DISINFORMATION AND FREEDOM OF EXPRESSION

WITNESS submission to United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, annual thematic report for 47th Session of Human Rights Council, June 2021

Submitted by: WITNESS

Contact: Sam Gregory, Program Director, sam@witness.org

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.

Introduction: WITNESS's highlights a comprehensive UNESCO/ITU recent guide on this topic to which we contributed. We then focus on a number of emerging areas of approach to address disinformation, including emerging forms of AI-driven disinformation such as deepfakes (section 9 through 13), and technical responses such as authenticity infrastructure that address a growing scope of online disinformation (1 through 8). We affirm the critical role of an affirmative right to record in response to state-sponsored disinformation and attempts to control truthful information (sections 14 through 16). We highlight the role of automated content moderation in compromising free expression and the need to preserve critical evidence that is removed from platforms (sections 17-18)). We make observations in section 19-21 on the questions of platform governance of disinformation, platform treatment of public figures, and the impact of COVID-19 on disinformation management by platforms.

In all cases we assess their impact on the rights to freedom of opinion and expression.

We begin by grounding in a broader set of frameworks for considering disinformation in the light of freedom of expression and opinion, connecting to a recent report for the UNESCO/ITU Broadband Commission for Sustainable Development to which WITNESS contributed.

Balancing countering digital disinformation while respecting freedom of opinion and expression

For an overall survey on disinformation in the context of freedom of expression, we reference the comprehensive study prepared for the Broadband Commission for Sustainable Development (UNESCO and ITU). *Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression* surveys issues with a global scope, includes an action-oriented suite of sector-specific actionable recommendations, and presents a 23-point framework to test disinformation responses. A copy of this report is provided as attachment to our submission.

Targeted analyses and recommendations to address the life cycle of online disinformation from production to transmission, reception and reproduction focus on:

- **Legislators and policy makers** (counter-disinformation campaigns, electoral-specific responses, the Freedom of Expression Assessment Framework) (Ch5, Ch8, Executive Summary)
- **Internet companies, producers and distributors** (content curation, technical and algorithmic, advertisement policy, demonetisation responses) (Ch6, Executive Summary)
- **Journalists, investigative researchers and fact checkers** (Ch4, Ch3, Ch2, Executive Summary)
- **Universities and applied and empirical researchers** (Ch3, Executive Summary)
- **Target audiences** (educational, ethical and normative, empowerment and credibility labeling responses) [Ch7, Executive Summary)

The findings are organized into a typology of 11 different categories of responses to disinformation – ranging from identification and investigatory responses, through to policy and legislative measures, technological steps, and educational approaches. For each category of response, the report includes a description of work being done around the world, by which actors, how it is funded and who or what is targeted. The report further analyses the underlying assumptions and theories of change behind these responses, while weighing the challenges and opportunities. Each category of response is also assessed in terms of its intersections with the universal human right of freedom of expression, with a particular focus on press freedom and access to information. Case studies of responses to COVID-19 disinformation are also presented within each category.

At the heart of this report is the need to balance responses to disinformation with respect for freedom of expression. The research shows us that this can be done.

This research of the Broadband Commission study is edited by Professor Kalina Bontcheva (University of Sheffield, UK) and Dr. Julie Posetti (International Center for Journalists,U.S.; Centre for Freedom of the Media, University of Sheffield/ Reuters Institute for the Study of Journalism, University of Oxford, UK). The other contributing authors are Denis Teyssou (Agence France Presse, France); Dr. Trisha Meyer (Vrije Universiteit Brussel, Belgium); Sam Gregory (WITNESS, U.S.); Clara Hanot (EU Disinfo Lab, Belgium); and Dr. Diana Maynard (University of Sheffield, UK).

WITNESS focuses the remainder of our submission on a number of emerging approaches to disinformation and an assessment of their impact on the right to freedom of opinion and expression.

Responses to disinformation and freedom of expression: Authenticity Infrastructure

1. 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 mis- and disinformation. Our report *Ticks or it Didn't Happen: Confronting Key Dilemmas in Building Authenticity Infrastructure for Multimedia* explores 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. These authenticity infrastructure (example, the Content Authenticity Initiative from Adobe and others) are technical mechanisms for normalizing the tracking the origins, manipulation and editing of multimedia as one proposed solution to problems of trust in online expression.

2. Until recently, discussion and technical development in this area were relatively niche. Over the past year, in response partially to concerns about mis- and disinformation, a number of initiatives 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, tracking edits and manipulations, and 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 of whether manipulation has occurred. Recent authenticity infrastructure initiatives, which seek to track manipulations and edits to media and to provide data at point of publication and sharing include the Content Authenticity Initiative (CAI) initiated by Adobe, New York Times and Twitter. The CAI examines 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 CAI emphasizing key trade offs from a global, human rights perspective. We advocated for features such as low technical barriers to entry and no requirement for identity as a basis of trust and inclusion of global, and flagged high-risk contexts and abusability of the framework as key concerns.

