WITNESS
SEE IT FILM IT CHANGE IT

Testimony of Sam Gregory, Executive Director, WITNESS

Before the United States House Committee on Oversight and Accountability, Subcommittee on Cybersecurity, Information Technology, and Government Innovation

'Advances in Deepfake Technology'

Date of hearing: November 8, 2023

Chairwoman Mace, Ranking Member Connolly and members of the House Oversight and Accountability Subcommittee on Cybersecurity, Information Technology and Government Innovation, thank you for the opportunity to testify today.

I am Sam Gregory, Executive Director of WITNESS.[1] Since 2018, WITNESS has led a global effort, *Prepare, Don't Panic*, to understand deepfake and synthetic media technologies, and more recently large language models (LLMs) and generative AI. In our consultations with different experts and communities, we have researched advances in deepfake technologies, assessed how deepfakes and their related harms are impacting society in the US and globally, and developed a set of recommendations to prepare accordingly.[2]

WITNESS efforts have included tracking technical developments, contribution to technical standards development,[3] engagement on detection and authenticity approaches that 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 to understand harmful misuses.[7] My testimony 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 and harmful attacks on themselves and their work.

Deepfake technologies, with their increasing 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 more broadly, our trust in the information we see and hear. In this statement I will cover these advances in deepfake technology, the relationship of these advances to harmful uses, and how Congress can mitigate these harms.

---

[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://journals.sagepub.com/doi/10.1177/14648849211060644 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/*

Advances in deepfake technologies

For purposes of this hearing, I will use the term 'deepfake technologies' to refer to synthetic media and generative AI technologies which draw on advances in artificial intelligence (AI) and related fields to create or manipulate entirely or partially synthetic or simulated images, video and audio.

Headlines in previous years often over-hyped the capabilities of deepfake technology to the detriment of public understanding. My organization, WITNESS, has for a number of years promoted a perspective of *Prepare, Don't Panic* in relation to deepfakes and generative AI. We have consistently pushed back against counter-productive hype, calling instead for investments in thoughtful preparation. However, the moment to address the advances in these technologies and their impact has now come.

First, let me share relevant advances in deepfake technologies, primarily focusing on describing trends in layperson terms.

I should begin by noting that improvements in technologies do not necessarily map to public understanding of such technologies. Some deepfake capabilities have become widely available. For instance, faking and cloning audio from a few seconds sample, generating an image from a text prompt, creating a synthetic avatar, producing a simple animated video from one photo, inserting a plausible 'nudified' body in place of a clothed body, and swapping faces with celebrities via an app. Despite public perceptions to the contrary, other formats—for example, realistic and convincing video recreation of a complex scene, as well as truly convincing face-swaps—remain out of the reach of most people without significant resources or additional technical skills, capacity for post-production and additional Computer Generated Imagery (CGI) effects, and computational resources.

Frames I have found useful in my work to understand the implications of these underlying technology shifts, and to then assess responses, include:

Commercialization, commoditization and accessibility: Deepfake technology is no longer just niche apps or code that require significant technical knowledge and resources, or that users must search-out. Image and audio creation tools are available within commonly accessed consumer and prosumer platforms including Google Search,[8] Microsoft Bing, and Adobe Firefly as well as in widely-used apps available in online stores. Apps provide access to some versions of video tools as do slightly less commonly used commercial platforms such as Runway ML. For those with a level of technical knowledge open-source code and other components of open source AI models are available on accessible repositories—alongside guidance from fellow users—while computational power

[8] Jay Peters, *Google's AI-powered search experience can now generate images*, 12 October 2023, The Verge https://www.theverge.com/2023/10/12/23913337/google-ai-powered-search-sge-images-written-drafts

can be accessed via cloud computing. On the discovery side, a search query on widely used search engines will provide a user with links to deepfake sexual imagery sites.[9]

Ease of use: Many deepfake technologies are increasingly easy to use. Image tools and certain video tools can be instructed with natural language prompts or by the provision of an image or video to adapt or work from. Audio tools are initiated by text or voice input. They do not require an individual to have their own coding skills, hardware or computational power or to invest in training or significant initial costs. With newer so-called diffusion models for generation, they are more flexible to the requests of the creator and more adaptable than previous Generative Adversarial Network (GAN) based models that characterized the initial face-swap approaches most commonly known to the public as 'deepfakes'.

Volume and variation: With audio and image particularly, it is possible to rapidly create a volume of similar images or variations on a particular image or audio. For example, a common element in the interface of a commonly used image generation platform like Midjourney is to create variations of an image, while text-to-audio and voice cloning tools allow rapid experimentation with audio versions.

