Shared UN considerations for online communications companies on the issues of countering disinformation and enhancing transparency

Arising from: UN informal interagency dialogues on disinformation and data transparency 2020-2021[1]

Note: *This document is a draft for discussion purposes that has not been formally edited or reviewed. As such it does not necessarily represent the official views of the United Nations or the agencies of the staff who have contributed.*

Reinforcing the right to seek and receive information by means of ensuring timely, local, relevant and accurate knowledge and data of public interest

In its work for peace, security, development and human rights, the UN family has serious concerns about the new threats posed by the spread of misinformation and disinformation at scale. This content is often spread through the platforms owned by large internet companies. While Alphabet (the parent company of Google and YouTube), Facebook, Twitter, and other social media, search and messaging platforms and companies have become major communications infrastructure, research and investigative journalism continue to expose the socio-technical vulnerabilities in these platforms in regard to their role of facilitating the spread of false and misleading content (disinformation and misinformation) at the expense of the flow of timely, accurate and factual information.

In accordance with the Universal Declaration of Human Rights, the UN has a mandate to protect and promote the rights to life, liberty and the security of person. International standards codify a range of associated rights such as the right to expression and access to information, to health, to environment, and to participation in democratic processes. The UN's Guiding Principles on Business and Human Rights recognizes the need for corporate responsibility to respect human rights; and the need for greater access by victims to effective remedy.

Full protection of human rights is challenged by many of the current business models for online content curation, as well as operational policies and practices for content moderation. This is compounded by lack of transparency about these models as well as about the enforcement of these policies by massive and transnational internet communications companies. In particular, the problems of disinformation and misinformation enabled and sometimes fueled by the companies constitute obstacles to the full enjoyment of human rights.

Against this background, in the context of the pandemic, the UN has called for effective public access to information related to COVID-19 and action to be taken against those who fuel xenophobia and hate speech.[2] The issue goes beyond the pandemic

[1] The informal interagency dialogues were convened by UNESCO and WHO, with the participation of EOSG, IOM, ITU, OHCHR, UN DESA, UN DGC, UNDP, UNEAD, UN ECA, UN ECLAC, UN EDA, UNEP, UN ESCWA, UN Global Pulse, UNICEF, UN Office of the Envoy on Technology, UN OSAPG, UN Women and UN75.

[2] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye. Disease pandemics and the freedom of opinion and expression. April 23, 2020. *https://undocs.org/A/HRC/44/49*; UNESCO. Journalism, press freedom and COVID-19. April 2020. *https://unesdoc.unesco.org/ark:/48223/pf0000373573*

however, as disinformation and misinformation impact on rights of voters, migrants and refugees, and minorities for example, as well as the wider public's right to a sustainable environment. At the same time, the experiences, both negative and positive, in regard to COVID-19 responses by internet communications companies provide a basis for considering stronger engagement by these actors in relation to these problems, and particularly by them enhancing transparency about the extent and engines of the problems as well as about the steps they are taking in relation to these.

The considerations below flow particularly from the right to freedom of expression which encompasses the right of access to information, and other universal human rights, as expressed in the Joint Declaration on Freedom of Expression and "Fake News," Disinformation, and Propaganda,[3] and Article 19 of the International Covenant on Civil and Political Rights.[4]

It is agreed within the UN family that a provisional path for mitigating disinformation- and misinformation-at-scale by internet communications companies could entail four categories of action within the context of their duty to respect human rights.[5] Accordingly, the following interdependent normative proposals apply - Transparency, Accountability, Accessibility and Co-operation:

1. **Transparency**[6]: Internet communications companies should:

   a. Go beyond current limited periodic transparency reports where these exist, and take steps to provide greater transparency regarding their curational model for content prioritization and reprioritization, including steps taken to inform users about possible options on how they themselves may shape their content feeds or search results listings.

   b. Increase the transparency of accounts attracting and those paying for advertising, including metadata on targeted categories and advertising spend, and extend beyond topics such as social issues, elections, or politics.

   c. Provide added context such as labels to accounts that are affiliated with corporations, agencies, politicians, media companies, and states, and make clear the criteria used to define the labels, the extent of their application, and records of appeal against such.

   d. Preserve and create a database of posts and accounts that have been removed (or subjected to other moderation treatment, to be accessed by regulators, journalists and researchers while respecting privacy, making an attempt to preserve the context in which the posts were shared).[7]

   e. Provide data on the spread of such content before it was subjected to moderation (including the number of users who have seen the post and sharing/engagement with the post prior to moderation).

