WITNESS—written evidence (LLM0050)

House of Lords Communications and Digital Select Committee inquiry: Large language models

About us: WITNESS is an international human rights organisation that helps people use video and technology to protect and defend their rights. 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. In consultation with civil society, academia and industry on four continents, we've identified the most pressing concerns about how deepfakes, synthetic media and generative AI are impacting the information ecosystem and society at large. As part of this process, we have also developed guidelines for principled action and recommendations to policy makers, technology companies, regulators and other stakeholders.

Summary

This submission responds to questions 2 (*What are the greatest opportunities and risks over the next three years?*) and 5 (c) (*How can the risk of unintended consequences be addressed?*), and puts forward a set of recommendations, based on our long standing work with human rights defenders, journalists, content creators, fact-checkers and technologists, to guard against the risks of large language models over the coming years. We specifically focus on the use of generative AI tools to produce audiovisual content. Our submission is informed by three decades of experience helping communities create trustworthy photo and video for human rights advocacy, protect themselves against the misuse of their content, and challenge mis- and disinformation that targets at-risk groups and individuals.[1]

We begin by explaining three overarching principles that should guide the assessment of the opportunities and risks of generative AI tools and the underlying large language models upon which they are built. These principles are: (1) centre those who are protecting human rights and democracy at the frontlines in the development of solutions, (2) place firm responsibility on stakeholders across the AI, technology and information pipeline; and (3) embed human rights standards, laws and practices in the development of technical solutions.

We then detail our proposals for addressing the opportunities and risks of large language models, particularly from the perspective of how AI-generated or edited audiovisual content can impact people's trust in such content. Our

---

[1] WITNESS https://www.witness.org/ For our work on emerging technologies, mis- and disinformation, see: https://www.gen-ai.witness.org/

recommendations are to: (1) encourage transparency in the production of AI content while balancing privacy and risks to users; (2) promote equity in access to solutions that can detect AI-generated or edited content; and (3) push for more investment and user-friendly tools that can assist with the verification of online audiovisual material.

We are grateful for the opportunity to provide evidence to the UK Communications and Digital Committee on Large Language Models.

OVERARCHING PRINCIPLES TO GUIDE THE ASSESSMENT OF THE OPPORTUNITIES AND RISKS OF LARGE LANGUAGE MODELS

1. Since 2018, WITNESS has been leading the first global effort to understand how deepfake technology is impacting communities at the frontlines of mis- and disinformation by convening leading human rights defenders, journalists, content creators, fact-checkers, technologists and other members of civil society across Africa, Brazil, Europe, South East Asia and the United States.[2] Over the past year, building upon these consultations and foundational research, we have also incorporated an analysis of the opportunities and risks of large language models and the generative AI tools that are built using these models.[3]

2. In deep collaboration with these stakeholders, we have identified three overarching principles that should guide the assessment of the opportunities and risks that generative AI brings to society.

Centre those who are protecting human rights and democracy at the frontlines in the development of solutions

3. With hyperbolic rhetoric undermining trust in visual media, human rights defenders, journalists and civil society actors collecting trustworthy information or debunking falsehoods will be among the most impacted by generative AI. Yet, emerging technologies are designed and deployed without the input from these groups, ignoring the threats and risks these technologies bring to communities already at a disadvantage or most affected by harms such as mis- and disinformation. Most importantly, many proposals fail to acknowledge the solutions that those who bear the burden of

---

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

[3] 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/

combating these threats and harms prioritise.[4] While these technologies originate primarily in western countries including the UK, they affect people globally. When they are developed, deployed, or regulated without an in-depth understanding of other local and national contexts, the people at the frontlines will be affected, regardless of their location. This is why it is crucial that the development and deployment of AI centres the voice of these communities.

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

4. All actors across the AI pipeline have a duty to insert safeguards and proactively address the harms their work may bring. These include:

   - those researching and building foundational large language models;
   - those commercialising generative AI tools that sit on top of large language models (such as text-to-image or text-to-video tools that allow users to describe an image or video they would like produced, and have the AI system generate it);
   - those creating synthetic media; and
   - those publishing, disseminating or distributing synthetic media (such as media outlets and platforms).

5. While strong investment in media literacy is crucial, the responsibility cannot be left solely on end-users to determine if the audiovisual content they are consuming is AI-generated or manipulated, as well as the larger context of such content.[5] There are existing frameworks, such as the Partnership on AI's Responsible Practices for Synthetic Media Framework ('the Framework'), which offers guidelines for developing, creating, sharing, and publishing synthetic media ethically and responsibly.[6] WITNESS was part of the Framework process from its beginning, helping guide the recommendations towards those building technology and infrastructure for synthetic media, those creating synthetic media, and those distributing or publishing synthetic media. This type of principles and frameworks, that are built on longstanding

---

4 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

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

6 Partnership on AI, R*esponsible Practices for Synthetic Media Framework.* https://syntheticmedia.partnershiponai.org/

engagement with these issues, could offer foundational guidance for new UK policies.