Multimodality: The concept of multimodality covers the idea that a tool can take inputs in one format, for example text, and output another format, for example, an image. The broader category of generative AI tools are increasingly multimodal, with text, image, video, audio and code functioning interchangeably as input or output.

Improving quality: Over the past year, the quality and customization of realistic image and audio generation have improved dramatically on commonly available consumer tools.[10] These shifts involve both the realism of content produced with synthetic media technologies that purports to show real individuals or lifelike scenes, as well as their fidelity to prompts or requests to produce non-realistic content - for example, fantastical scenes - content that matches the prompt.

Personalization and impersonation: The capacity to produce images, audio and videos that manipulate or synthesize a known real individual is broadly available. In the case of images and audio, limited data is required to produce a novel simulated image or audio track of a known individual (for example, a single image, or a minute of audio). Future research on audio indicates that this will be reduced further to a few seconds of input audio, and will also then retain voice, emotion and acoustic environment.[11] While video remains harder to do in complex real-world scenarios, consumer apps can swap an

[9] Matt Burgess, *Deepfake Porn Is Out of Control*, WIRED, 16 October 2023, https://www.wired.com/story/deepfake-porn-is-out-of-control/
[10] See for example in this comparison over one year of the evolution of one such tool, MidJourney.AI Art & Photography, *Midjourney ALL Versions from V1 to V5.2 Comparison*, YouTube,10 August 2023, https://www.youtube.com/watch?v=_6_waARUr6k
[11] *Vall-E (X): A neural codec language model for speech synthesis*, Microsoft, https://www.microsoft.com/en-us/research/project/vall-e-x

individual's face onto another's body or animate a person from a single photo. Live deepfakes are increasingly a concern, including in injection attacks that could compromise biometric systems.[12] Matching lip movements to new words in a video or a translated version is feasible and platforms like YouTube have indicated wider availability in 2024, oriented towards language accessibility, and increasing access more broadly.

To create these user-facing shifts, a series of technical advances in generative adversarial networks (GANs), diffusion models, variational autoencoders and transformer models are increasingly being explored in combination, or used for particular purposes for which they perform well, in order to improve the speed and flexibility of the synthetic production process and the quality of the output. Further underlying technical trends in the generation of media include "the increased use and improvement of multimodal models, such as the merging of LLMs and diffusion models; the improved ability to lift a 2D image to 3D to enable the realistic generation of video based on a single image; faster and tunable methods for real time modified video generation; and models that require less input data to customize results, such as synthetic audio that captures the characteristics of an individual with just a few seconds of reference data."[13]

Looking ahead, future trends suggest that the tools currently available are 'the worst they'll ever be'. Likely advances include continued improvements in the ease of instructing these tools in plain language; reduction in the quantity of input media required; higher ability to tailor those outputs and more realistic outputs; improved audio with simulation of voice, emotion and acoustic environment; and eventually similar advances in video to what we now see in audio and images.

The combination of LLMs, other AI/machine learning tools and deepfake technologies will enable targeting of consumers and others with synthetic content that is more specifically tailored to them, and provide interactive personalization for a given context, individual consumer, specific user or audience in existing social media contexts, as well as in emerging formats for communications such as extended reality (XR), virtual reality (VR) and augmented reality (AR).[14]

---

[12] Kevin Carta, Claude Barral, Nadia El Mrabet, Stefane Mouille, *Video injection attacks on remote digital identity verification solution using face recognition*, Proceedings of the 13th International Multi-Conference on Complexity, Informatics and Cybernetics (IMCIC 2022) https://www.iiis.org/CDs2022/CD2022Spring/papers/ZA639OX.pdf

[13] Joint CSI, *Contextualizing Deepfake Threats to Organizations*, National Security Agency/Federal Bureau of Investigation/Cybersecurity and Infrastructure Security Agency, 12 September 2023, https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

[14] 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/

The domestic and global context of harmful misuses of audiovisual generative AI and deepfake technologies

While there are undoubtedly creative and commercial benefits to deepfake technologies, there are already harms in the US and globally, with disproportionate impacts on groups already at-risk of discrimination or vulnerable to offline harms.

WITNESS global consultations have consistently identified a concern across countries: threats from deepfake technologies will disproportionately impact those who are already at risk, because of their ethnicity, gender, sexual orientation, profession, or belonging to a social group. These risks will often build on existing dynamics of harmful uses with previous technologies, and the AI-driven dynamics of the information ecosystem, in particular on platforms where these communities have experienced differential and/or disparate impact to them.

