Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Submitted on: February 2, 2024

Submitted by: WITNESS

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

INTRODUCTION

WITNESS is an international human rights organization that helps people use video and technology to protect and defend their rights.[1] Our Technology Threats and Opportunities Team engages early on with emerging technologies that have the potential to enhance or undermine society's trust in audiovisual content. Building upon years of WITNESS' foundational research and global advocacy on synthetic media, we've been preparing for the impact of artificial intelligence (AI) on our ability to discern the truth.

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 citizens and communities in the US and globally, and to prepare accordingly.[2] These efforts have included contribution to the development of technical standards,[3] pioneering work facilitating real-time analysis of suspected deepfakes that can have important consequences for democracy and human rights,[4] policy input to technology companies and legislators,[5] experimentation with generative AI tools for human rights advocacy,[6] public education and advocacy,[7] US congressional testimonies,[8] and in-depth consultations with activists, journalists, content creators, technologists and other members of civil society.[9]

In response to the US National Institute of Standards and Technology's (NIST) *Request for Information to assist in carrying out several of its responsibilities under the Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* issued on October 30, 2023, we have divided our submission into two sections.

In the first section we overview a number of risks with current approaches to AI transparency, including indirect disclosure mechanisms such as watermarking, fingerprinting, and signed metadata. We also highlight the importance of centering the experience of those at the frontlines of human rights and democracy. In the second section we outline risks and limitations of current AI detections tools and share what we have learned through our experience working with leading AI detection experts.

---

[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] 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/
[5] Sam Gregory, Raquel Vazquez Llorente, Regulating Transparency in Audiovisual Generative AI: How Legislators Can Center Human Rights, Tech Policy Press, October 2023, https://www.techpolicy.press/regulating-transparency-in-audiovisual-generative-ai-how-legislators-can-center-human-rights/
[6] shirin anlen and Raquel Vazquez Llorente, Using Generative AI for Human Rights Advocacy, WITNESS, June 2023, https://blog.witness.org/2023/06/using-generative-ai-for-human-rights-advocacy/
[7] Sam Gregory, When AI can fake reality, who can you trust? TED Democracy, November 2023, https://www.ted.com/talks/sam_gregory_when_ai_can_fake_reality_who_can_you_trust
[8] U.S. Senate Committee on Commerce, Science, and Transportation, 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, https://www.commerce.senate.gov/services/files/DAD2163A-EF02-41B5-B7BA-2BA8B568C977
[9] 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

1. Risks with current approaches to AI transparency

The importance of centering the experiences of those protecting human rights and democracy

In this section we provide an overview of a number of risks and harms that can come from how audiovisual generative AI tools are developed and used, as well as mitigation recommendations. These tools, with their potential to create realistic image, audio and video simulations at scale, as well as personalized content, are having far-reaching implications for people in the US and globally. They can exacerbate existing problems such as the undermining of trust in the information people see and hear. This in turn risks eroding our democracy more broadly.

At the moment, emerging technologies including generative AI are being designed, developed, and deployed without the input of those who will be most impacted by them. Communities like human rights defenders and journalists are at the frontlines of democracy and human rights, and they play an essential role in fighting for and protecting these rights. For democracies to develop and flourish it is crucial that people can trust what they see and hear in critical government, business, and personal communications, as well as the documentation of events on the ground. It is also important that people are adequately informed about what they see and hear in order to realize the creative and innovative potential inherent in generative AI. When powerful technologies are developed without a comprehensive understanding of local and national contexts, individuals at the frontlines will inevitably face harm. Therefore, it is of utmost importance that the input of these communities should actively drive the development and inform the deployment of such technologies.

WITNESS' experience is 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. We have been researching the impact that deepfakes and other emerging technologies are having on frontline communities since 2018 and have run numerous workshops with these communities to hear about their experience and the solutions they are prioritizing. At these workshops, we have repeatedly heard how the main overarching concern for communities across countries is that threats from synthetic media will disproportionately impact those who are already at risk because of their ethnicity, gender, sexual orientation, profession, or belonging to a social group. 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. As we have seen with increasing frequency over the past year, women in particular already face widespread threats from non-consensual sexual images. These images do not require high-quality or complex production to be harmful. This reflects both the biases in the design of 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 have heard how the fear of synthetic media, combined with the confusion about its capabilities and the lack of knowledge to detect AI-manipulation, are exploited to dismiss authentic information with claims it is falsified.[10]

It is not realistic to expect the ordinary person 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. 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.

As media, communication and content production are becoming more and more 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.[11] Transparency approaches can also support better control for individuals and others on how their data is used in AI models.

