CODE OF CONDUCT

FOR INFORMATION INTEGRITY ON DIGITAL PLATFORMS

Working Group 3
61ST Graduate Study Programme
UN Office at Geneva

United Nations

United Nations Geneva

Code of Conduct for Information Integrity on Digital Platforms

Distr.: General
1 4 July 2023

United Nations 61st Graduate Study Programme
UNOG

Adopted by UN GSP 61
on 14 July 2023

Preamble

The Members of Working Group 3,

Considering the Secretary General's Our Common Agenda Policy Brief 8

Underscoring the aspirations of the international community to uphold information integrity in the peaceful use of communications technologies for the common good of humankind and to further the sustainable development of all countries, irrespective of their economic, scientific and technological progress,

Noting the rapid spread of misinformation, disinformation and hate speech on digital platforms worldwide, resulting in an 'infodemic' that harms human lives,

Highlighting the essential cooperation of States and other stakeholders in regulating, formulating and adhering to new universal measures for information integrity on digital platforms,

Propose this Code of Conduct for consideration and adoption by UNHQ New York within the formal drafting process for the future UN Code of Conduct on Information Integrity, which is to be presented at the Summit of the Future in 2024

Andessa Maria Alves Santos
Brazil

Mahamade Hamine Ouedraogo
Burkina Faso

Elizabeth Collin-Paré
Canada

Melanie Echeverria Bracamonte
Guatemala

Ayush Garg
India

Sirine El Halabi
Lebanon

Ahmed Hussain
Pakistan

Luisa B. T. Azevedo
Portugal

Phirapat Mangkhalasiri
Thailand

Diana Kuznetsova
Ukraine

Samuel Se-Hyun Lee
United States of America

Victoria Pasquet
Uruguay

Introduction

A CODE OF CONDUCT

The Secretary General's Our Common Agenda Policy Brief 8 on Information Integrity on Digital Platforms calls for a Code of Conduct to be presented at the Summit of the Future.

The present document is put forward by Working Group 3 of the United Nations 61st Graduate Study Programme, which has developed a Code of Conduct under the premise that the path towards stronger information integrity needs to be human rights-based, multi-stakeholder & multidimensional. The regulation of information integrity on digital platforms is a complex domain as it entails finely delineating the extent to which the right to freedom of speech can have limitations in the light of the increasing spread of misleading, false, or hateful online content with the potential to cause serious harm.

The UN High Commissioner for Human Rights, Volker Turk in his address to the 53rd session of the Human Rights Council, said "*The limitation of any kind of speech or expression must, as a fundamental principle, remain an exception – particularly since laws limiting speech are often misused by those in power, including to stifle debate on critical issues. But on the other hand, an act of speech, in the specific circumstances in which it occurs, can constitute incitement to action on the part of others — in some cases, very violent and discriminatory action*".

This Code of Conduct encourages all stakeholders, including Member States and digital platforms, to foster a digital environment that champions truth, serves as a verified filter against mis- and disinformation, and is unwavering in the fight against hate speech. It also purports to ensure that, in the pursuit of said objectives, the rights to freedom of thought, conscience, and expression, to privacy and nondiscrimination; maintain all their vigor throughout the digital space.

PURPOSE OF THIS DOCUMENT

The Code establishes guidelines for the regulation of the information diffused through digital platforms. It seeks to support independent media, protect it from unwarranted influences and safeguard its role as a critical watchdog of truthful information. It aims to promote transparency, so as to ensure digital platforms are trustworthy and the consent of its users is free and informed. It aspires to strengthen user empowerment, so that they have control over their personal data, they can discern truth from fabrication and opinions from hate and are able to freely express their ideas without fear of undue reprisal. It intends to enhance research and data access, to aid the process of devising effective, data-driven responses to the challenges of the future. It delineates scaled-up responses, commensurate with the gravity and scale of the issues at hand.

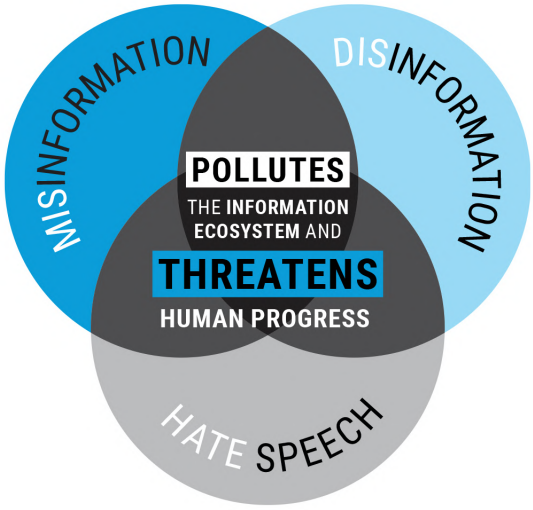
The Code proposes stronger disincentives for those who spread mis- and disinformation or engage in hate speech. It promotes a respectful digital dialogue instead and strives to enhance trust & safety online, in order to reinforce the belief that technology should be a positive force for all of humanity and not a stronghold for viral lies & amplified contempt toward individuals or groups. We are at a crossroads of the digital revolution. This Code of Conduct embodies the necessary principles for collectively shaping a united digital world that helps unite rather than divide; respect rather than denigrate. It encapsulates a shared vision of the digital landscape of the future: safer, respectful, inclusive, and truthful.

This Code of Conduct is, finally, a living document. As the digital world continues to evolve, so too must our efforts to ensure its integrity. We shall stand prepared to adapt, to learn, and to refine our approaches in the pursuit of promoting human rights across all borders and dimensions of human life.

