WITNESS—written evidence (FEO0070)

House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online

WITNESS (witness.org) works closely with human rights defenders, ordinary people and civic journalists worldwide who use video and technology to protect and defend human rights. Our work is global in scope including work in the US, Europe, Latin America, Sub-Saharan Africa, Middle East and North Africa and South and Southeast Asia. We collaborate with human rights defenders and civic journalists to develop and share knowledge about how to safely, ethically, and effectively produce and utilize their own video, as well as archive and organize video created by others. Driven by our understanding of what is happening in a range of high-risk global communities, we advocate at a systems level to technology companies, multi-stakeholder bodies and regulators to enable greater freedom of expression and other rights, to make existing tools and policies work better for human rights defenders and all users, and to ensure emerging technologies and policies reflect global human rights values and needs. This work impinges -- among others -- on these critical questions of the Inquiry:

6. To what extent should users be allowed anonymity online?
7. How can technology be used to help protect the freedom of expression?
8. How do the design and norms of platforms influence the freedom of expression? How can platforms create environments that reduce the propensity for online harms?
10. How can content moderation systems be improved? Are users of online platforms sufficiently able to appeal moderation decisions with which they disagree? What role should regulators play?

In this submission we provide input related to our work on

- Visual anonymity online

- Emerging authenticity infrastructure and its implications for freedom of expression

- Emerging audiovisual manipulation trends, e.g. deepfakes and implications for freedom of expression

- Platform responsibilities to create ‘evidence lockers’ to preserve critical human rights evidence in the light of human and algorithmic content moderation decision-making

Visual anonymity and freedom of expression online

1. WITNESS’s experience with human rights defenders globally indicates the critical importance of allowing both anonymity and persistent pseudonymous identity online. Widespread research on the implications of enforcing real name policies, and their inefficacy (see Jillian C York summary, 2021).[1] Critical to an active right to both free expression and to privacy is the right to communicate anonymously. Of course, this is not an absolute right – after all, anonymity can also be used, for example to cover criminal activity. However, the active presence of options to have anonymity and no a priori restrictions on anonymity enables freedom of expression and supports the right to privacy. WITNESS’s own focus has been on the importance of promoting and supporting options for visual anonymity online. By visual anonymity we mean the ability to more easily control how images of your face are shared and distributed online, recognizing the growing dominance of audiovisual-centric platforms.

2. A right to visual anonymity draws on the rationales in human rights standards for the need for anonymity for participation by at-risk individuals, but notes that the implementation of this right is challenging in a world of persistent faceprints and increasing facial recognition (see Human Rights Made Visible,[2] Gregory 2012). ‘Visual anonymity’ may sound like a contradiction in terms, but people often wish to speak out and to “be seen” on primarily visual communications platforms but still wish to conceal their face, the background of their home or not expose others without their consent. With increasing facial recognition, they should not be faced with the challenge that any expression of something politically unpopular, whistle-blow or speak out in some other manner, can be correlated – whether they like it or not - to the other 99 percent of their online and offline identity. Policy discussions on freedom of expression should consider how the lack of informational self-determination over how that one’s faceprint relates to discussions on regulation and banning of facial recognition. Platforms also have a responsibility to provide tools to enable visual anonymity -- as YouTube[3] and Signal[4] have done.

Responses to misinformation and freedom of expression: Authenticity Infrastructure

3. WITNESS has focused extensively on questions of how technology is used to protect freedom of expression and high public-interest credible, trustworthy information, particularly in the face of online and offline misinformation and disinformation. Our report ‘Ticks or it Didn’t Happen: Confronting Key Dilemmas in Building Authenticity Infrastructure for Multimedia’[5] explores

[1] https://jilliancyork.com/2021/01/14/everything-old-is-new-part-2-why-online-anonymity-matters/

[2] https://www.researchgate.net/publication/338518491_Human_Rights_Made_Visible_New_Dimensions_to_Anonymity_Consent_and_Intentionality

[3] https://www.wired.com/2016/02/visual-anonymity/

[4] https://signal.org/blog/blur-tools/

key dilemmas and trade-offs around privacy, freedom of expression and inclusion in relation to emerging proposals and technical infrastructure for more robust ‘authenticity infrastructure’ for multimedia: technical mechanisms for normalizing the tracking the origins, manipulation and editing of multimedia as one proposed solution to problems of trust in online expression.