3. WITNESS has seen the value of these mechanisms to enhance trust in footage through our own experience building tools in this area such as ProofMode, 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 by enhancing the trustworthiness and probative value of key media and in countering mis/disinformation. Yet we are also acutely conscious of the risks as these approaches move from niche technologies used by journalists and rights defenders into the public tech infrastructure. We outline these risks below.

4. Authenticity infrastructure has important implications for how trust is assigned, or not, to online accounts, and on whom the burden of proof is increasingly placed to prove media is 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 technology, good battery life, GPS or connectivity can compound existing information inequities.

5. 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'.

6. Authenticity infrastructure builders must make choices on whether to insist on persistent identity or allow anonymous or pseudonymous speech as part of validating trustworthiness and integrity of online speech. This reflects many of the existing tensions around persistent real-name identity that have led to the exclusion of critical dissident and human rights voices from platforms like Facebook. One critical element that WITNESS pushed for in our involvement in the Content Authenticity Initiative CAI was that identity not be essential to the infrastructure of authenticity, and that selective, clearly indicated redaction of key audiovisual data (e.g. the ability to blur faces but protect other authenticity and context data) was critical as a part of any standard. As the Special Rapporteur reviews technical responses to protect and enhance freedom of expression, it is critical to focus on preserving options for pseudonymity and anonymity and ensuring emerging infrastructure options do not compromise privacy.

7. It is well-observed that new technologies often produce so-called 'ratchet effects' through which the utility and credibility of previous technologies is undermined. In the case of authenticity infrastructure, these ratchet effects may both discredit individual users who are unable to use authenticity infrastructure due to older technology, lack of consistent GPS, or online access-- as well as disadvantage smaller media outlets, community journalists, and others who are unable to adopt these approaches as rapidly. Expectations of provenance and 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 developed as a default, de facto, or legal obligation.

8. 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 very plausible possibility. Secondly, legislators in democratic contexts should be aware of the possibility of 'legislative opportunism' in authoritarian and non-democractic countries that will take their well-meaning approaches to regulating free speech and mitigating misinformation and disinformation in democratic contexts and adapt them into opportunistic 'fake news' laws and regulations that disadvantage human rights speech. This is the trend we are seeing already globally and should be centered as a key concern in any integration of authenticity infrastructure to a greater degree. Infrastructure possibilities are potentially appealing to governments given the increasing range of 'securitized' fake news laws (see Gabriela Lim, 2020). These laws 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 unrealistic to imagine 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.

Responses to disinformation and freedom of expression: Deepfakes

9. WITNESS has coordinated 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 mis- and disinformation and freedom of expression. For the purposes of this submission, WITNESS highlights three leading concerns we have consistently heard in a comprehensive process of feedback gathering with researchers, technologists and impacted communities. This series of deepfake preparedness convenings in the United States, Brazil, Sub-Saharan Africa and Southeast Asia comprise the only globally-oriented, human rights-led effort 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*). Each of these findings has implications for online freedom of expression and legislative, technical and educational responses.

10. A key problem: 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.

11. A key problem**:** 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 above. 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.

12. A needed solution: 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.

13. A needed solution: Technology companies have a role to play in ensuring access to tools to detect new forms of audiovisual manipulation, as well as provide consumer-friendly tools for detecting existing forms of 'shallowfake' manipulations, or mis-contextualized, mis-captioned or lightly edited photos and videos. Companies 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 inside platforms 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?') -- otherwise existing issues of inequity in terms of dealing with mis/disinformation threats will be perpetuated.

Right to Record in the context of countering disinformation and misinformation

14. The right to record is critical to the ability of civil society to counter government disinformation. The suppression of the right to record (to film authorities, the police and the military in the course of their public duties) impacts the ability of civilians to challenge disinformation in public or state-sponsored narratives. It threatens one of the primary tools of free expression available to a growing number of individuals worldwide - the capacity to film and share information that challenges disinformation, or asserts truth.