What is information integrity?

Information integrity refers to the accuracy, consistency and reliability of information. It is threatened by disinformation, misinformation and hate speech. While there are no universally accepted definitions of these terms, United Nations entities have developed working definitions.

The Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression refers to disinformation as "false information that is disseminated intentionally to cause serious social harm". Disinformation is described by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as false or misleading content that can cause specific harm, irrespective of motivations, awareness or behaviors.



For the purposes of the Code of Conduct, the difference between mis- and disinformation lies with intent. Disinformation is information that is not only inaccurate, but is also intended to deceive and is spread in order to inflict harm. Disinformation can be spread by State or non-State actors in multiple contexts, including during armed conflict, and can affect all areas of development, from peace and security to human rights, public health, humanitarian aid and climate action.

Misinformation refers to the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods. Misinformation can be rooted in disinformation as deliberate lies and misleading narratives are weaponized over time, fed into the public discourse and passed on unwittingly.[4] In practice, the distinction between mis- and disinformation can be difficult to determine.

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".

Mis- and disinformation and hate speech are related but distinct phenomena, with certain areas of overlap and difference in how they can be identified, mitigated and addressed. All three pollute the information ecosystem and threaten human progress.

Threats to information integrity are not new. Falsehoods and hatred have long been spread for political or financial gain. Yet in the digital age these operations can be conducted on a previously unthinkable scale. Powerful communication tools can now spread content instantly across the globe, creating a problem so widespread that online platforms themselves are at times unable to grasp its full extent. The lack of governmentally agreed definitions of these terms should not result in inertia. We must do all we can to curb the harms they cause.

Information integrity and digital platforms

Digital platforms should be integral players in the drive to uphold information integrity. While certain traditional media can also be sources of mis- and disinformation, the velocity, volume and virality of their spread via digital channels warrants an urgent and tailored response. For the purposes of the present Code of Conduct, 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.

Mis- and disinformation are created by a wide range of actors, with various motivations, who by and large are able to remain anonymous. Coordinated disinformation campaigns by State and non-State actors have exploited flawed digital systems to promote harmful narratives, with serious repercussions.

Many States have launched initiatives to regulate digital platforms, with at least 70 such laws adopted or considered in the last four years. At their core, legislative approaches typically involve a narrow scope of remedies to define and remove harmful content. By focusing on the removal of harmful content, some States have introduced flawed and overbroad legislation that has in effect silenced "protected speech", which is per- mitted under international law. Other responses, such as blanket Internet shutdowns and bans on platforms, may lack legal basis and infringe human rights. Many States and political figures have used alleged concerns over information integrity as a pretext to restrict access to information, to discredit and restrict reporting, and to target journalists and opponents. State actors have also pressured platforms to do their bidding under the guise of tackling mis- and disinformation.

The risks inherent in the regulation of expression require a carefully tailored approach that complies with the requirements of legality, necessity and proportionality under human rights law, even when there is a legitimate public interest purpose.

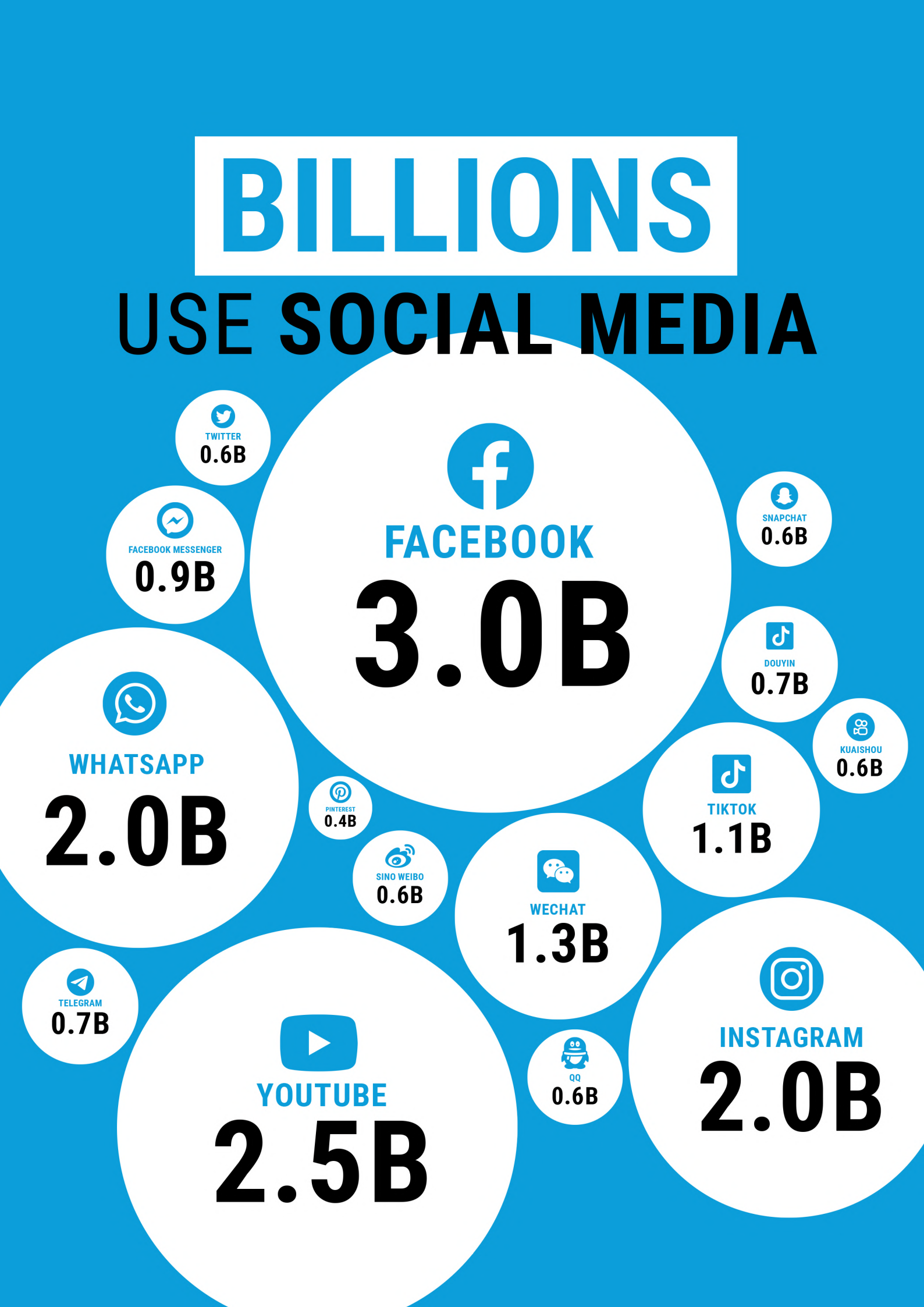
Disinformation is also big business. Both "dark" and mainstream public relations firms, contracted by States, political figures and the private sector, are key sources of false and misleading content. One tactic, among others, has been to publish content to fake cloned versions of news sites to make articles seem like they are from legitimate sources. This shadowy business is extremely difficult to track and research so that the true scale of the problem is unknown. Individuals, too, spread false claims to peddle products or services for profit, often targeting vulnerable groups during times of crisis and insecurity.

