend and plausible deniability around real content, rather than contributing to clarity.[39]

However, from WITNESS's experience they are a critical element—alongside the incorporation of provenance data and media literacy—when it comes to real-world scenarios where journalists, civil society and governments are attempting to discern how content has been created and manipulated. As we have seen in our work supporting forensic analysis of high profile global cases, there is a gap between on one side the needs of journalists, civil society leaders and election officials, and on the other side the availability of detection skills, resources and tools that are timely, effective and grounded in local contexts. These issues highlight the 'detection equity' gap that exists—the tools to detect AI-generated media are not available to the people who need them the most. Further research into improving detection capabilities remains critical as well as ensuring those who access tools also have the knowledge and skills to use them.

[38] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, Luisa Verdoliva, *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), On The Detection of Synthetic Images Generated by Diffusion Models, https://arxiv.org/abs/2211.00680*; Luisa Verdoliva, *Media Forensics and Deepfakes: An Overview*, 2020, https://arxiv.org/abs/2001.06564

[39] Sam Gregory*, 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/ ;* Nilesh Christopher, *ibid*; 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/

## Conclusion

Significant evolutions in volume, ease of access, personalization and malicious usage of generative AI reflect both the potential for creativity but also the heightened harms from audiovisual generative AI and deepfakes–including the plausible deniability that these tools enable, undermining consumers' trust in the information ecosystem. I have highlighted in this statement the need to focus on existing harms as identified by those on the frontlines of deepfakes and synthetic media, and to center the voices of those affected by an AI-powered information landscape. I encourage this Subcommittee and legislators to go beyond a risks-based approach, and push for a rights-based framework in order to prevent and mitigate accidental harms, misuse harms, supply chain harms and structural harms. In this regard, approaches to transparency in audiovisual content production that incorporate strong privacy measures can protect personal information, safeguard democracy around the world, and promote creative production.

Sam Gregory

Executive Director

WITNESS
SEE IT FILM IT
CHANGE IT