Responsibility and risks of indirect disclosure mechanisms

Pipeline responsibility

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.[12] To ensure that people are able to understand the content they are seeing and hearing, all actors across the AI and media distribution pipeline need to be responsible for providing transparency. 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

As a starting point, it is important that approaches to transparency do not place the sole burden on ordinary people to identify AI-generated or edited content.

---

[10] U.S. Senate Committee on Commerce, Science, and Transportation, 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, https://www.commerce.senate.gov/services/files/DAD2163A-EF02-41B5-B7BA-2BA8B568C977

[11] 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/

[12] 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

Synthetic media transparency terminology

There is now a significant trend in AI governance towards a pipeline approach and a focus on developing technical approaches to providing transparency to content that was generated or edited using AI. For example, 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.[13] In October 2023, US President Biden released a comprehensive Executive Order concerning artificial intelligence that aims to advance the "safe, secure, and trustworthy development and utilization of artificial intelligence".[14] 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.[15] The EU AI Act includes significant requirements for disclosing deepfakes and machine-generated content from foundation models, including potentially requiring providers of certain types of AI systems to embed technical solutions that enable marking content that has been generated or edited by an AI system.

While there are a number of approaches to transparency, with each carrying its own risks and potential harms, there is significant and unhelpful confusion around the terminology used to describe these various approaches to AI-generated or edited content transparency.[16] The Partnership on AI's Glossary for Synthetic Media Transparency Methods provides definitions around a number of key transparency terms. WITNESS took part in a series of workshops that PAI ran and directly fed into the creation of this glossary. As policymakers including in Executive Order 14110 continue to explore methods for disclosing content that has been touched by AI, a harmonization of terminology is important.

PAI's Glossary for Synthetic Media Transparency Methods defines 'watermarking' in the following way[17]:

- Visible watermarks which are modifications made to a piece of synthetic content that are detectable to the human eye or ear and do not require the use of a watermark detector to interpret them.

[13] 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

[14] The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 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/

[15] 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

[16] 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/

[17] Partnership on AI, Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure, December 2023, https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

- Invisible watermarks which are modifications made to a piece of content that are imperceptible to the human eye or ear and can only be identified by a watermark detector.

Another transparency term that is used to show use of AI in content is 'fingerprinting'. Fingerprinting is defined in PAI's Glossary as the proactive process by which a hash is generated for a piece of content for the purpose of identifying that content at a later date. Such hashes must be stored in a database in order to verify future content against the original. Unlike watermarking, this hash is not embedded in the content file itself - and can come in two forms:

- Cryptographic hashing which is an exact-match form of hashing where the hash for a piece of synthetic content will not match if the content has been modified in any way.
- Perceptual hashing which is a probabilistic-match form of hashing where the hash for a piece of synthetic content is resilient to minor perturbations.

A final technological approach that is used to provide transparency around if a piece of content was AI-generated or edited is metadata. PAI's Glossary defines metadata as information about the origin, structure, and/or editing history of a piece of content that is proactively attached to the content itself. One approach to adding metadata is to sign it, which means that the information is proactively attached to AI-generated or edited content using cryptographic signatures. Another approach is to leave the metadata unsigned, which means that while information about how a piece of content was created is still proactively attached to the content, it is not validated by a trusted signer certificate.

Implementing a comprehensive approach to transparency will require a combination of these methods.

Synthetic media transparency methods and their risks

Visible and invisible approaches to watermarking

Visible watermarks can be useful in specific scenarios such as AI-based imagery or production within election advertising. However, they are often easily cropped, scaled out, masked or removed, and specialized tools can remove them without leaving a trace.

Current approaches do not take into consideration what type of information would make the ordinary person trust a visible watermark or the AI-generated or edited content. For example, in some contexts visible watermarks should provide more than just binary labeling of AI-generated or edited or human-created. They could include meaningful context about the content and help people to understand why AI was used in the content's production. However, visible (and/or audible) watermarks alone are inadequate for reflecting to people if and how AI was used in an image or video, especially as AI tools become more broadly adapted in our media environment.

Indirect disclosure mechanisms using invisible watermarking can provide information about how a piece of content was created with a particular dataset, technology, or tool. Invisible watermarks, like

Google's SynthID, 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; and they do not provide nuanced information on how media was edited beyond the initial creation with an AI-generative tool. 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.

Approaches that focus instead on leaving traces in the underlying data of a model—'radioactive data'—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.