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 & hate speech.

Facing a decline in advertising revenue, digital platforms are seeking alternative avenues for profit beyond the attention economy. For example, paid verification plans, whereby accounts can buy a seal of approval previously used to denote authenticity, have raised serious concerns for information integrity given the potential for abuse by disinformation actors

BILLIONS

USE SOCIAL MEDIA



Towards a United Nations Code of Conduct

This United Nations GSP Code of Conduct for Information Integrity on Digital Platforms is built upon the following principles:

- **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**
- **Enhanced trust and safety**

Commitment to Information Integrity

1.1 To combat information pollution, mis- and disinformation, and to promote information integrity, the following actions shall be taken at different levels:

Individuals:

- Encourage individuals to verify the accuracy of information before sharing it.
- Establish mechanisms for individuals to report misleading or false content.
- Promote media literacy and critical thinking from a young age and throughout the lifecycle

Media outlets:

- Establish mandatory, ethical and fact-checking standards for the media.
- Require the media to publish corrections or retractions in case of errors.
- Enhance transparency regarding the sources of information used by the media.
- Encourage cooperation and synergies for cross media verification of information.

Digital platforms:

- Impose the swift removal of clearly false, misleading, or harmful content.
- Require digital companies to comply with ethical information sharing laws in the Member States in which they operate.
- Demand that recommendation algorithms prioritize reliable and verified content.
- Strengthen monitoring of accounts and bots involved in disseminating disinformation.

Respect for Human Rights

2.1 Universal Declaration of Human Rights

The UDHR outlined a collection of principles and rights that are universally applicable for every human being. After 75 years, it remains the cornerstone of the human rights system. Member States, digital platforms and other stakeholders must be aware of the following effects that the dissemination of mis- and disinformation, and hate speech are having on human rights:

Right to Non Discrimination: *Article 2*

Mis-, disinformation, and hate speech can motivate discrimination by spreading prejudices and through hate speech's targeting of certain groups based on race, color, ethnicity, gender, language, religion or political opinion, and other characteristics.

Right to a Fair Trial: *Article 10*

The dissemination of mis- and disinformation has the capacity to erode the fundamental right to a fair trial.

Right to Privacy: *Article 12*

The dissemination of mis- and disinformation can lead to privacy violations, which can expose individuals to emotional or physical harm.

Right to Freedom of Thought, Conscience & Expression: *Article 18*

The dissemination of mis- and disinformation can impede individuals' capacity to openly express their thoughts and opinions while also hindering their access to reliable information that is crucial for the development of ideas.

Right to Freedom of Expression *Article 19*

The right to freedom of expression is not an absolute right. The dissemination of mis- and disinformation can cause physical harm.

Right to Information: *Article 19*

The dissemination of mis- and disinformation can hinder the right to access and share accrued information.

Right to Health: *Article 25*

The dissemination of mis- and disinformation concerning health can undermine public health initiatives, hinder access to reliable medical information, impact mental health, and contribute to the proliferation of inaccurate treatments, as was most recently seen during the COVID-19 pandemic.

Right to Education: *Article 26*

The dissemination of mis- and disinformation can mislead and distort facts, impacting the quality of education and adversely impact future generations.

Right to Participate in Cultural Life
Article 27

The dissemination of mis-, disinformation, and hate speech can perpetuate stereotypes, marginalize individuals from cultural activities, and create false cultural narratives.

2.2 Sustainable Development Goals

Member States, digital platforms, and other stakeholders should recognize that the dissemination of mis- and disinformation affects the fulfillment of the SDGs, particularly *climate action*.