4. Until recently discussion and technical development in this area was relatively niche. However, this year a number of initiatives have launched that aim to develop more widely shared technical standards for tracking media provenance and authenticity. Authenticity infrastructure includes tools for capture and media origins, for tracking edits and manipulations and for maintaining this information once media is published and shared. Commercial and non-profit ‘verified capture' tools and apps provide additional markers of indexicality to a photo or video shot on a mobile device as well as indications if manipulation has occurred. Authenticity infrastructure for tracking manipulations and edits to media and for providing data at point of publication and sharing include initiatives such as the Content Authenticity Initiative[6] (CAI) initiated by Adobe, New York Times and Twitter. The CAI looks at how original content, manipulations, changes, edits and additions can be tracked in a shared standard for photos and video, and includes stakeholders from the smartphone witnessing context including TruePic and WITNESS. Approaches to tracking provenance and attribution of mainstream media images include Project Origin from a consortium including the BBC, CBC, Facebook, First Draft, Google/YouTube, Microsoft, Reuters, The Hindu, Twitter and others looking at sustaining attribution of mainstream media content. WITNESS participated in the development of the initial White Paper for the Content Authenticity Initiative emphasizing key tradeoffs from a global, human rights perspective[7] and advocating for such features as low technical barriers to entry, no requirement for identity as a basis of trust and inclusion of global, high-risk contexts and abusability of the framework as key concerns.

5. WITNESS has seen the value of these mechanisms for enhancing trust in footage from our own experience building tools in this area, and from the ongoing needs of human rights defenders and civic journalists to defend the integrity of their content. We therefore approach the development of authenticity infrastructure as having benefits for high public interest, high-risk expression and in countering mis/disinformation while being acutely conscious of the risks as these approaches move from niche technologies used by journalists and rights defenders into the public tech infrastructure.

6. Authenticity infrastructure has important implications for which online accounts will be trusted and which not, and on whom the burden of proof is

[5] https://lab.witness.org/ticks-or-it-didnt-happen/

[6] http://contentauthenticity.org/

[7] https://blog.witness.org/2020/05/authenticity-infrastructure/

increased to prove untampered, show origins or confirm manipulation. Who is "incidentally" – via technical barriers or access questions -- or deliberately excluded or chilled? Intertwining authenticity claims and trust with access to newer tech, good battery life, GPS or connectivity can compound existing information inequities.

7. As with all infrastructure, we must be alert to how provenance architecture could be weaponized to delegitimize the accounts of people who must make complicated choices about visibility and invisibility, anonymity and pseudonymity on a case by case basis when they participate online. For example, human rights activists must navigate life-or-death decisions about affiliating themselves to footage that may show evidence of violations and navigate a 'friction between dual desires for visibility and obscurity'.[8]

8. Authenticity infrastructure choices on whether to make either persistent identity or providing an identity as part of validating trustworthiness and integrity of online speech similarly reflect many of the existing tensions around persistent real-name identity that have led to critical dissident and human rights voices being excluded from platforms like Facebook. One critical element that WITNESS pushed for in our involvement in the Content Authenticity Initiative CAI[9] was that identity not be essential to the infrastructure of authenticity, and also that selective clearly indicated redaction of key audiovisual data, for example the ability to blur faces but protect other authenticity and context data, was critical as a part of any standard. It is critical that as the Inquiry considers technical responses to protect and enhance freedom of expression it focuses on preserving options for pseudonymity and anonymity and ensuring emerging infrastructure options do not compromise this.