   f. Provide greater algorithmic transparency of how information is automatically ranked, sorted, and presented to users, especially prominently displaying the source of all science and health-related content.

   g. Implement counterfactual explanations for decisions-making algorithms which can help provide users with effective information about the effects of such algorithms, while avoiding

3 United Nations Human Rights Office of the High Commissioner. "Freedom of Expression Monitors Issue Joint Declaration on 'Fake News', Disinformation and Propaganda," March 3, 2017. *https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=21287&LangID=E*

4 United Nations General Assembly. "International Covenant on Civil and Political Rights," March 23, 1976. *https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx*

5 Other UN initiatives in this area include, among others: the OHCHR B-Tech project implementing the United Nations Guiding Principles on Business and Human rights in the technology space (*https://www.ohchr.org/EN/Issues/Business/Pages/B-TechProject.aspx*); the UN Secretary General's Roadmap on Digital Cooperation (*https://www.un.org/en/content/digital-cooperation-roadmap/*); the UN Privacy Policy Group (*https://www.unglobalpulse.org/policy/un-privacy-policy-group/*); the Internet Governance Forum (*https://www.intgovforum.org/multilingual/*); and the UN Strategy and Plan of Action on Hate Speech (*https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml*)

6 Internal note: the concept of transparency may be further reviewed and further elaborated in the next version of this paper.

7 Emma Llanso. "COVID-19 Content Moderation Research Letter – in English, Spanish, & Arabic." Center for Democracy & Technology, April 22, 2020. *https://cdt.org/insights/covid-19-content-moderation-research-letter/*; .UNECE. "Environmental Democracy in Times of COVID-19," 2020. *https://www.unece.org/info/media/executive-secretary-blog/2020/environmental-democracy-in-times-of-covid-19/doc.html*

property infringement or code and data exploitability.

h. Engage with and explain to users the conclusions of algorithmic decision-making, which can promote and amplify content with disinformation, to help achieve transparency about black box models and their outcomes.

2. <u>Accountability:</u> Internet communications companies should:

a. Build upon existing governance arrangements to ensure strategic societal and technical transformative change is able to anticipate potentially negative issues and mitigate negative outcomes and impacts, including through collaborative partnerships with thematic experts from the UN and other sectors;

b. Demonstrate comprehensively and consistently, as distinct from occasional announcements, that they are taking strong action to prevent the wide scale spread of false and misleading content on their platforms or search results.

c. Increase efforts to educate users about false and misleading content by reporting on coordinated behavior by fake accounts and the use of bots.[8]

d. Ensure that their terms of service are applied equally to all account holders and content consistently, regardless of political and social status or influence, and including political propaganda that includes health and other types of misinformation.

e. Put attention and resources into these issues across all countries of operation, not just a limited number of focal countries, and make available data about their treatment of content issues in minority languages.

f. Engage in open and inclusive multi-stakeholder participation in the development and evaluation of codes of conduct and community standards.

3. <u>Accessibility:</u> The internet communications companies should:

a. Be more responsive in terms of amplifying authoritative public interest content to be more accessible to the public. This includes visibly redirecting individuals seeking information about COVID-19 and other topics of public interest to accurate information and increasing the visibility and reach of trusted official and journalistic content professionals.

b. Take immediate and evident action to use labels, delays, or other treatment of harmful information, and complement this by using recommendation and trending algorithmic systems to elevate timely, local, and accurate information.

4. <u>Cooperation:</u> The internet communications companies should:

a. Work with UN actors and other expert stakeholders at country level who are concerned with taking pre-emptive actions regarding management and response to mis/disinformation such as health misinformation trends, interference in political/democratic processes including elections, and anti-migrant hostility.

b. Allow journalists and credible bodies to review the ethical dimensions of algorithms used to find or distribute content, to locate and investigate bias and discrimination, such as false political advertisements targeted at voter suppression of minority groups.

c. Engage information commissions, electoral management bodies, and data regulators in transparency initiatives to allow for an impartial exchange of information and to ensure accountability in content curation, especially in times of crisis when disinformation multiplies. ◆

[8] Donovan, Joan. 2020. "Concrete Recommendations for Cutting Through Misinformation During the COVID-19 Pandemic." American Journal of Public Health 110 (S3): S286–87. *https://doi.org/10.2105/AJPH.2020.305922*