The detrimental effects of spreading false or misleading information extend to multiple aspects, such as –

- Impeding poverty eradication efforts
- Exacerbating societal divisions, with a particularly harmful impact on marginalized, vulnerable communities
- Perpetuating economic, social, and political exclusion
- Impacting quality of education
- Fueling violence against women
- Detrimental to Global economic growth
- Impeding innovation & development
- Creating a polarized society
- Negatively influencing development
- Undermining the importance of a sustainable environment
- Hampering efforts to address the climate change crisis
- Negatively Influencing elections
- Disrupting the work carried out by institutions
- Impeding partnerships & cooperation on SDGs.

2.3 Principles and Commitments

This Code of Conduct stipulates principles and commitments that Member States, digital platforms, and other stakeholders must implement to address harmful impacts of threats to information integrity, in compliance with international law and standards.

Member States should –

- Counter mis-, disinformation, and hate speech by respecting, protecting, and fulfilling human rights, especially the rights to freedom of opinion, freedom of expression, and the right of access to information
- Protect the fundamental rights of users of digital platforms, particularly the right to privacy & personal data self-determination, and assure the implementation of enforcement mechanisms, remedies, and other judicial, administrative, and legal measures to fulfill human rights or redress human rights violations.
- Apply a multistakeholder approach to counter mis-, disinformation, and hate speech.
- Member States should take appropriate steps to prevent, investigate, punish, and redress abuse of human rights on digital platforms.
- Guarantee the right to digital literacy, from an early age, as an enabling right to freedom of expression and as part of the human right to education.
- Member States should also ensure right to participate in public affairs.

- Ensure effective access to information, including diverse sources and independent media. Laws pertaining to information transparency and responses to combat threats to information integrity, such as defamation, cyber bullying, and harassment, should align with established human rights norms and standards.
- Incorporate international standards and laws related to information integrity and freedom of expression in national legislations, especially regarding limits on incitement to discrimination, hostility and violence.
- Refrain from internet shutdowns and bans on platforms or media outlets unless these actions are deemed absolutely necessary and comply with the requirements of general restrictions of rights under international law, including necessity, proportionality, and legality.
- Ensure that regulations encompass both public and private actors with special attention to digital platforms. Power imbalances should be addressed between digital platforms and individuals.

Digital Platforms should –

Implement and disclose policies of information integrity to guarantee the access of information and the transparency of:

- Collection and use of data
- Content moderation
- Advertising
- Implementation of human rights impact assessments and due diligence
- Content removals and suspension of online accounts
- Use of algorithms
- Governmental requests in relation to mis- and disinformation and hate speech
- Assure the implementation of redress mechanisms and independent reviews according to the 'United Nations Guiding Principles on Business and Human Rights'.
- Install independent oversight bodies and mechanisms conformed by technical, legal, and operational experts from different fields to take decisions regarding content moderation and elimination and the respect of human rights

Other Stakeholders –

International non-governmental organizations, influential thinkers, youth representatives, local and community leaders, the United Nations System and its partners, and others should:

- Promote diversity, security, and inclusion in the digital space
- Undertake efforts to establish and support reporting and fact-checking mechanisms
- Refrain from committing, spreading, or endorsing mis- & disinformation and hate speech.
- Encourage digital users' empowerment.

Support for Independent Media

3.1 To maintain Editorial Independence

Member States should –

- Establish a permanent government budget dedicated to funding public media initiatives which would guarantee autonomy of editorial content
- Condemn and refrain from the practice of censorship of the media
- Develop and implement legal frameworks that guarantee the protection of media freedom and prevention of censorship. Member States should create clear legislative ground for legal action towards media outlets that are proven to repeatedly and voluntarily disseminate mis-, disinformation, and hate speech.

Media outlets should –

- Diversify their funding sources in order to decrease their dependence on several large sponsors
- Prioritize funding sources that do not compromise the integrity of published content by exercising influence over editorial decisions.
- Maintain clear guidelines on fact-checking, verification and collaboration of news reports.

3.2 To strengthen a Plural Media Landscape

Member States should –

- Develop national strategies to guarantee an enabling environment for journalists to thrive
- Provide funding opportunities to small and medium media initiatives that produce and publish content in local languages and represent minorities
- Develop and implement legal frameworks that increase the hiring of journalists from diverse backgrounds

Media outlets should –

- Increase diversity inside their newsrooms through diverse hirings
- Ensure content accessibility for people with disabilities (i.e., sign language)
- Develop and implement legal frameworks that enable negotiations between technology companies and media outlets regarding the payment of journalistic content on digital platforms.

3.3 To empower Fact-Checking

Member States should –

- Facilitate information transparency and access to fact-checking initiatives
- Create funding opportunities for local fact-checking initiatives, across national and local languages
- Provide training programmes to fact-checkers in alignment with international fact-checking standards.
- Provide funding opportunities to small and medium media initiatives that produce and publish content in local languages and represent minorities

Media outlets should –

- Grow and continuously train their fact-checking workforce
- Invest in new fact-checking technology resources, especially in rapidly evolving artificial intelligence technologies
- Pursue and adhere to the international fact-checking standards
- Develop coordinating mechanisms within the journalistic community to enable cooperation and rapid fact checking in times of crises and emergencies.

3.4 To increase Protection of Journalists

Member States should –

- Support awareness-raising initiatives aimed at the population as a whole about the importance of respecting media freedom
- Provide continuous training of the judiciary and law enforcement on international human rights, international humanitarian law obligations, and commitment regarding the safety of journalists
- Condemn and not in any way engage in online or offline attacks, harassment, and violence against journalists and other media professionals
- Create an efficient process to properly investigate and prosecute online or offline attacks against journalists and other media professionals
- Immediately release any journalist or media professional who has been arbitrarily detained or incarcerated without due legal process.

