Input to a Virtual Global Consultation to Develop a UN Code of Conduct for Information Integrity

IT for Change

January 2024

IT for Change

Table of Contents

Input to the Virtual Global Consultation to Develop a UN Code of Conduct for Information Integrity

IT for Change[1]

January 2024

We commend the initiative of the United Nations to develop a code of conduct for information integrity on digital platforms in close consultation with a broad range of civil society groups. The Code of Conduct is intended to be developed based on the set of nine principles proposed in the United Nations Secretary-General's 'Our Common Agenda' policy brief number eight, issued in June 2023. Below, we provide our comments and actionable suggestions for implementing these principles. Additionally, we propose the inclusion of the principle of 'Strong Accountability and Liability Framework' to bind the commitment of digital platforms to information integrity.

I. Commitment to Information Integrity

The UN Secretary-General's policy brief defines information integrity as the accuracy, consistency, and reliability of information. It emphasizes that information integrity is threatened by disinformation, misinformation, and hate speech. While the policy brief itself acknowledges there are no universally accepted definitions for these terms, various United Nations entities have developed working definitions. It is important for a Code of Conduct on Information Integrity to establish clear definitions for these contested terms, eliminating any ambiguities in the interpretation, practice, and implementation of the Code.

Apart from the social, political, and other strategic imperatives, there needs to be an explicit recognition of the economic aspects within the information governance of the platformized internet. The attention economy logic, upon which digital platforms have built their business model, tends to prioritize the circulation of false, misleading, and toxic content to generate higher user engagement, and consequently, higher revenue.

The amplification of disinformation due to algorithmic virality, which tends to prioritize profits over people, poses a crisis for democracy. Stakeholders committed to information integrity must seek solutions that address the impunity of large platform corporations through mechanisms of public

[1] For additional information, please contact Anita Gurumurthy (anita@itforchange.net) and Merrin Muhammed Ashraf (merrin@itforchange.net).

accountability, oversight, and due process. Additionally, efforts should be made to encourage platform models to foster pluralistic public spheres.

II. Respect for Human Rights

A Multilateral System should:

(i) Evolve a binding, human rights-based content governance paradigm for the transnational communications agora of the internet that holds states and corporations to account for human rights violations.

Member States should:

(i) Be transparent about the content removal or restriction orders issued to digital platforms. These orders must follow the due legal processes, be subject to an independent judicial authority, and the basis for such orders must be published;
(ii) In case they deem it necessary to impose restrictions, such as internet shutdown, ensure that these restrictions are proportional, non-discriminatory, and undertaken only as necessary for transparently reported and legitimate aims, in accordance with international human rights law; and
(iii) In pursuance of the UN Guiding Principles on Business and Human Rights, enact legislative and policy measures to mandating digital platforms to respect the human rights of both users and non-users of the platform.

Platforms should:

(i) Ensure the integration of human rights and due process considerations into all stages of the design process, as well as in content moderation and curation policies and practices. This should be ensured irrespective of Member States' ability or willingness to fulfil their duty to protect human rights.

III. Support for Independent Media

Member States should:

(i) Establish regulatory mechanisms for fair revenue sharing, ensuring that social media platforms fairly compensate media publishers for sharing content to promote trusted journalism in the digital age, as seen in Canada and Australia;
(ii) Mandate data portability and interoperability between platforms to facilitate the emergence of new digital platforms with diverse value propositions, offering users varied information and viewpoints. An example is the Digital Markets Act of the European Union; and

(iii) Require platforms hosting news to optimize curation algorithms for diversity, advancing the goal of securing information as a public good.

Member States are also recommended to:

(i) Develop policies to encourage a diverse and plural ecology of digital media, interoperating over common protocols, and possibly, independent client-side applications;
(ii) Consider providing public funding for a meaningful alternative to for-profit communication platforms, with a civic mission of providing citizens with a diverse and global view of the world and ensuring arenas where timely, accurate, local knowledge is always available; and
(iii) Engage with platforms and technologists to explore the possibility of functionally separating content hosting from curation functions, and developing a marketplace of alternative recommender systems.

IV. Transparency Efforts

Member States should mandate that platforms comply with meaningful and proactive disclosure in respect of the following:

(i) Means used for content moderation, including a specification of the use of automated tools, training data used for the development and improvement of such tools, procedures for quality assurance or evaluation to improve the decisions made using such tools, and measures taken to mitigate any harm from incorrect decisions;
(ii) Deployment of any proprietary algorithm to curate content for the users; the logic behind determining what content appears on a user's feed and what is hidden; and the nature of personal data of users that is collected and used by the recommendation algorithm;
(iii) Platform's policies and practices regarding the placement of advertisements;
(iv) Number of human moderators, their expertise, and employment status; and
(v) Types of complaints received, and the number of complaints under each type, time taken for resolution, appeal process, the outcome of appeals, and the specific action taken by platforms in relation to a complaint and the reasons for the same.

The effectiveness of digital platforms' transparency mechanisms should be independently evaluated against international standards, such as the 26 high-level principles on transparency in the digital age set forth by UNESCO.

V. User Empowerment

Platforms should:

(i) Make information accessible for users to understand the various products, services, and tools provided. Platforms should empower users to make informed decisions about the content they share and consume. Information should be provided in users' own languages, taking into consideration their ages and disabilities; and
(ii) In case of vital public interest, such as public health, elections, social security services, suicide prevention, and support for victims of violence, credible official sources of public interest information should be highlighted and prioritized by algorithms and recommender systems of digital platforms.