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

6. Legislation, regulation and other norms, as well as company policies and technical infrastructures, should have human rights standards baked in—especially as satire, art and other forms of creative expression test the boundaries of existing legislation and policies.[7] The Coalition for Content Provenance and Authenticity (C2PA), which is developing technical specifications to make it easier to identify how, where and by whom a piece of media may have been created, and the modifications it may have undergone while disseminated, is an example of how human rights standards can inform the develop of technical specifications. As a co-chair to the C2PA Threats and Harms Taskforce, WITNESS has successfully advocated for globally-driven human rights perspectives and practical experiences to be reflected in the technical standard.[8]

7. 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.[9] While these approaches can help journalists, activists, human rights defenders and others to ensure societies are able to ascertain how a piece of content was created, and if and how it has been modified; they can also lead to potential harms to a broad range of individuals and communities, especially those at the frontlines of defending human rights.[10] For instance, governments could require provenance schemes that capture personally identifiable information to augment surveillance and stifle freedom of expression.[11]

8. People using generative AI tools to create audiovisual content should not be required to forfeit their right to privacy to adopt these emerging technologies. The UK government has the opportunity to ensure that provenance requirements and standards are developed in-line with global

---

7 shirin anlen and Raquel Vazquez Llorente, *Using Generative AI for Human Rights Advocacy* (June 2023) https://blog.witness.org/2023/06/using-generative-ai-for-human-rights-advocacy/ See also: WITNESS, *Report: Just Joking! Deepfakes, Satire and the Politics of Synthetic Media* (2022) https://cocreationstudio.mit.edu/just-joking/

8 Jacobo Castellanos, *WITNESS and the C2PA Harms and Misuse Assessment Process.* WITNESS (2021) https://blog.witness.org/2021/12/witness-and-the-c2pa-harms-and-misuse-assessment-process/ ; The Coalition for Content Provenance and Authenticity, *C2PA Harms Modelling*. https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html

9 Ibid.

10 Sam Gregory, *Tracing trust: Why we must build authenticity infrastructure that works for all.* WITNESS (2020) https://blog.witness.org/2020/05/authenticity-infrastructure/

11 List of potential harms of the C2PA specifications: https://c2pa.org/specifications/specifications/1.0/security/_attachments/Due_Diligence_Actions.pdf Also see: Gabrielle Lim and Samantha Bradshaw, *Chilling Legislation: Tracking the Impact of "Fake News" Laws on Press Freedom Internationally.* Center for International Media Assistance (July 2023) https://www.cima.ned.org/publication/chilling-legislation/

human rights laws, do not include the automated collection of personal data, and are able to be opted-in or out of. Further, the tools should be built with accessibility in mind, and in a way that allows all levels of technical knowhow to opt-in or out and have people's identities protected.[12]

9. In March 2023, WITNESS highlighted these points in our response to the Office of the United Nations High Commissioner for Human Rights' call for input on the relationship between human rights and technical standard-setting processes for new and emerging digital technologies.[13]

RECOMMENDATIONS TO ADDRESS THE OPPORTUNITIES AND RISKS OF LARGE LANGUAGE MODELS

10. With the above guiding principles in mind, WITNESS has developed three recommendations that can help the UK promote a beneficial deployment of large language models, and position itself as a world-leading centre for AI safety.

Encourage transparency in the production of AI content while balancing privacy and risks to users

11. The audiovisual content we consume is increasingly being touched by AI. In a world with wider access to AI tools that simplify the generation or edition of photos, videos, and audio, it is important for the public to be able to understand if a piece of media was created or altered using AI. In July 2023, seven leading AI companies agreed 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.[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]

---

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

13 WITNESS, *Submission to call for input: The relationship between human rights and technical standard-setting processes for new and emerging digital technologies* (2023) https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-csos-witness-4-input.pdf; Full report by The Office of the United Nations High Commissioner for Human Rights, *Human rights and technical standard-setting processes for new and emerging digital technologies* (June 2023) https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/sessions-regular/session53/advance-versions/A_HRC_53_42_AdvanceUneditedVersion.docx

14 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/

12. WITNESS understands the term *disclosure* to refer to the process of communicating transparently and effectively about image and video synthesis and manipulation. The Partnership on AI's Responsible Practices for Synthetic Media Framework describes direct forms of disclosure as 'visible to the eye', and include methods such as applying a visible label marking the content as AI-generated, adding disclaimers, and watermarking AI-generated content.[16] Indirect forms of disclosure are embedded text, image, or other information in AI-generated digital content in a way that is imperceptible to the human eye.[17] The Framework provides examples such as applying cryptographic provenance to generated content (e.g. C2PA standard), embedding traceable elements to training data and the content generated, or adding single-frame disclosure statements in videos.