Media outlets should –

- Provide legal and psychosocial assistance to any employee who has been the victim of online or offline harassment, attacks, or violence
- Provide continuous training of cyber-safety and self-safety, especially, but not exclusively, in armed conflict contexts.

Digital Platforms should –

- Develop and implement concrete plans of action to combat online harassment, attacks, or violence targeting journalists and media workers
- Collaborate efficiently with media outlets and the States on the investigation of online harassment, attacks, or violence targeting journalists and media workers, with the view of promptly identifying perpetrators and preventing future recurrences.

Increased Transparency

Member States should regulate minimum transparency standards for digital platforms' policies and practices. These should be available, accessible, concise, and intelligible to users by using plain and clear language. User consent should not be valid if the information preceding it is not clear and understandable to an average user.

For this purpose, Member States should work with digital platforms to implement the following measures, which include, but are not limited to:

4.1 Algorithms

- Prioritize the use of intelligible algorithms, whenever possible
- Disclose, when known, the criteria and information utilized for algorithmic decision-making

4.2 Data

- Take appropriate measures to provide users with any information referring to the collection, processing, and storage of users' personal data
- Keep the procedures through which user requests related to the use of personal information are processed as short, straightforward, and as simple as possible
- Provide the identity and contact details of the entities accountable for the processing of personal information
- Develop certification mechanisms and data protection seals or marks, which allow data subjects to assess the level of data protection of sites to which they are redirected through ads and/or promotions.

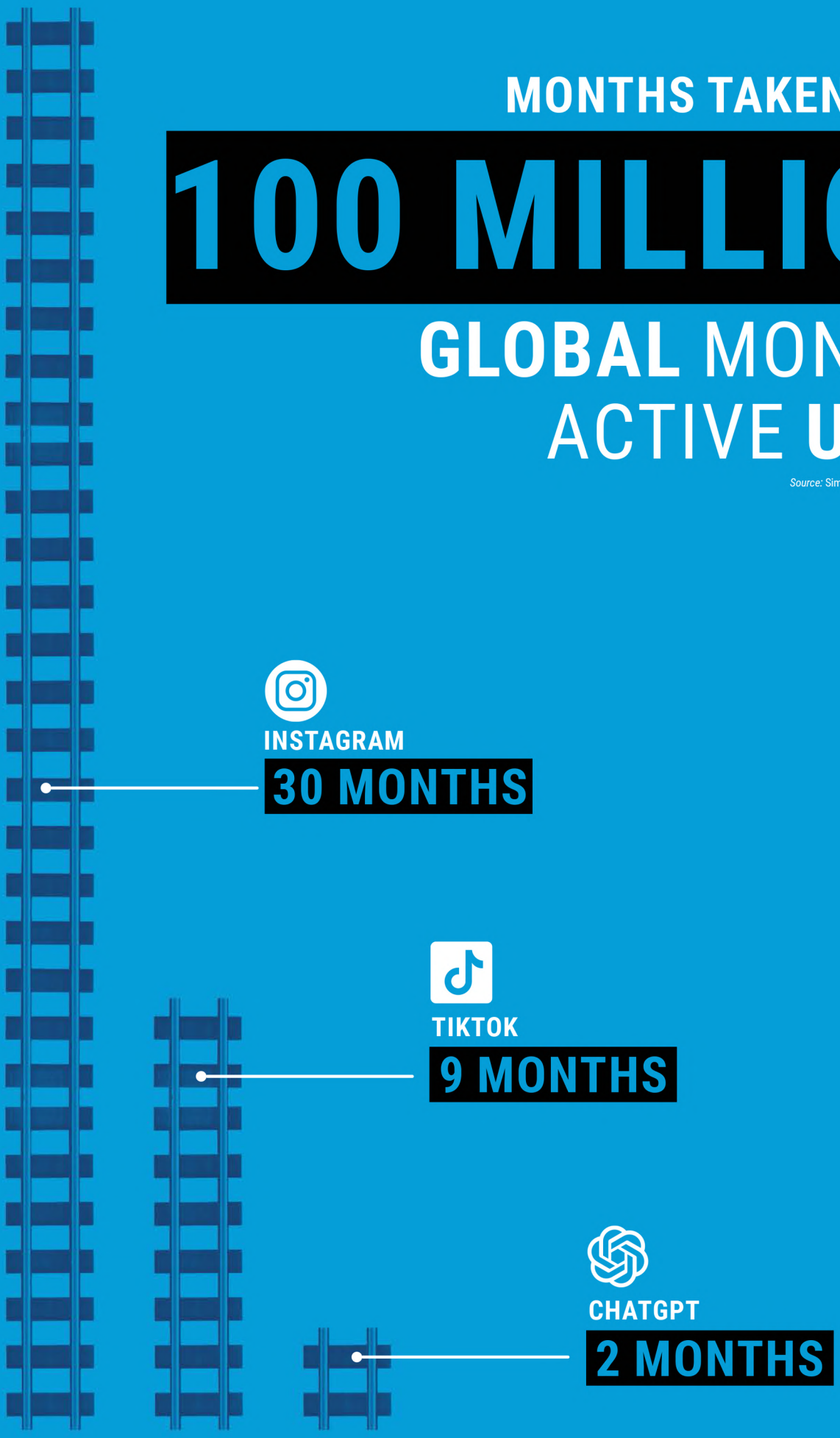
4.3 Content Moderation

- Make the standards to be observed in user-generated content readily available for users, including clear criteria for content filtering on the grounds of mis- and disinformation, and hate speech
- Communicate the changes in said standards and guidelines effectively to users well in advance of their implementation
- Explicitly disclose when artificial intelligence is employed for content moderation
- Provide access to appropriate, effective, simple, and transparent procedures to contest bans, suspensions, account deletions, or other sanctions

- Decide on sanctions and communicate them to users within reasonable time frames and on an individual basis
- Publish quantitative and qualitative information about the status and results of appeals received, treated, accepted, and rejected.
- Report periodically to Member States on the outcomes of their policies and practices aimed specifically at combatting mis- and disinformation and hate speech.