Existing harms are exacerbated by deepfake technologies. Women already face widespread threats from non-consensual sexual images or release of intimate partner images that do not require high-quality or complex production to be harmful. Non-consensual sexual deepfake images and videos are currently used to target private citizens and public figures, particularly women. The scale of this is growing: a recent analysis estimated that even based on public site-sharing of these videos (rather than videos posted on social media, those shared privately, or manipulated photos) by the end of this year, more videos will have been produced in 2023 than the total number of every other year combined.[15] A recent trend documented globally - recently in the US, Spain and Brazil - is a broadening scope of this misuse in educational contexts targeting young women and girls.[16] The volume of AI-generated child sexual abuse material (CSAM) is also increasing and has raised concern among states Attorneys General across the US.[17]

Actors and others have had their likenesses stolen to use in non-satirical commercial contexts.[18] Existing imposter scams are already a primary source of consumer complaints

---

[15] Matt Burgess, ibid; Megan Farokhmanesh, *The Debate on Deepfake Porn Misses the Point,* WIRED, 1 March 2023 https://www.wired.com/story/deepfakes-twitch-streamers-qtcinderella-atrioc-pokimane/

[16] Manuel Viejo, *In Spain, dozens of girls are reporting AI-generated nude photos of them being circulated at school: 'My heart skipped a beat'*, El País, 18 September 2023 https://english.elpais.com/international/2023-09-18/in-spain-dozens-of-girls-are-reporting-ai-generated-nude-photos-of-them-being-circulated-at-school-my-heart-skipped-a-beat.html ; April Rubin, *Teens exploited by fake nudes illustrate threat of unregulated AI*, Axios, 3 November 2023 https://www.axios.com/2023/11/03/ai-deepfake-nude-images-new-jersey-high-school ; Pranshu Verma, *AI fake nudes are booming. It's ruining real teens' lives,* 5 November 2023, The Washington Post https://www.washingtonpost.com/technology/2023/11/05/ai-deepfake-porn-teens-women-impact/; Por Paula Ferreira, *Polícia Civil do Rio identifica parte dos jovens que criaram nudes falsos de alunas de colégio*, Estadão, 2 November 2023, https://www.estadao.com.br/educacao/policia-civil-do-rio-identifica-parte-de-jovens-que-criaram-nudes-falsos-de-alunas-de-colegio/

[17] 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/

[18] Anumita Kaur, *MrBeast, Tom Hanks, Gayle King warn of online deepfake ads,* October 3, 2023, https://www.washingtonpost.com/entertainment/2023/10/03/tom-hanks-ai-ad-deepfake/

to the FTC.[19] Now AI-simulated audio scams are proliferating.[20] AI audio used to deceive has occurred in high profile political contexts recently in Sudan, Slovakia and the UK.[21] Imposter profiles in social media such as LinkedIn as well as global covert influence campaign contexts and the use of lifelike synthetic avatars posing as real are more frequent.[22] Deepfake video 'injection attacks' permit spoofing of biometric systems.[23] In an organizational and business context, a recent joint CSI bulletin identified threats related to brand integrity, impersonation and deceptive communications for fraudulent access.[24]

WITNESS has been observing - in a range of cases where WITNESS has been asked to review cases, and in broader information ecosystems - how claims of AI-generation are used to dismiss real content from human rights defenders and journalists.[25] A number of these recent cases have involved audio, where because of the advances in technology, a claim that real content is AI-generated content - as noted above - is a plausible excuse, not a hypothetical. There are significant challenges ahead in both identifying what is faked, but also confirming what is authentic.

Text-to-image tools perpetuate existing patterns of prejudice, bias or discriminatory representation present in their training data.[26] Additionally, creatives and artists have had their creative work and 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.[27]

One significant concern we have heard in consultations is how 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. In addition, the *potential* threats brought by synthetic media and generative AI have motivated governments to suggest laws suppressing free expression and dissent,

---

19 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

20 Pranshu Verma, *They thought loved ones were calling for help. It was an AI scam,* 5 March 2023, The Washington Post, https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/

21 See Jack Goodman and Monahad Hashim, AI: Voice cloning tech emerges in Sudan civil war, 5 October 2023, BBC, https://www.bbc.com/news/world-africa-66987869; Morgan Meaker, Slovakia's Election Deepfakes Show AI Is a Danger to Democracy, 3 October 2023, WIRED, https://www.wired.co.uk/article/slovakia-election-deepfakes; Morgan Meaker *Deepfake Audio Is a Political Nightmare*, 9 October 2023, WIRED, https://www.wired.com/story/deepfake-audio-keir-starmer/