Fingerprinting

Cryptographic and perceptual hashing are well-established techniques used to identify and track digital content. These methods are instrumental in platforms like YouTube, through its ContentID system, and are also used in databases to identify instances of Terrorist and Violent Extremist Content (TVEC) and Child Sexual Abuse Material (CSAM). Additionally, these techniques can link synthetic media to its source data—for instance, as in the C2PA approach outlined below, to reconnect a file to its related metadata. When it comes to verifying the authenticity of media post-creation, especially media identified as synthetically generated, these hashing methods are employed to mark and track such content in databases. However, the reliability of these methods is subject to certain limitations, similar to those discussed in the section on detection methods below.

Metadata

Signed metadata-based standards such as the Coalition for Content Provenance and Authenticity (C2PA) specifications 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. Information about how and when a piece of content was created is added to the content when it is created and edited. This information is then accessible to others to view and acts as a way to verify the source and potential veracity of the content. 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, the C2PA specifications have been integrated into capture devices such as cameras, utilizing a process known as 'authenticated capture', which could be used to automatically record information about how and when a piece of audiovisual content was taken.

The C2PA specifications are currently being more broadly adopted. For example, Microsoft has been working on implementing this form of indirect disclosure to content generated using its AI systems,[18] and Adobe has started to provide it via its Content Credentials approach.[19] The use of cryptographic signature and metadata-based standards has also been recognized as an important risk-mitigating mechanism in emerging regulation, such as in the case of the Executive Order on AI, and especially in the context of elections, as evidenced by announcements from Microsoft and Open AI.[20]

WITNESS is a member of the C2PA, actively participates in the Technical Working Group, and currently acts as a co-chair of the Threats and Harms Taskforce. In this context, we have advocated for globally-driven human rights perspectives and practical experiences to be reflected in the technical specifications.[21] In WITNESS' experience, one of the main risks of indirect disclosure approaches to transparency is that global civil society and human rights organizations are not sufficiently consulted in how the technical standards are developed and deployed. One obstacle that may inhibit participation from civil society and human rights organizations is the notion that standards-setting processes are strictly technical–as opposed to technical processes imbued by interests and worldviews. This often results in spaces that value technically trained specialists and undermine experience from other stakeholders, including the same people that may be most affected by the standards being designed. To mitigate the risk of potentially harmful standards being broadly rolled out, we recommend that there be further development of targeted and thoughtful mechanisms that can bridge different lived, technical and professional experiences with the intention of upholding human rights concerns.

Risks exclusive to open source models

There are also challenges exclusive to open source generative AI models in designing and implementing visible and invisible watermarks. These include technical challenges with standardizing how different forms of synthetic media transparency methods could be applied to AI-generated or edited content, given the ability for people to tweak open source systems. This also opens up regulatory challenges around ensuring downstream accountability is possible.

[18] 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
[19] Adobe Content Credentials, https://helpx.adobe.com/creative-cloud/help/content-credentials.html
[20] See for example: Microsoft, Microsoft announces new steps to help protect elections, November 2023, https://blogs.microsoft.com/on-the-issues/2023/11/07/microsoft-elections-2024-ai-voting-mtac/ and OpenAI, How OpenAI is approaching 2024 worldwide elections, January 2024,: https://openai.com/blog/how-openai-is-approaching-2024-worldwide-elections
[21] The Coalition for Content Provenance and Authenticity, C2PA Harms Modelling, https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html

Economic and security implications of indirect disclosure mechanisms

Privacy

To safeguard US Constitutional and human rights, approaches to AI media transparency and indirect 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 (where used) 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. For example, it is not necessary to include certain data points, such as the country in which the person who generated a piece of content is located. Data points that could be helpful in illustrating *how* a piece of content was generated is information such as the software, model and version used to generate the content, as well as potentially the date and time the content was generated.

Since 2019, WITNESS has been raising concerns about the potential harms that could arise from the inclusion of personal data in metadata-based solutions to AI content transparency.[22] The US government has the opportunity to ensure that disclosure 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.

AI-generated and edited content will be edited and combined with human-generated content in media and communicative production. While a requirement to include signed metadata that indicates if content was AI-generated or edited could be a legal requirement in certain cases, metadata that reveals how content was changed over time should not extend to content created outside of AI-based tools, which should always be opt-in. We encourage NIST to create a specific task force to research the balance of fundamental rights like privacy more deeply, and to also focus on the intersection of human-generated content with AI-generated content.