Member states should:

(i) Institute legal frameworks for transparency obligations of platforms with clear limits on algorithmic personalization to protect user's autonomy and human rights. This is critical given the overwhelming role that platforms play in organizing public discourse. Democratic discourse in digital society depends on autonomy-enhancing platform infrastructure, ensuring that user choice is structured through adequate transparency, and normative and ethical limits to algorithmic personalization.

VI. Strengthened Research and Data Access

Information governance regimes must empower civic-publics to be watchdogs of digital age democracy.

Towards this goal, Member States should:

(i) Require digital platforms to share with regulators, independent researchers, academics, journalists, and advocacy groups, the data necessary to understand the impact of digital platforms, especially when public interest considerations are involved. For example, Article 31 of EU's Digital Services Act;
(ii) Provide clear guidance on how and under what conditions data sharing by digital platforms may be deemed necessary, proportionate, and reasonable for research purposes; and
(iii) In cases where user privacy may be affected, prescribe appropriate safeguards, including vetting of requesting individuals/entities, and specify how data should be shared (such as anonymizing datasets through measures like de-identification and sampling before sharing).

Platforms should:

(i) Comply with data sharing mandates and facilitate data access for regulators, independent researchers, academics, journalists, and advocacy groups on an ongoing basis, through automated means, such as application programming interfaces (APIs), or other open and accessible technical solutions allowing the analysis of the said data; and
(ii) Build reliable interfaces for data access and provide disaggregated data based on gender and other relevant intersecting factors, such as race, ethnicity, age, socio-economic status, disability, etc.

VII. Scaled-up Responses

Platforms should:

(i) Invest in developing their own mechanisms for fact-checking, instead of transferring the costs of cleaning up corporate irresponsibility to civil society; and
(ii) Engage in periodic dialogues and consultations with local communities, especially vulnerable and marginalized individuals and groups, civil society, and academia, to build an understanding of societal harm arising from their content policies and practices. This understanding will enable them to work towards minimizing or eliminating exposure to harmful content. Targeted outreach, respect for cultural diversity, and the use of inclusive language and formats can facilitate effective participation in such dialogues and consultations.

Member States should:

(i) Ensure the independence and integrity of the members of fact-checking units established by them so that they are not under the influence of the political executive and can function in a non-partisan and fair manner.

Member States are also recommended to:

(i) Launch a citizens' dialogue to determine what, if any, new charters of rights, institutions, or regulatory frameworks may be necessary to ensure that the algorithmic curation of news and information complements societal norms, international human rights agreements, and public expectations.

VIII. Stronger Disincentives

An explicit call, made by the UN Secretary-General in his policy brief, to digital platforms to move away from business models that prioritize profit over human rights is very welcome. However, it is not prudent to rely on the voluntary initiative of profit-driven technology companies to achieve this.

Member States should take legal and policy measures to create disincentives for platforms to act in ways harmful to information integrity and human rights in their pursuit of revenue and profit.

These measures can include:

(i) Requiring platforms to make a separation between those in charge of enforcement of content moderation rules of the platform and those whose job performance is measured against other metrics, such as product growth and political lobbying;
(ii) Requiring disclosure of the nature and extent of involvement of third-parties in their content moderation and levying fines on companies that fail to disclose information on outside parties influencing individual content moderation decisions; and
(iii) Instituting mechanisms for public accountability of private governance decision-makers.

Measures such as the above need to be instituted and enforced by Member States through democratic legislative processes.

IX. Enhanced Trust and Safety

Platforms should:

(i) Conduct periodic risk assessments and submit reports to an independent regulator to determine, identify, and address any actual or potential harm or human rights impact of their operations or actions. Such assessments should also be undertaken prior to any major design changes, major decisions, or changes in operations, or new activity or relationship, or in response to a major event or any change in the operating environment;
(ii) Institute mechanisms to proactively detect illegal and harmful content that is generated using AI tools and either remove it or label it as AI-generated content as appropriate. In addition to technological measures, platforms should also engage with human reviewers and collaborate with fact-checkers to determine the authenticity of a piece of content; and
(iii) Adopt measures, such as triggering an internal viral circuit breaker, to prevent the algorithmic amplification of AI-generated unlawful or harmful content.

Member States should:

(i) Mandate AI actors to disclose and combat any kind of stereotyping in the outcomes of AI systems to ensure that training data sets do not foster cultural, economic, or social inequalities, prejudice, the spreading of disinformation and misinformation, and disruption of freedom of expression and access to information.

X. Other (proposal for a new principle not already addressed)

We recommend the inclusion of 'Strong Accountability and Liability Framework' as a principle to bind the commitment of digital platforms to information integrity. Regulatory responses have to move beyond self-regulation and reactive content take-downs towards imposing systemic responsibility or a statutory duty of care on platform owners for addressing the individual and social harms stemming from their techno-design choices.

In this regard, multilateral systems should adopt a binding consensus to enforce corporate accountability for preventing hate speech and incitement to discrimination, hostility, and violence in platform environments, including algorithmic content moderation and curation.

Member States should adopt comprehensive ex-ante and ex-post accountability measures, regulated by legislation. Ex-ante measures encompass periodic risk assessments, human rights, and gender impact audits, addressing systemic risks, privacy and safety by design, robust user reporting, and proactive transparency. Independent regulators appointed by governments should oversee the enforceability of these measures.

Ex-post measures include developing a liability framework to hold platforms and those directly in charge of and responsible for the conduct of business accountable for enabling or facilitating harms, including disinformation, hate speech, incitement to violence, and for any systematic or deliberate failure to take steps to prevent or mitigate the harm, despite actual knowledge of it.