9. It is well-observed that new technologies often produce so-called 'ratchet effects' in terms of expectations of usage over previous iterations. In the case of authenticity infrastructure these ratchet effects may both discredit individual users who, for reasons of technology, GPS, online access, are unable to use authenticity infrastructure -- as well as disadvantage smaller media outlets, community journalism and others who are unable to adopt these approaches as rapidly. Expectations of provenance and of increased technical markers of authenticity must not be leverageable against vulnerable populations that cannot or choose not to use them. The possibility of an 'implied falsehood effect' (a version of the 'implied truth effect' where content not labelled as falsehood is considered true) is a risk if authenticity infrastructure is not accompanied by public education and media literacy efforts, and if it is developed as a default or defacto or legal obligation.

[8] https://twentysix.fibreculturejournal.org/fcjmesh-005-technology-and-citizen-witnessing-navigating-the-friction-between-dual-desires-for-visibility-and-obscurity/

[9] https://blog.witness.org/2020/08/adobe-content-authenticity-initiative-approach-authenticity-infrastructure-media-manipulation/

10. A particular responsibility falls on democratically-elected legislatures to consider how authenticity infrastructure may perpetuate two trends. One is the growth of surveillance capitalism premised on increasing amounts of personal data at the intersection of mobile platforms, audiovisual media and private platforms mediating: although initial efforts at authenticity infrastructure are not premised on sharing increased data into data-mining efforts, this is a possibility that is very plausible. Secondly, legislators should be aware of the possibility of 'legislative opportunism' that will take their approaches to regulating free speech and mitigating misinformation and disinformation in democratic contexts and adopt 'fake news' laws and regulations that disadvantage human rights speech. This is the trend we are seeing already globally and should be of concern in any integration of authenticity infrastructure to a greater degree. Infrastructure possibilities are potentially appealing to governments given the increasing range of these 'securitized' fake news laws (see Gabriela Lim, 2020)[10] that articulate arguments for speech control in terms of national security or public health infrastructures that can confirm who is responsible for 'rumours', 'hoaxes' or dissident speech. It is not an unrealistic leap of imagination to see how this type of authenticity infrastructure could be weaponized via the legislative opportunism of 'fake news' laws against journalists and dissidents to impose requirements for identity, data disclosure, and required authenticity infrastructure signals to post.

Content moderation and the need for preservation of critical human rights content

11. The broad area of how commercial content moderation disadvantages and harms vulnerable users, marginalized populations and human rights defenders is an area of critical concern to WITNESS, where we have seen the impact of hate speech, incitement to violence and misinformation permitted while legitimate speech and human rights defenders have their content and accounts removed or suspended (see our submission[11] to the UN Special Rapporteur on Freedom of Expression in regard to Content Moderation in the Digital Age). Human rights speech is highly unstable on commercial platforms, particularly when it occurs outside the US and Europe where platform action is inconsistent, under-resourced and subject to limited appeal. Foundationally we see the need for content moderation approaches to be based in international human rights standards, with consistency in approach and transparency in policy, process and appeals (in line with the Santa Clara Principles).[12] In the context of this inquiry, WITNESS will focus on one dimension of this - the need for better mechanisms for preservation of critical human rights content that is removed by algorithms or by human oversight from platforms.

[10] https://datasociety.net/library/securitize-counter-securitize/

[11] https://www.ohchr.org/Documents/Issues/Opinion/ContentRegulation/Witness.pdf

