



SUMMIT OF THE FUTURE INFORMATION CLEARINGHOUSE

BULLETIN NO. 15:

Our Common Agenda Policy Brief 8 – Information Integrity on Digital Platforms



PROJECT TEAM: Eshana Amarasinghe (*Lead Author, Bulletin No. 15*), Eliane El Haber, Fergus Watt, Ishaan Shah, Jebilson Raja Joslin, Jeffery Huffines and Mwendwa Kiogora

Policy Brief 8 – Information Integrity on Digital Platforms

ABOUT: Building on the proposals presented in Our Common Agenda report, the Secretary-General (SG) is publishing a [series](#) of Policy Briefs over 2023 to serve as inputs into the preparations for the Summit of the Future (SOTF). The Policy Brief on [information integrity on digital platforms](#) is the eighth one in that series.

EXECUTIVE SUMMARY:

Information integrity is increasingly under threat of mis- and disinformation and hate speech, which inflict a range of harms; from worsening tensions in conflict areas to undermining the climate emergency to worsening the economic and social exclusion of vulnerable groups. Member states have adopted legislation to promote information integrity while protecting users' freedom of expression on digital platforms. However, several challenges remain in terms of protecting data, empowering users and improving transparency, among other issues. To this end, the Secretary General proposes the United Nations Code of Conduct, which aims to action the following: a commitment to information integrity, respect for human rights, support for independent media, increased transparency, user empowerment, strengthened research and data access, scaled up responses, stronger disincentives to enabling disinformation and enhanced trust and safety. Principles for a UN Code of Conduct can be found on page 4 of this brief.

What is Information Integrity? – Information integrity refers to the accuracy, consistency and reliability of information. It is threatened by disinformation, misinformation and hate speech.¹ For the purposes of this brief, disinformation is information that is not only inaccurate, but is also intended to deceive and is spread in order to inflict harm. Misinformation refers to the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods.

Information Integrity and Digital Platforms – Digital platforms should be integral players in the drive to uphold information integrity; the velocity, volume and virality of their spread via digital channels warrants an urgent and tailored response.²

By focusing on the removal of harmful content *on digital platforms*, some States have introduced flawed and overbroad legislation that has in effect silenced “protected speech”, which is permitted under international law.

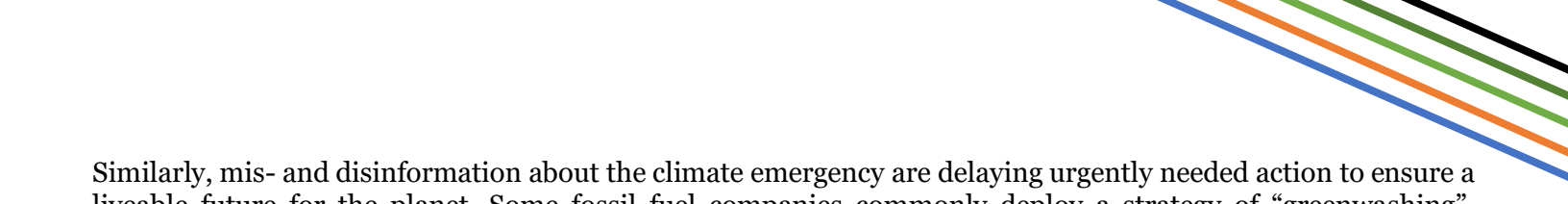
A dominant approach in the current business models of most digital platforms hinges on the “attention economy”. Algorithms are designed to prioritize content that keeps users' attention, thereby maximizing engagement and advertising revenue. Inaccurate and hateful content designed to polarize users and generate strong emotions is often that which generates the most engagement, with the result that algorithms have been known to reward and amplify mis- and disinformation and hate speech.

What harm is being caused by online mis- and disinformation and hate speech? – Mis- and disinformation can be dangerous and potentially deadly, especially in times of crisis, emergency or conflict. For example, many victims of COVID-19 refused to get vaccinated or take basic health precautions after being exposed to mis- and disinformation online.

Weaponized information has the ability to “amplify tensions and divisions in times of emergency, crisis, key political moments or armed conflict” (A/77/287, para. 6.). Countries in the midst of conflict, or with otherwise volatile contexts that are often less lucrative markets, have not been allocated sufficient resources for content moderation or user assistance.