4.4 Advertising

- Disclose advertising policies and practices, as well as the sources of funding for all advertisements
- Separate distinctly, utilizing appropriate labels, paid promotional content from independent editorial or news content
- Disclose the criteria for targeted advertising upon the request of both advertisers and the target audience, including specific information about targeting algorithms, the audience reached, and the overall effectiveness of advertisements
- Periodically publish policies, aimed at restricting the revenues of entities spreading mis- and disinformation and hate speech, related to the scrutiny of advertisement placements as well as their outcomes
- Provide users with simple mechanisms to report misleading advertisements and take prompt action when such reports are made
- Offer tools and features to enhance users' understanding and control of the advertisements they are exposed to
- Work in conjunction with independent bodies to periodically audit advertising practices.


MONTHS TAKEN TO ADD
100 MILLION
GLOBAL MONTHLY ACTIVE USERS
Source: Similarweb, using data from Sensor Tower
INSTAGRAM
30 MONTHS
TIKTOK
9 MONTHS
CHATGPT
2 MONTHS

User Empowerment

5.1 Promotion of Government Transparency and Accountability

- Member States should ensure public access to accurate, transparent, and credibly sourced government information
- Use descriptive texts and sign language to cater to persons with different disabilities and enhance accessibility
- Implement a robust system of checks and balances that monitors and ensures the credibility of disseminated information on digital platforms. Non-compliance by digital platforms should trigger corrective measures by Member States to uphold the principles of transparency.

5.2 Freedom of Opinion

- Member States and digital platforms should have an effective and responsive content evaluation process to swiftly identify false content or content that has a clear intent to misinform
- Member States should pursue a judicial framework, pending an independent judicial process, where governments can challenge decisions by digital platforms to filter certain content on a case-by-case basis

5.3 Digital Tools for Empowered Interaction

- Digital platforms should provide tools that facilitate user customization, enabling interactive online experiences, content discovery, and access to varied news sources
- Make available user-friendly and accessible tools for reporting disinformation, thus helping to dilute the visibility of disinformation and to improve accessibility to trustworthy content
- Present terms and conditions agreements under digital platforms in plain and simple language for users' ease of understanding
- List terms and conditions related to human rights, including the right to privacy, at the beginning of agreements to emphasize their importance to the user

5.4 Prioritization of Relavant and Authentic Information

- Digital platforms should invest in technological means to prioritize relevant, authentic, authoritative, scientific, and verified information in search, feeds, or other automatically ranked distribution channels
- Transparency is vital, allowing users to understand why they are shown specific content or advertisements
- Give visibility to media news outlets that have a record of truthful and unbiased reporting. Digital platforms can consult with relevant domestic bodies of Member States to identify trustworthy media outlets, which can include publicly funded outlets

5.5 Trust Indicators

- Digital Platforms should collaborate with news media associations to develop and implement indicators of trustworthiness, such as media ownership and verified identity, established with journalistic principles.
- Transparency about these indicators can facilitate user assessment of content

5.6 Youth Focused Digital Literacy and Rights Awareness

- All relevant stakeholders should partner in efforts to enhance an understanding of everyone's rights in online spaces, the workings of digital platforms, their personal data usage, and ways to identify and respond to mis- and disinformation
- Develop initiatives to teach digital literacy for children as soon as they are in contact with digital platforms. Training literacy
- Digital literacy programmes should include a human rights education component.
- Programmes should continue from kindergarten education to late stages of life and be frequently evaluated to evolve with digital mediums in real time.
- Develop specific child-friendly materials and tools to ensure the message is conveyed effectively, with nuanced attention given to young people, adolescents, and children of various ages.

5.7 User Rights and Privacy

- Users should be clearly informed about their privacy options from the beginning of their use of a digital platform and have the ability to opt-out at any time
- Upholding the "right to be forgotten" and enabling users to exert control over their shared data is crucial
- The implementation of principles empowering users to retract or select specific aspects of shared data must be prioritized.

5.8 Cultural Sensitivity and Inclusivity

- Member States and digital platforms should dedicate resources to ensure that terms and conditions agreements and other user empowerment materials are in a local language understood by the local user population.
- Considering the diversity of digital users around the world, cultural context and differences should inform all aspects of user empowerment.
- Culturally sensitive and inclusive practices should ensure equal access and opportunities to all users, including youth.