22 Shannon Bond, *That smiling LinkedIn profile face might be a computer-generated fake*, NPR, 27 March 2023, https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles ; Grafika, *Deepfake It Till You Make It*, February 2023, https://graphika.com/reports/deepfake-it-till-you-make-it; Florantonia Singer, *They're not TV anchors, they're avatars: How Venezuela is using AI-generated propaganda*, 22 February 2023, El Pais, https://english.elpais.com/international/2023-02-22/theyre-not-tv-anchors-theyre-avatars-how-venezuela-is-using-ai-generated-propaganda.html

23 Carta et al., ibid.

24 Joint CSI, ibid.

25 Nilesh Christopher, ibid.

26 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

27 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

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.[28]

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 existing harms and broader AI issues need to be central to the discussion and prioritization of solutions.[29]

Options for harm reduction: provenance and detection

There are a range of potential mitigations that Congress should be familiar with, relating to the advances in deepfake technology I described earlier. In this written testimony I focus on two particular areas of technical intervention: *provenance, watermarking and labeling/disclosure approaches*, and then on *detection of synthetic content*.

Given the technological shifts I have highlighted, it is unreasonable to expect consumers and citizens to be able to 'spot' deceptive and realistic imagery and voices. Guidance to 'look for the six-fingered hands,' or inspect visual errors in a Pope in a puffer jacket, or to look to see if a suspect image does not blink, or listen super-close to the audio to hope to hear an error are not sufficient and do not help in the long run or even medium-term. These heuristics are the current Achilles heel or temporary failings of a process, not long-term durable or scalable guidance.

Both recent and longer-standing research related to audiovisual fakery including deepfakes also indicates that humans do trust the realism cues of audio and video,[30] cannot identify machine-generated speech cloning accurately,[31] do not recognize simulated human faces,[32] do not fare well spotting face-swapped faces,[33] and retain false memories of deepfakes.[34] As the Federal Trade Commission (FTC) has already noted,[35] most of the challenges and risks associated with generative AI cannot be addressed by the consumer or individual acting alone.

---

[28] 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
[29] Sam Gregory, Journalism, December 2021, *ibid*.
[30] 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
[31] 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
[32] 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
[33] 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
[34] 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
[35] 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

Single technical solutions will also not be sufficient. 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. In a broader context of AI governance, these should be complemented by processes of documentation and transparency for foundation models, pre-release testing, third-party auditing, and pre/post-release human rights impact assessments.

The opportunity in disclosure of use of deepfake technologies[36]

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 broader generative AI technology holds, that consumers are informed about what they see and hear.

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. 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.[37]

WITNESS has participated in the Partnership on AI's Responsible Practices for Synthetic Media Framework.[38] This Framework describes *direct* forms of disclosure as those methods that are 'visible to the eye', such as labels marking the content with a note such as 'Made with AI', or adding context disclaimers. *Indirect* forms of disclosure are not perceptible to the human eye and include cryptographically-signed embedded metadata or other information that is machine readable or presentable and shows the production process content over time or embeds durable watermarking elements into either or both the training data and the content captured or generated.[39] 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.

---

[36] This section draws on testimony provided in a Senate Commerce Committee Subcommittee on Consumer Protection, Product Safety and Data Security hearing, *The Need for Transparency in Artificial Intelligence*, 12 September 2023, https://www.commerce.senate.gov/2023/9/the-need-for-transparency-in-artificial-intelligence

[37] Henry Ajder and Joshua Glick, *Just Joking! Deepfakes, satire, and the politics of synthetic media*, WITNESS and MIT, December 2012, https://cocreationstudio.mit.edu/just-joking/

[38] 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/

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

Visible, direct disclosure signals or labels can be useful in specific scenarios such as AI-based imagery or production within election advertising. However, visible watermarks are often easily cropped, scaled out, masked or removed, and specialized tools can remove them without leaving a trace. As a result, they are 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 deepfake technology is used. Labels also bring up questions about their interpretability and accessibility by different audiences,[40] from the format of the label, to their placement or language they employ.