[12] https://santaclaraprinciples.org/

12. Recent events in relation to the loss of large amounts of online footage on YouTube showing potential war crimes in Syria (documented on an ongoing basis by the human rights group Mnemonic),[13] as well as the challenges of international fact-finding and judicial bodies in accessing Facebook content related to crimes in Myanmar illustrate the importance of ensuring that any legislative response to platforms include approaches to ensure that legitimate actors with a public interest in understanding critical issues that are documented on platforms have access to relevant content that has been taken down from those platforms through their content moderation processes. WITNESS and other human rights groups have described this in terms of 'evidence lockers' for critical footage that may otherwise be both correctly and incorrectly be removed under platform policies and relevant legislation as being in violation of policies on terrorist or violent extremist content or of policies on graphic violence (see also Alexa Koenig, Big Tech Can Help Bring War Criminals to Justice: Social Media Companies Need to Preserve Evidence of Abuse,[14] 2020 and Human Rights Watch 'Video Unavailable: Social Media Platforms Remove Evidence of War Crimes').[15] This social media content is of critical importance for accountability in a growing number of human rights scenarios and it is incumbent on civil society, platforms and government to collaborate to identify appropriate models and mechanisms for preserving it.

Responses to misinformation and freedom of expression: Deepfakes

13. WITNESS has led one of the leading global efforts to prepare better for the threat of emerging forms of audiovisual manipulation that make it harder to discern manipulated and synthesized video, audio and text from real. Popularly known as 'deepfakes', these forms of AI-generated representations of people doing and saying things they never did, events that never occurred and people who never existed have been subject to significant rhetorical hype in terms of their impact on misinformation, disinformation and freedom of expression. For the purposes of this Inquiry, WITNESS highlights three dimensions we have consistently heard in a series of expert convenings and meetings with researchers, technologists and impacted communities that comprise the only globally-oriented, human rights-led attempt to understand how to best assess threats and solutions to evolving visual misinformation and disinformation (further information on this research effort available at 'Prepare, Don't Panic: Synthetic Media and Deepfakes').[16] Each of these has implications for online freedom of expression and to legislative, technical and educational responses.

[13] https://mnemonic.org/

[14] https://www.foreignaffairs.com/articles/united-states/2020-11-11/big-tech-can-help-bring-war-criminals-justice

[15] https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes

[16] https://lab.witness.org/projects/synthetic-media-and-deep-fakes/

- The growing prevalence of non-consensual sexual images and image-based abuse directed towards women, including the increasing accessibility of AI-based generative approaches is an existing scaled harm that impacts participation in the public sphere and free expression. Legislators and technologists need to prioritize strategies for responding to these harms.

- The rhetorical claims of audiovisual manipulation including claims that compromising audio and video is 'a deepfake' are already being used to challenge critical human rights and civic evidence of state violence, corruption and official misconduct in each of the regions where WITNESS conducted expert convening work (including the US, Brazil,[17] Sub-Saharan Africa[18] and Southeast Asia).[19] This ability to claim plausible deniability on any compromising content and to exercise the so-called 'liar's dividend' are threats to a more diverse civic sphere and to free expression.

- Rhetorical claims of pervasive audiovisual manipulation have led to increasing public scepticism of real content - governments should invest in broad-based media literacy efforts that support stronger capacities at a community influencer level as well as among individual citizens and residents to better discern and interrogate online content and behaviour.

- Technology companies have a role to play in providing access to tools for detection of new forms of audiovisual manipulation as well as better ensuring access to consumer-friendly tools for detecting existing forms of 'shallowfake' manipulation such as mis-contextualized, mis-captioned or lightly edited photos and videos. They should provide as broad-based access as possible to detection tools for deepfakes, as well as build tools such as reverse video search and context-provision in platform to enable greater ease in identifying and mitigating existing shallowfakes. However, ensuring access for diverse media and civic actors globally is a concern WITNESS heard in consultations ('What's needed in deepfakes detection?')[20] -- otherwise existing issues of inequity in terms of dealing with mis/disinformation threats will be perpetuated.

15 January 2021

[17] https://lab.witness.org/brazil-deepfakes-prepare-now/

[18] https://blog.witness.org/2020/02/report-pretoria-deepfakes-workshop/

[19] https://lab.witness.org/asia-deepfakes-prepare-now/

[20] https://blog.witness.org/2020/04/whats-needed-deepfakes-detection/