5.8 Monitoring and Accountability

- Member States and digital platforms should bear responsibility for the enforcement and success of provisions regarding ongoing monitoring and assessment, which is essential for effective user empowerment
- Solicit feedback from users routinely and integrate received feedback into future policies and practices
- Seek user feedback through focus groups that represent and reflect the user population in a certain country. These periodic consultations should be thoroughly conducted to understand users' feedback and incorporate it into Member States and digital platform policy and practice

Strengthened Research and Data Access

6.1 Member States should:

- Actively invest in and support independent research, studying the prevalence and impact of mis- and disinformation, as well as hate speech, on digital platforms, across various countries, regions, and languages.
- Encourage partnerships and research collaboration between digital platforms, civil society, and academia to jointly combat mis- and disinformation and hate speech.
- Establish long-term, diversified, and sustainable funding mechanisms, including public-private partnerships, for the research on mis- and disinformation and hate speech across all digital platforms. Funding should be structured in a manner that ensures the independence, integrity, and credibility of the research, with clear agreements to prevent any undue influence

6.2 Digital Platforms should:

- Provide data access to researchers and academics with due respect for user privacy. This will foster a better understanding of the extent and nature of the harm caused while respecting data protection and human rights
- Commit to transparent data-sharing practices with guidelines in place specifying what data can be shared, how it will be shared, and under what circumstances. Accountability should be enhanced through periodic reporting.

6.3 Civil Society, Member States and digital platforms should:

- Advocate and take necessary steps for the full participation of civil society and academia in combating mis- and disinformation and hate speech on digital platforms through dialogues & participation in decision-making processes.
- Emphasize the respect for data protection laws adhering to privacy safegaurd techniques such as data encryption and anonymization under robust ethical guidelines.

Scaled up Responses

7.1 Allocating resources and further investment

Member States should –

- Allocate sufficient resources to tested media outlets that have a proven track records of spreading factual information
- Encourage independent third-party fact checking services, reallocating funds to them while preserving and respecting the parties' independence and integrity

Digital platforms should –

- Allocate sufficient resources into their own mis- and disinformation content detection
- Harness the growing power of AI to prevent and reduce the spread of mis- and disinformation. AI algorithms are powerful tools to detect mis- and disinformation.

7.2 Collaboration

Member States should –

- Collaborate with digital platforms, academia, think tanks, researchers, civil society organizations, and international organizations in order to reduce the spread of mis- & disinformation
- Collaborate with established media outlets to effectively combat mis- and disinformation across all platforms

Media outlets should –

- Implement algorithm strategies that promote fact-checking content in order to burst digital bubbles and echo chambers and increase the reach to a significant percentage of the broader public

7.3 Training and capacity building

Member States should –

- Invest in researching and developing curricula that help integrate digital literacy into all educational institutions
- Ensure all citizens learning and adapting with digital literacy

Digital platforms & Academia should –

- Provide opportunities, financially and logistically support, and encourage the work of independent researchers working on preserving information integrity.

Stronger Disincentives

All stakeholders acknowledge the important role that businesses have played in innovation and development of the digital platforms that allow people to connect worldwide. Furthermore, digital platforms recognize their corporate responsibility to respect human rights. They should strive to integrate the following best practices:

8.1 Engagement on human rights, privacy and safety

Digital platforms should –

- Ensure that their human content moderators receive the necessary psychological support and care
- Ensure the protection of user information
- Provide a safe environment for users.
- Develop content management methods using ethical artificial intelligence technology for future enhancements

8.2 Engagement on advertisements

Member States should –

- Ensure that advertisements are not placed next to online mis-, disinformation, or hate speech
- Prohibit the promotion of advertising containing disinformation
- Ensure transparency regarding financing of political advertisements on their platforms.

8.3 Engagement on paid advertising

Digital platforms should –

- Mark clearly all media content on their platform that includes paid advertising and advertorial content
- Take concrete measures to prevent and address the spread of content that contains mis- and disinformation, as well as hate speech

Member State & digital platforms must –

- Provide access to effective remedies to users through judicial and non-judicial grievance mechanisms in accordance with the principles elaborated on the "Guiding Principles on Business and Human Rights, 2011"

Enhanced Trust and Safety

9.1 Digital security and trust

Member States & Digital platforms should –

- Take measures to secure citizens' data, building institutional and public trust
- Prioritize safety and privacy by incorporating these elements into the product design of Digital Platforms. This can be achieved through sufficient resourcing of in-house trust and safety expertise
- Ensure security by requiring Digital Platforms to implement privacy features such as two-factor authentication & alerts
- Enhance digital trust, which refers to the level of confidence that users have in the security, privacy, and reliability of digital platforms. Such Platforms should establish transparent data handling policies wherein users should clearly know how their data is being used and managed
- Implement laws and procedures that ensure user protection, industry safety standards, and digital platform accountability for any data breaches.

9.2 Use of ethical AI in digital platforms

Digital platforms should –

- Make efforts to train artificial intelligence to moderate content in as many languages as possible around the world, making all digital platforms accessible to all humanity
- Expertise of the *UN Inter-Agency Working Group on Artificial Intelligence* (IAWG-AI) should be taken into consideration when developing the ethical protocols for the use of AI algorithms, pooled datasets, interoperability standards, regulatory sandboxes, and computational capacity by digital platforms

Member States should –

- Appoint Technology Ambassadors that act as a focal point between their respective States and digital platforms to address and tackle issues pertaining to the ethical use, understanding, and legal framework of AI as applicable to national frameworks
- Encourage digital platforms to incentivize collaboration around data and AI and turn the focus away from competitive approaches. This can be done by adopting the principle of '*Open, Free and Secure digital future for All*'

9.3 Inclusion of digital platforms into an information integrity framework

Member States should –

- Member States and digital platforms should ensure that digital platforms meet the three-part test of legality, proportionality and necessity when applying restrictions to or limiting online content to tackle mis-, disinformation and hate speech. The UN Rabat Action Plan (2012) can be used by digital platforms in this regard
- Digital platforms, Member States, and stakeholders should work together in ensuring the UN's aim of achieving universal connectivity by 2030 is met. Connecting all people in all languages to all digital platforms should be a core mechanism of achieving information integrity

Digital platforms should –

- Digital companies should provide for adequate and appropriate human resources for content moderation on digital platforms. They should ensure the availability of good health support for the human content moderators across platforms
- All stakeholders should support an information integrity framework that is trustworthy, human- rights based, safe, sustainable, and promotes peace
- To build digital trust, companies and stakeholders should adhere to the principles of '*Fairness, Transparency, Non-discrimination'* , and respect for Human Rights' on digital platforms.