15. WITNESS's project on the Right to Record notes that the United Nations Human Rights Council has previously explicitly recognized the Right to Record in their resolution A/HRC/38/L.16, on "The promotion and protection of human rights in the context of peaceful protests." The resolution, submitted by Costa Rica and Switzerland, also contained other important language supporting the rights of witnesses and human rights defenders. A 2015 report on " The use of information and communications technologies to secure the right to life" from Christof Heynes, former Special Rapporteur on extrajudicial, summary or arbitrary executions, directs States to respect and protect the "the individual's right to make a recording of a public event, including the conduct of law enforcement officials." In a 2016 joint report, Heynes was joined by Maina Kiai, former Special Rapporteur on the rights to freedom of peaceful assembly and of association in a report on the proper management of assemblies to expand on this

> "All persons enjoy the right to observe, and by extension monitor, assemblies. This right is derived from the right to seek and receive information, which is protected under article 19 (2) of the International Covenant on Civil and Political Rights. The concept of monitoring encapsulates not only the act of observing an assembly, but also the active collection, verification and immediate use of information to address human rights problems.

16. Moving forward, WITNESS has seen significant efforts to suppress the right to record during the COVID-19 pandemic, and notes that many countries fail to affirmatively protect this right.

Content moderation and the need to preserve critical human rights content

17. The ways in which commercial content moderation disadvantages and harms vulnerable users, marginalized populations and human rights defenders is an area of critical concern to WITNESS. Over the past decade, we have seen the impact of hate speech, incitement to violence, and misinformation left on platforms and social media while legitimate speech and human rights defenders have their content and accounts removed or suspended (see our detailed submission to the UN Special Rapporteur on Freedom of Expression in regard to Content Moderation in the Digital Age, his thematic report in

2018). 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). In the context of this submission, WITNESS will focus on one dimension of this: the need for strengthened mechanisms to preserve critical human rights content and evidence of harmful disinformation that is removed by algorithms or by human oversight from platforms.

18. In recent years, we have witnessed 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), as well as the challenges of international fact-finding and judicial bodies in accessing Facebook content related to crimes in Myanmar. These events illustrate the importance of ensuring that any legislative response to platforms includes approaches to guarantee 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, whether with due process and respect for freedom of expression or in its absence. WITNESS and other human rights groups have called for 'evidence lockers' to house critical footage that may otherwise be both correctly and incorrectly removed under platform policies or relevant legislation, due to violating policies on "terrorist or violent extremist content" or graphic violence (see also Alexa Koenig, Big Tech Can Help Bring War Criminals to Justice: Social Media Companies Need to Preserve Evidence of Abuse, 2020 and Human Rights Watch 'Video Unavailable: Social Media Platforms Remove Evidence of War Crimes'). Given the critical importance of this social media content for accountability in a growing number of human rights scenarios, it is incumbent on civil society, platforms and government to collaborate to identify appropriate models and mechanisms for preserving it.

Current topic: Role of platforms in relation to disinformation and public figures

19. More often than not, world leaders who incite violence and hatred online get away with it. This has been the case globally and Trump's deplatforming shed increased attention on the incongruity of this response compared to actions in contexts such as Brazil, India, and the Philippines. On other forms of disinformation, COVID-19 has opened the door to decision-making by platforms to crack down on inaccurate claims made by President Jair Bolsonaro of Brazil and former President Trump in the US. In doing so, Twitter, Facebook and Google showed that they were prepared to hold influential politicians to the same standards applied to everyday people. Now that this precedent exists, there should be no valid reason to allow public figures to contravene these rules, especially when clear harms have, and will continue to, result from their behavior.

20. Until now, most platforms have provided public interest exceptions for leaders who share false information or incite hate or have acted only on particular posts. Moving forward, leaders should be subject to equal or greater scrutiny when they push boundaries on platforms, not less. With power comes responsibility, and freedom of speech does not guarantee freedom of reach. At the same time, we must demand transparency on how decisions are made for both for leaders and ordinary users, and hew to human rights principles of legitimacy, proportionality, and specificity rather than over-broad, inconsistent deplatforming. Moving forward, we must demand to see these approaches

applied consistently and with an understanding of the context outside the US, rather than reinforcing a trend of US exceptionalism that has come to define platform accountability. Our WITNESS submission to the Facebook Oversight Board consideration of the Trump suspension further outlines these issues.

Current topic: Platforms in the specific context of responses to COVID-19 disinformation

21. In WITNESS's assessment of COVID-19 Misinformation and Disinformation Responses, we provide a framework for assessing the actions of platforms in countering COVID-19 related misinformation and disinformation. We provide an overview of some of the key trends in platform response to misinformation, disinformation and harmful speech during COVID-19, and develop a set of criteria through which the wide range of possible reactions can be assessed, using a framework based on human rights and our experience working with marginalized communities and human rights defenders globally. We then apply this framework to highlight areas in which response has been strong and should be expanded into other information domains, along with key gaps and concerns which should be addressed as circumstances evolve. This leads us to a set of recommendations for what companies should do now, what they should continue to do in the future, and what they should stop doing altogether.