13. Disclosing if content was generated is one way to foster a safer AI ecosystem. However, there are a number of considerations when developing and deploying disclosure systems. Using direct forms of disclosure, such as labels or watermarks, to signal explicitly to viewers that they are looking at AI-generated or edited content can help people understand what they are consuming. However there are limitations to this approach.[18] For example, these markings tend to be small and easily missed, and there is not necessarily always space to provide meaningful context on how the media was created or why. Further, when a piece of media is distributed across politicised and closed social media groups, its creators lose control of how it is framed, interpreted, and shared even when it had been originally labelled.

14. Solutions that track the provenance of media, including watermarking, should be understood as providing signals of trust, but not as an absolute confirmation of the 'truth'. While provenance data can help verify a piece of media, ascertaining the truth still requires further analysis of the actual content. Using labels, watermarks and other provenance data to deliver binary results about the truthfulness of content can have catastrophic consequences for communities whose content's authenticity and integrity may be on the line.[19]

15. WITNESS has been researching the sociological impacts of different forms of

---

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 Partnership on AI, R*esponsible Practices for Synthetic Media Framework.* https://syntheticmedia.partnershiponai.org/

17 The most recent example is SynthID, released by Google on August 29 (2023) https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid

18 Katerina Cizek, shirin anlen, *The Thorny Art of Deepfake Labeling*. WIRED (May 2023) https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/ ; Sue Halpern, *Will Biden's Meetings with A.I. Companies Make Any Difference?* The New Yorker (July 2023) https://www.newyorker.com/news/daily-comment/will-bidens-meetings-with-ai-companies-make-any-difference

19 Raquel Vazquez Llorente, *Trusting Video in the Age of Generative A*I. Commonplace (June 2023) https://commonplace.knowledgefutures.org/pub/9q6dd6lg/release/2

watermarks. We are currently exploring how different types of watermarks applied to audiovisual media can impact people's trust in online content, the accessibility of different forms of watermarks, and where responsibility to apply watermarks lies. At a time when policy makers and companies are focusing on how to *technically* implement watermarks, we have chosen to examine three key socio-technical areas that are crucial to the success of such endeavours and are of equivalent importance to the ongoing technical discussions, yet remain under-explored. We are able to provide the Committee further evidence about these points, orally or in writing, if desired.

Promote equity in access to solutions that can detect AI-generated or edited content

16. Detection tools may 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 can play an important role in the beneficial deployment of generative AI, and in mitigating risks. However, existing detection tools require expert input to assess the results and are not generalisable across multiple synthesis technologies and techniques. As such, detection tools can lead to unintentioned confusion and exclusion. For example, we have seen how the use by the general public of detection tools has contributed to increased doubt around real footage, rather than contributing to clarity.[20] Further, detection tools need to be trained on data related to the scenarios in which they are deployed in order to provide useful insights.

17. 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. Companies and governments should support further research into improving detection capabilities, and ensure that those who need them also have the knowledge and skills to use them. To contribute to this effort, WITNESS is currently piloting a Deepfakes Rapid Response Force that connects members from the International Fact-Checking Network (IFCN) with world-leading experts in media forensics and AI generation and manipulation. The Force analyses some of the most difficult deepfake cases and claims of deepfakes, and provides a timely assessment to local journalists and fact-checkers.[21]

Push for more investment and user-friendly tools that can assist with the

---

20 Sam Gregory, *Pre-Empting a Crisis: Deepfake Detection Skills and Global Access to Media Forensics Tools*. WITNESS (2021) https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/ ; Sam Gregory, *The World Needs Deepfake Experts to Stem This Chaos*. WIRED (June 2021) https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/

21 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/

verification of online audiovisual material

18. Reverse image and video search helps discover visually similar images or videos from around the web. Companies should invest in the implementation of platform-level intuitive reverse image and video search so that people can understand how content is circulating on a platform. 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 optimised towards commercial applications, such as online shopping or protecting copyright, rather than curbing mis- and disinformation.[22] At present, no widely accessible tools exist that allow people to conduct reverse video searches, although research is underway.[23]

19. Although AI-generated content is an emerging technology, the threats it poses are not new. In WITNESS' years of organising global workshops, a primary concern that has arisen repeatedly is the mis-contextualization, mis-attribution, or simple editing of video and audio circulated on social media platforms ('shallowfakes'). Accessible reverse video search would allow people to, in effect, simply click a button and conduct a search to see where a video was originally posted and how it has been shared or edited over time. This type of functionality would allow researchers to better train detection tools and also allow a less technical audience to benefit from the tools.[24]

20. Most of the cases brought to our Deepfakes Rapid Response Force were not escalated to the media forensics experts because the content was either mis-contextualised or an unsophisticated manipulation, rather than an example of technically complex media. This powerfully illustrates one of the reasons why WITNESS advocates for platforms and messaging apps to support further research, development, and deployment of more accessible tools that can explain and contextualise shallowfakes. WITNESS is available to provide more evidence, orally or in writing, about how image and video copy detection can advance human rights and help combat mis- and disinformation.

5 September 2023

---

22 See for example Google Lens: https://lens.google/

23 Most recent research is available at Visual Copy Detection Workshop (2023) https://sites.google.com/view/vcdw2023/

24 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/