Anonymity

Building trust in content must allow for anonymity and redaction. Immutability and inability to edit content do not reflect the realities of people, how and why media is made, or that certain redaction may be needed in sensitive content.[23] An immutable record, for example by writing records into the

[22] 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/
[23] Raquel Vazquez Llorente, Trusting Video in the Age of Generative AI, Commonplace, June 2023,

blockchain or by blocking changes to digital objects, may not allow for the ability to blur someone's face whose identity may need to be protected.[24] A permanent record cannot be deleted, whereas immutable ones can. 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.[25] While specifications like the C2PA focus on protecting privacy and don't mandate the inclusion of people's identities or personally identifiable information, this privacy requirement needs to be protected during widespread adoption. We should be wary of how these transparency 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.[26] Signed metadata for both AI and user-generated content provides signals—i.e. additional information about a piece of content—but does not prove truth. These signals should be complementary to other processes of digital and media literacy that consumers choose to use, such as seeking information from a variety of sources, to help them triage questions they may have, and that are available to other parties engaging with the content, including potentially platforms. 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.[27]

2. Risks and limitations of current AI detection tools

To reflect the objectives of the Executive Order 14110 of ensuring the AI is developed and used safely, it is crucial to examine the risks of current tools aimed at helping to detect AI-generated or edited content. In the previous section we outlined the risks of current approaches to AI transparency, in particular methods of indirect disclosure such as watermarking, fingerprinting, and signed metadata. In this section we provide an overview of our experience of the current limitations of detecting synthetic content.

---

https://commonplace.knowledgefutures.org/pub/9q6dd6lg/release/2

[24] See for instance: https://docs.aws.amazon.com/AmazonS3/latest/userguide/object-lock.html

[25] 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

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

[27] U.S. Senate Committee on Commerce, Science, and Transportation, 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, https://www.commerce.senate.gov/services/files/DAD2163A-EF02-41B5-B7BA-2BA8B568C977

AI detection tools

AI detection tools allow people to run a piece of content through it and receive information about the likelihood this material had been generated or edited by an AI system. As such, these tools could play an important role in the broader solution and help to ensure that people are able to understand the context of the content they are consuming, as well as counter false claims that human-generated content is AI-generated.

Detection tools are also important for content believed to be AI-generated that does not have signed metadata or that has been manipulated with counter-forensics approaches.

Risks and limitations of current AI detection tools

However, existing detection of audiovisual generative AI and deepfakes outputs is flawed. Existing detection models frequently require expert input to assess the results. Often they are not generalizable across multiple synthesis technologies and techniques. They can also require personalization to a particular person in order to more accurately protect people from fraudulent voices or imagery. As such, detection tools can lead to unintentional confusion and exclusion. WITNESS has seen how the use 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.[28]

Lack of access to detection tools

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

In our experience, detection tools are a critical element—alongside the incorporation of signed metadata and media literacy—when it comes to real-world scenarios where journalists, civil society, and ordinary citizens are attempting to discern how content has been created and edited.

WITNESS is currently piloting a Deepfake Rapid Response Force that allows International Fact-Checking Network members to escalate cases of suspected deepfakes, and get a timely assessment on the authenticity or origin of the content.[29]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 and civil society leaders, 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

[28] 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/

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

'detection equity' gap that exists—the tools to detect AI-generated media are not available to those at the frontline of democracy and human rights and who need these tools the most.

Limitations of current AI detection models

Since we began operating the Deepfake Rapid Response Force, we have had nearly 30 cases referred to us. Recently, we have seen increasing numbers of audio cases, and these cases have proven more challenging for the Force to analyze. One reason for this is because it can require Force members with audio expertise to first train their models on authentic audio of the person speaking in the alleged deepfake before being able to run an analysis of the potentially deepfake audio clip. The models also tend to be trained primarily in English, and analysis of audio clips in other languages have proven difficult.

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

Reverse image and video search helps discover visually similar images or videos from around the web. Since content tends to spread across different platforms, companies should also develop and support infrastructure that allows people to cross-check audiovisual content across a number of platforms simultaneously, as this functionality would allow people to track how content is shared across different platforms. While a handful of reverse image search tools exist, those more available are largely optimized towards commercial applications, such as online shopping or protecting copyright, rather than curbing mis- and disinformation.[30] At present, no widely accessible tools exist that allow people to conduct reverse video searches, although research is underway.[31] Easy to use reverse video search would allow people to, in effect, simply click a button and conduct a search to see where a video was originally posted and how it has been shared or edited over time.[32]

[30] See for example Google Lens: https://lens.google/
[31] Most recent research is available at Visual Copy Detection Workshop, 2023, https://sites.google.com/view/vcdw2023/
[32] Sam Gregory, Shallowfakes are rampant: Tools to spot them must be equally accessible, The Hill, August 2022, https://thehill.com/opinion/technology/3616877-shallowfakes-are-rampant-tools-to-spot-them-must-be-equally-accessible/