¹ Hate speech, according to the working definition in the United Nations Strategy and Plan of Action on Hate Speech, is “any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor”.

² For the purposes of the present brief, the term “digital platform” refers to a digital service that facilitates interactions between two or more users, covering a wide range of activities, from social media and search engines to messaging apps. Typically, they collect data about their users and their interactions.



Similarly, mis- and disinformation about the climate emergency are delaying urgently needed action to ensure a liveable future for the planet. Some fossil fuel companies commonly deploy a strategy of “greenwashing”, misleading the public into believing that a company or entity is doing more to protect the environment, and less to harm it, than it is.

Mis- and disinformation are having a profound impact on democracy, weakening trust in democratic institutions and independent media, and dampening participation in political and public affairs. Throughout the electoral cycle, exposure to false and misleading information can rob voters of the chance to make informed choices. States and political leaders have proved to be potent sources of disinformation, deliberately and strategically spreading falsehoods to maintain or secure power, or undermine democratic processes in other countries.

Marginalized and vulnerable groups are also frequent targets of mis- and disinformation and hate speech, resulting in their further social, economic and political exclusion. These attacks undermine political participation and weaken democratic institutions and human rights, including the freedom of expression and access to information of these groups.

Mis- and disinformation also cross-pollinate between and within platforms and traditional media, becoming even more complex to track and address if not detected at the source. At the same time, the rise of digital platforms has precipitated a dramatic decline in trustworthy, independent media. “Newswashing” – whereby sponsored content is dressed up to look like reported news stories – is often inadequately signposted when posted to digital platforms, lending it a veneer of legitimacy.

Disinformation is also having a direct impact on the work of the United Nations. *In particular*, on the Organization’s operational safety, effectiveness and ability to deliver. Mis- and disinformation can also be used to target humanitarian actors and hamper life-saving operations in conflict areas.

Relevant international legal frameworks – In its resolution 76/227, adopted in 2021, the General Assembly emphasized that all forms of disinformation can negatively impact the enjoyment of human rights and fundamental freedoms, as well as the attainment of the Sustainable Development Goals. Similarly, in its resolution 49/21, adopted in 2022, the Human Rights Council affirmed that disinformation can negatively affect the enjoyment and realization of all human rights.

Article 19 of the Universal Declaration of Human Rights and article 19 (2) of the *Covenant on Civil and Political Rights* protect the right to freedom of expression, including the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, and through any media. Linked to freedom of expression, freedom of information is itself a right.

Freedom of expression and access to information may be subject to certain restrictions that meet specific criteria laid out in article 19 (3) of the *Covenant*. States cannot add additional grounds or restrict expression beyond what is permissible under international law.

Hate speech has been a precursor to atrocity crimes, including genocide. The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence, adopted in 2012, provides practical legal and policy guidance to States on how best to implement article 20 (2) of the *Covenant* and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination, which prohibit certain forms of hate speech.

How can we strengthen information integrity? – Efforts to achieve the Sustainable Development Goals are fundamental to building a world in which trust can be restored. In crafting responses, it is important not to lose sight of the tremendous value digital platforms bring to the world.

❖ **Regulatory Responses -**

- a. Digital Services Act - establishes new rules for users, digital platforms and businesses operating online within the European Union. Measures taken aim at illegal online content, goods and services and provide a mechanism for users both to flag illegal content and to challenge moderation decisions that go against them. They require digital platforms to

- improve transparency, especially on the use and nature of recommendation algorithms, and for larger platforms to provide researchers with access to data.
- b. Code of Practice on Discrimination - sets out principles and commitments for online platforms and the advertising sector to counter the spread of disinformation online in the European Union. These include voluntary commitments to help demonetize disinformation, both by preventing the dissemination of advertisements containing disinformation and by avoiding the placement of advertisements alongside content containing disinformation. Signatories also agreed to label political advertising more clearly...and to create searchable databases of political advertisements. Furthermore, they committed to share information about malicious manipulative behaviours used to spread disinformation...and regularly update and implement policies to tackle them.³

❖ **Digital Platform Responses** - Several of the larger platforms have publicly committed to uphold the Guiding Principles on Business and Human Rights, but gaps persist in policy, transparency and implementation. Some platforms do not enforce their own standards and, to varying degrees, allow and amplify lies and hate.