INFORMATION INTEGRITY AND THE SUSTAINABLE DEVELOPMENT GOALS

As shown, threats to information integrity can have a negative impact on the achievement of the Sustainable Development Goals.

Mis- and disinformation continue to have implications for poverty eradication efforts and the global economy. Economic hardship can also fuel the spread of polarizing and hateful lies, including about marginalized groups. The cost-of-living crisis has been particularly fertile ground for the dissemination of disinformation falsely blaming the switch to renewable energy for soaring energy costs or job losses, for instance.

Threats to information integrity can compound global hunger, including by exacerbating conflict, climate change, disasters, poverty and inequality. Disinformation can deflect attention and distract from the challenges to global food security posed by conflicts.

During the coronavirus disease (COVID-19) pandemic, an infodemic of related mis- and disinformation undermined public health measures and vaccination drives. The threat to children's health and well-being resulting from exposure to harmful content persists.

Mis- and disinformation and hate speech can have an adverse impact on access to quality education, in particular for marginalized groups, including young women and girls. Access to information and digital media literacy drives to increase resilience will play a key role in limiting the societal impact of online harms.

Gender-based hate speech and disinformation seek to systematically subjugate women by silencing them and pushing them out of the public sphere. They can have devastating consequences, from suppressing women's voices and fuelling self-censorship, to causing professional and reputational damage and inciting physical violence.

Two billion people live without safely managed drinking water services. Mis- and disinformation about the safety of drinking water and sanitation can have dangerous health consequences.

Climate mis- and disinformation, much of it seeded by the fossil fuel industry, are undermining the urgent transition to cleaner forms of energy production, narrowing the closing window to deliver a sustainable future for all.

Research has pointed to the detrimental impacts of mis- and disinformation and hate speech on the economy.

Mis- and disinformation and hate speech, and overbroad responses to these phenomena, can have a detrimental impact on innovation, including by limiting the potential of marginalized groups and making digital spaces less equal and inclusive.

Mis- and disinformation and hate speech spread online are polarizing societies and targeting already marginalized and vulnerable communities, and can result in their further social, economic and political exclusion.

Efforts to make cities and communities more sustainable can be undermined by disinformation that denies or deflects attention away from the impacts of human activity on the environment.

Activists behind initiatives to foster a circular economy and boost zero-waste practices have been targeted through online hate speech and disinformation.

Climate disinformation, and the inertia that it encourages, is undermining efforts to take urgent action to address the climate crisis, including by impeding the crucial shift from polluting fossil fuels to renewable energy and urgent investments in climate resilience.

Mis- and disinformation can have a negative impact on efforts to conserve and sustainably use the oceans, seas and marine resources.

Environmental activists working to protect life on land have been targeted by online hate and disinformation campaigns, with real-life consequences. Climate mis- and disinformation are undermining climate action efforts.

Disinformation and hate speech have been used to influence elections and public narratives and sow confusion. They have been used to undermine adversaries, thwart peacemaking efforts, incite violence, prolong conflict and damage trust in the rule of law. Efforts to promote peaceful and inclusive societies, and the role of the United Nations in supporting peace and security, have been seriously affected as a result.

Mis- and disinformation and hate speech can hinder meaningful partnerships to achieve the Goals, while resources diverted to address the problem can weaken efforts to leave no one behind.

Conclusion

Strengthening information integrity on digital platforms is an urgent priority for the international community. From health and gender equality to peace, justice, education and climate action, measures that limit the impact of mis- and dis- information and hate speech will boost efforts to achieve a sustainable future and leave no one be- hind. Even with action at the national level, these problems can only be fully addressed through stronger global cooperation. The core ideas outlined in this Code of Conduct demonstrate that the path towards stronger information integrity needs to be human rights-based, multi-stakeholder, and multidimensional. They have been distilled into a number of principles to be considered for a United Nations Code of Conduct for Information Integrity on Digital Platforms at the Summit of the Future; that will provide a blueprint for bolstering information integrity while vigorously upholding human rights.

AUTHORED BY

BRAZIL
Andessa Maria Santos

BURKINA FASO
Mahamade Hamine

CANADA
Elizabeth Collin-Pare

GUATEMALA
Melanie Echeverria

INDIA
Ayush Garg

LEBANON
Sirine El Halabi

PAKISTAN
Ahmed Hussain

PORTUGAL
Luisa Azevedo

THAILAND
Phirapat Mangkhalasiri

UKRAINE
Diana Kuznetsova

UNITED STATES OF AMERICA
Samuel Se-Hyun Lee

URUGUAY
Victoria Pasquet

EDITED BY

Miladin Bogetic
UN Information Service

Working Group 3
61ST Graduate Study Programme

UNOG