Cryptographic signature and provenance-based standards track the production process of content over time, and enable the reconnection of a piece of content to a set of metadata if that is removed. They also make it hard to tamper with them without leaving evidence of the attempt. One such initiative is the standard proposed by the Coalition for Content Provenance and Authenticity (C2PA). Microsoft has been working on implementing provenance data on AI content using C2PA specifications,[41] and Adobe has started to provide it via its Content Credentials approach.[42] 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, in a process known as 'authenticated capture'. 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.[43]

Invisible watermarks, like Google's SynthID,[44] generally focus on embedding a digital watermark directly into the pixels of AI-generated images, making it imperceptible to the human eye. These types of approaches are increasingly robust to modifications like cropping, adding filters, changing colors and lossy compression schemes. However, they are not yet interoperable across watermarking and detection techniques. Without standardization, watermarks created by an image generation model may not be detected confidently enough by a content distribution platform, for instance. Similarly, the utility of invisible watermarking may be restricted beyond closed systems. According to Everypixel Journal, more than 11 billion images have been created using models from three open source repositories.[45] In these situations, invisible watermarks can be removed by deleting

[40] Kat Cizek and shirin anlen, *The Thorny Art of Deepfake Labeling*, 5 May 2023, WIRED https://www.wired.com/story/the-thorny-art-of-deepfake-labeling

[41] 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

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

[43] 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.*

[44] *SynthID*, Google Deepmind, https://www.deepmind.com/synthid

[45] *AI Has Already Created As Many Images As Photographers Have Taken in 150 Years*. Statistics for 2023, Everypixel Journal, https://journal.everypixel.com/ai-image-statistics

the line that generates it. Promising research by Meta on Stable Signature roots the watermark in the model and allows tracing the image back to where it was created, even being able to deal with various versions of the same model.[46]

Technical interventions at the dataset level can help indicate the origin of a piece of content from a particular deepfake technology approach. However, many datasets are already in use and do not include these signals. Additionally, small companies and independent developers may not have the capacity and ability to develop this type of watermarking. Dataset-level watermarks also require their application across broad data collections, which brings questions around ownership. As we are seeing from copyright lawsuits, the original content creators have usually not consented to add their content to a training dataset. Given the current data infrastructure, they are unlikely to be involved in the decision to watermark their content.

Such interventions will take place within an information environment with more complex creative and communicative production. Most audiovisual content we create and consume will involve AI. This makes binary solutions oriented towards identifying, signaling or detecting "AI or not" hard to implement.

Reflecting this complexity, 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.[47]

Comprehensive approaches will require the integration of detection and provenance across the pipelines of AI design, content production and information distribution.[48] 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 both.[49] 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;

---

[46] AI at Meta, *Stable Signature: A new method for watermarking images created by open source generative AI*, 6 October 2023, https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/

[47] Raquel Vazquez Llorente and Sam Gregory, *Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights*, Tech Policy Press, October 2023, https://techpolicy.press/regulating-transparency-in-audiovisual-generative-ai-how-legislators-can-center-human-rights/

[48] 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/

[49] 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

- 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.

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.[50] It will be hard to address AI content in isolation from this broader question of media provenance.

Similarly, 'seeing' both invisible watermarks and provenance metadata that are imperceptible to the eye will require consumer-facing tools. However, the average person 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. Implementing this approach to transparency will require standardized, 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) - as proposed for further Department of Commerce-led research in the recent White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[51]

How to provide rights-respecting disclosure

To safeguard Constitutional and human rights, approaches to provenance and disclosure should 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 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**'

[50]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 /
[51]White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Section 4.5, 30 October 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

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.[52] The US government has the opportunity to ensure that provenance requirements and standards for deepfake technologies protect civil rights and First Amendment rights, are developed in-line with global human rights standards, 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 very specific 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, particularly for human-generated content. 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.[53] 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.[54] While specifications like the C2PA focus on protecting privacy and do not 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.[55] 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.[56] 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

[52] 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/
[53] Raquel Vazquez Llorente, *Trusting Video in the Age of Generative AI*, Commonplace, June 2023, https://commonplace.knowledgefutures.org/pub/9q6dd6lg/release/2
[54] 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
[55] Sam Gregory, *Ticks Or It Didn't Happen*, WITNESS, December 2019, https://lab.witness.org/ticks-or-it-didnt-happen/
[56] Sam Gregory, Journalism, December 2021, *ibid.*

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. These technical measures 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.

The role of detection alongside provenance

Detection tools are necessary for content believed to be AI-generated that does not have provenance information or that has been manipulated with counter-forensics approaches. They also work in concert with the signals provided by provenance and disclosure mechanisms.