Most digital platforms have some kind of system of self-regulation, moderation or oversight mechanisms in place, yet transparency around content removal policy and practice remains a challenge. Translation of moderation tools and oversight mechanisms into local languages is incomplete across platforms, a recent survey found. At the same time, moderation is often outsourced and woefully underresourced in languages other than English. Moderators report being constantly exposed to violent and disturbing content, and being given a matter of seconds to determine if a reported post violates company policy. Automated content moderation systems can play an essential role, but are exposed to possible bias based on the data and structures used to train them. They also have high rates of error in English and even worse success rates across other languages. A number of digital platforms employ trust and safety, human rights and information integrity teams, yet these experts are often not included at the earliest stages of product development and are often the first jobs cut during cost-saving measures.

❖ **Data Access** - Existing research and resources remain heavily skewed towards the United States of America and Europe. This is partly because researchers lack access to platforms and their data. The tools needed for effective research of the limited data provided by the platforms also tend to be designed with marketing in mind and are largely prohibitively expensive. A shift by the platforms from an “access by request” approach to “disclosure by default”, with necessary safeguards for privacy, would allow researchers to properly evaluate harms.

❖ **User Empowerment** - Platform users, including marginalized groups, should be encouraged, included and involved in the policy space. As digital natives, young people, in particular young women, and children are already often the targets of mis- and disinformation and hate speech and will be directly affected by emerging and new platforms. Younger users can speak from experience about the differentiated impact of various proposals and their potential flaws.

Improved critical thinking skills can make users more resilient against digital manipulation. The United Nations Verified initiative has successfully deployed a range of tactics, including targeting messaging for users, pre-bunking – warning users about falsehoods before they encounter them – and digital literacy drives.

❖ **Disincentives** - The current business models of most digital platforms prioritize engagement above human rights, privacy and safety. Proposals seek to address the profitability of disinformation, ensure full transparency around monetization of content and independent risk assessments, and disincentivize those involved in online advertising from enabling disinformation. Advertisers can also pressure digital

³ In February 2023, signatories to the Code of Practice published their first baseline reports on how they are implementing the commitments. The reports provided insight into the extent to which advertising revenue was prevented from flowing to disinformation actors and other detected manipulative behaviour, including a large-scale coordinated effort to manipulate public opinion about the war in Ukraine in several European countries.

platforms to step up action to protect information integrity and can refrain from advertising with media outlets that fuel hatred and spread disinformation.

- ❖ **Independent Media** - New measures in dozens of countries continue to undermine press freedom. With 2.7 billion people still offline, a further priority is to strengthen independent media, boost the prevalence of fact-checking initiatives and underpin reliable and accurate reporting in the public interest. Ethical reporters, with quality training and working conditions, have the skills to restore balance in the face of mis- and disinformation.
- ❖ **Future Proofing** - While holding almost unimaginable potential to address global challenges, there are serious and urgent concerns about the equally powerful potential of recent advances in artificial intelligence – including image generators and video deepfakes – to threaten information integrity. It is essential that user privacy, security, transparency and safety by design are integrated into all new technologies and products at the outset.
- ❖ **United Nations Responses** - The United Nations Strategy and Plan of Action on Hate Speech sets out strategic guidance for the Organization to address hate speech at the national and global levels. In February 2023, UNESCO hosted the Internet for Trust conference to discuss a set of draft global guidelines for regulating digital platforms, due to be finalized later this year.

Towards a United Nations Code of Conduct - The United Nations Code of Conduct for Information Integrity on Digital Platforms would build upon the following principles:

Commitment to information integrity:

- a. All stakeholders should refrain from using, supporting or amplifying disinformation and hate speech for any purpose;

Respect for human rights:

- b. Member States should:
 - i. Ensure that responses to mis- and disinformation and hate speech are consistent with international law, including international human rights law, and are not misused to block any legitimate expression of views or opinion, including through blanket Internet shutdowns or bans on platforms or media outlets
 - ii. Undertake regulatory measures to protect the fundamental rights of users of digital platforms, including enforcement mechanisms, with full transparency as to the requirements placed on technology companies;
- c. All stakeholders should comply with the Guiding Principles on Business and Human Rights

Support for independent media:

- d. Member States should guarantee a free, viable, independent and plural media landscape with strong protections for journalists and independent media, and support the establishment, funding and training of independent fact-checking organizations in local languages;
- e. News media should ensure accurate and ethical independent reporting supported by quality training and adequate working conditions in line with international labour and human rights norms and standards

Increased transparency:

- f. Digital platforms should:
 - i. Ensure meaningful transparency regarding algorithms, data, content moderation and advertising;
 - ii. Publish and publicize accessible policies on mis- and disinformation and hate speech, and report on the prevalence of coordinated disinformation on their services and the efficacy of policies to counter such operations;

- g.** News media should ensure meaningful transparency of funding sources and advertising policies, and clearly distinguish editorial content from paid advertising, including when publishing to digital platforms;

User empowerment:

- h.** Member States should ensure public access to accurate, transparent, and credibly sourced government information, particularly information that serves the public interest, including all aspects of the Sustainable Development Goals;
- i.** Digital platforms should ensure transparent user empowerment and protection, giving people greater choice over the content that they see and how their data is used. They should enable users to prove identity and authenticity free of monetary or privacy tradeoffs and establish transparent user complaint and reporting processes supported by independent, well publicized and accessible complaint review mechanisms;
- j.** All stakeholders should invest in robust digital literacy drives to empower users of all ages to better understand how digital platforms work, how their personal data might be used, and to identify and respond to mis- and disinformation and hate speech. Particular attention should be given to ensuring that young people, adolescents and children are fully aware of their rights in online spaces;

Strengthened research and data access:

- k.** Member States should invest in and support independent research on the prevalence and impact of mis- and disinformation and hate speech across countries and languages, particularly in underserved contexts and in languages other than English, allowing civil society and academia to operate freely and safely;
- l.** Digital platforms should:
 - i.** Allow researchers and academics access to data, while respecting user privacy.
 - ii.** Ensure the full participation of civil society in efforts to address mis- and disinformation and hate speech;

Scaled up responses:


- m.** All stakeholders should:
 - i.** Allocate resources to address and report on the origins, spread and impact of mis- and disinformation and hate speech, while respecting human rights norms and standards and further invest in fact-checking capabilities across countries and contexts;
 - ii.** Form broad coalitions on information integrity, bringing together different expertise and approaches to help to bridge the gap between local organizations and technology companies operating at a global scale;
 - iii.** Promote training and capacity-building to develop understanding of how mis- and disinformation and hate speech manifest and to strengthen prevention and mitigation strategies;

Stronger disincentives:

- n.** Digital platforms should move away from business models that prioritize engagement above human rights, privacy and safety;
- o.** Advertisers and digital platforms should ensure that advertisements are not placed next to online mis- or disinformation or hate speech, and that advertising containing disinformation is not promoted;
- p.** News media should ensure that all paid advertising and advertorial content is clearly marked as such and is free of mis- and disinformation and hate speech;

Enhanced trust and safety:

- q.** Digital platforms should:
 - i.** Ensure safety and privacy by design in all products, including through adequate resourcing of in-house trust and safety expertise, alongside consistent application of policies across countries and languages;

- 
- ii. Invest in human and artificial intelligence content moderation systems in all languages used in countries of operation, and ensure content reporting mechanisms are transparent, with an accelerated response rate, especially in conflict settings;
 - r. All stakeholders should take urgent and immediate measures to ensure the safe, secure, responsible, ethical and human rights-compliant use of artificial intelligence and address the implications of recent advances in this field for the spread of mis- and disinformation and hate speech.

Consultations will continue with stakeholders to further refine the content of the Code of Conduct, as well as to identify concrete methodologies to operationalize its principles.

Next steps - The United Nations Secretariat will undertake broad consultations with a range of stakeholders on the development of the United Nations Code of Conduct, including mechanisms for follow-up and implementation. To support and inform the Code, the United Nations Secretariat may carry out in-depth studies to enhance understanding of information integrity globally, especially in under researched parts of the world.

The Secretary-General will establish dedicated capacity in the United Nations Secretariat to scale up the response to online mis- and disinformation and hate speech. Based on expert monitoring and analysis, such capacity would develop tailored communication strategies to anticipate and/or rapidly address threats before they spiral into online and offline harm, and support capacity-building of United Nations staff and Member States. It would support efforts of Member States, digital platforms and other stakeholders to adhere to and implement the Code when finalized.