In the direct experience of my own organization in supporting analysis of high-profile suspected deepfakes encountered globally, it is challenging to do rapid, high-quality media forensics analysis. Detection resources and signals of AI usage 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.[57] 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.

In WITNESS' work we also see 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".[58] In our work analyzing claims of deepfakes, incidents of the liar's dividend are the most prevalent.

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.[59] Scaled detection of deepfake technology outputs across multiple modes of

[57] 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.
[58] 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
[59] 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; João Phillipe Cardenuto et al., *The Age of Synthetic Realities: Challenges and Opportunities,* APSIPA Trans. on Signal and Information Processing (invited paper under review), June 2023, https://arxiv.org/abs/2306.11503

production and in real-world conditions is challenging[60], and even under best case scenarios where multiple techniques are combined is unlikely to reach more than 85-90% accuracy. Existing 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. Use by the general public of detection tools - often found online - 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.[61]

However, from WITNESS's experience these detection tools 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 a 'detection equity' gap —the tools to detect AI-generated media are not available to some of the people who need them most. Further research into improving detection capabilities remains critical as well as ensuring the applicability of these tools to a range of real-world users and ensuring that those who access tools also have the knowledge and skills to use them.

Recommendations: How Congress can address technical advances in deepfake technology

Significant evolutions in the volume of creation, ease of access, and personalization of deepfake technologies bring the potential for creativity but also mounting threats. I have highlighted in this statement the need to focus on how advances in deepfake technologies relate to existing harms as identified by those directly impacted or at heightened risk.

I encourage this Subcommittee and legislators to support responsible detection and provenance approaches that protect privacy and free expression, as well as to consider targeted legislation on known harms, including non-consensual synthetic sexual and intimate images.

---

[60] João Phillipe Cardenuto et al., *The Age of Synthetic Realities: Challenges and Opportunities*, submitted APSIPA Trans. on Signal and Information Processing, June 2023, https://arxiv.org/abs/2306.11503 ; Nicola Jones, *How to stop AI deepfakes from sinking society - and science*, Nature Vol. 621, 28 September 2023 https://www.nature.com/articles/d41586-023-02990-y

[61] 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/

In particular, I urge Congress to consider:

1. **Targeted legislation on immediate harms including non-consensual synthetic sexual and intimate imagery**: There is an opportunity for Congress to lead on a federal approach to to synthesized sexual and intimate images, as this existing problem grows in scope. Consultation with groups disproportionately experiencing the existing and emerging harms from deepfake technologies, as well as key civil society actors, should help you craft appropriate steps in this area, as well as around child sexual abuse material (CSAM). Deepfake technology also forces us to think about our "likeness" rights, and how to ensure robust protections for speech, including satirical use, while protecting individuals from commercial and personal abuse.

2. **Provenance and disclosure**: People will soon not be able to 'spot' a wide range of deepfake content with their eyes or ears, so solutions are needed to proactively add and show the provenance of AI content, and, if the user opts for it, human-generated content too. Visible or audio cues to the viewer such as a disclaimer or a label are unlikely to be sufficient. Other complementary techniques disclose reliable information on the 'recipe' of AI-made media that can be accessed easily and shown to the user. These disclosure approaches should provide information to consumers, and not be a punitive measure to "single out" AI content nor be understood to indicate deception. Critically, these approaches should protect privacy, and not collect by default personally-identifiable information. For content that is not AI generated, we should be wary of how any provenance approach can be misused for surveillance and stifling freedom of speech.

3. **Detection**: Alongside indicating how the content we consume was made, there is a continued need for after-the-fact detection for content believed to be AI-generated. From WITNESS's experience, the skills and tools to detect AI-generated media remain unavailable to the people who need them the most: journalists, rights defenders and election officials domestically and globally. It remains critical to support federal research and investment in this area to improve detection overall and to close this 'detection equity' and access gap.

Lastly, these measures to manage advances in deepfake technologies would be supported by bigger picture legislation related to both AI governance and data privacy.

- **Pipeline responsibility:** For the public to understand how deepfake technologies are used in the media we consume, and for provenance and detection approaches to be effective, we need a pipeline of responsibility. Accountability should be distributed across technology actors involved in the production of AI technologies

more broadly, including from the foundation models to those designing and deploying software and apps, and to platforms that disseminate content.

- **Data privacy legislation**: AI legislation should incorporate strong privacy protections, and Congress should consider comprehensive data privacy legislation. These safeguards will not only support the responsible implementation of legislation on transparency in media production, but also help mitigate existing and future harms.

Sam Gregory

Executive Director