GLOBSEC
IDEAS SHAPING THE WORLD

United Nations Code of Conduct for Information Integrity on Digital Platforms

Policy Brief

www.globsec.org

Introduction

In an era defined by the pervasive influence of digital platforms, the rapid dissemination of information has become both a cornerstone of connectivity and a potential catalyst for societal challenges. As we navigate the vast landscape of the digital age, the threat posed by these platforms in terms of the spread of harmful content and disinformation looms large, demanding a nuanced understanding and a strategic approach to mitigate its impact.

Digital platforms, encompassing social media, messaging applications, and content-sharing networks, have undeniably revolutionized the way information is produced, consumed, and shared. However, the democratization of information dissemination has come at a cost, as these platforms have become breeding grounds for the proliferation of harmful content and the rapid spread of disinformation. The interconnected nature of these platforms amplifies the reach and velocity at which such content can permeate societies, often with far-reaching consequences for public discourse, political stability, and individual well-being.

The threat of harmful content encompasses a spectrum ranging from explicit violence, hate speech, and extremist ideologies to more subtle forms such as cyberbullying and online harassment. The unrestricted dissemination of such content not only poses a direct threat to the safety of individuals but also erodes the fabric of societal cohesion, leading to **extremely polarized** societies.

Simultaneously, the negative influence of disinformation on digital platforms has emerged as a formidable challenge. Whether fueled by state actors, malicious entities, or simply the unintended consequences of algorithmic biases, disinformation campaigns have the potential to manipulate public opinion, undermine trust in institutions, and even impact democratic processes.

Recognizing the detrimental impact of disinformation and harmful content, including hate speech and incitement to violence on societies worldwide, GLOBSEC welcomes the United Nations' initiative to develop a voluntary Code of Conduct for Information Integrity on Digital Platforms. Leveraging our experience in studying the impact of disinformation on societies in Central, Eastern and Southeast Europe (e.g. through regular **GLOBSEC Trends** polling), and our involvement in co-shaping EU-level regulation as a co-signatory of the EU Code of Practice on Disinformation (CoP) and in drafting the strengthened code adopted in 2022, we are well positioned to offer data-based insights into the global framework.

Reflections on the UN Code of Conduct Commitments

The Code of Conduct for Information Integrity on Digital Platforms is being developed by the UN in the context of preparations for the Summit of the Future. The Code of Conduct aims to provide a gold standard for guiding action to strengthen information integrity. Member States will be invited to implement the Code of Conduct at the national level. The UN is conducting consultations with stakeholders to further refine the content of the Code of Conduct, as well as to identify concrete methodologies to operationalize its principles. Therefore, GLOBSEC would like to contribute to these efforts with the following recommendations to the UN divided based on commitments outlined in the Code:

Commitment to information integrity:

- **Focus on demonetization:** According to several studies published by **ProPublica**, **Global Disinformation Index**, or a **consortium** of Slovak and Czech researchers, disinformation is often spread for economic benefit. We consider it important to recognize the *demonetization of disinformation* as one of the key aims of the Code, which is missing in the proposed text.

- **Define stakeholders and their role:** The clarification of who is considered as a stakeholder under the Code is essential for the Code's implementation. We therefore suggest adding a definition of a "stakeholder" to the Code, with the potential to reference definitions used in **UNESCO's Guidelines for the Governance of Digital Platforms** ("The Guidelines").

- **Require equal enforcement of terms and conditions:** To protect users and the information environment from harmful content, digital platforms should be required to uniformly enforce internal guidelines for all content published on their platforms, including by political representatives. At the moment, for example, Facebook treats content published by politicians as **"newsworthy" content** that should, as a rule, be seen and heard, even if breaking the community standards. Considering that the users of digital platforms are not able to see balanced content due to the algorithmic setup promoting content with the highest engagement potential, we consider the exemption for political speech as dangerous, as it allows for hate speech and violent content to be used as a part of pre-election campaigning. Such an exemption should only be applied in exceptional cases when necessary for informing citizens and/or the international community.

Commitment to respect for human rights:

- **Require incorporation of human rights protection in terms and conditions:** While emphasizing the importance of protecting human rights, the Code currently places the responsibility on the Member States, overlooking the potential violation of human rights by the state actors. To address this issue, we propose including provisions that mandate stakeholders, including digital platforms, to incorporate human rights protection in their terms and conditions, statutes or mission, and ensure the fulfilment of these obligations in the services they provide. Such provisions would enhance user protection against laws that may ostensibly target disinformation but are employed as tools to restrict online freedom of expression (such as the Turkish **Disinformation Law**, criticized by opposition parties or the Venice Commission), as well as against mass amplification of hateful content, as witnessed in the **case** of hate speech targeting Rohingya people in Myanmar by accounts linked to the military and far-right extremists, resulting in mass casualties and a refugee crisis.

- **Reference international human rights laws:** While the **EU's Code of Practice on Disinformation** leaves the definition of illegal content to member states, the UN Code of Conduct is set to be universally applied, including in authoritative regimes. Therefore, it would be preferable for it to draw upon international human rights law instruments to delineate the illegality of online content.

Commitment to support for independent media:

- **Define independent media and fact-checkers:** To prevent Member States from arbitrarily defining independent media or fact-checkers, the Code should draw upon existing initiatives and Codes as references. Examples include the International Fact-Checking Network, ensuring that organizations engaged in fact-checking or combating mis- and disinformation adhere to the highest standards of methodology, ethics, and transparency.

- **Establish a UN-based Fund for Information Integrity:** In pursuit of equitable and unbiased resource distribution, GLOBSEC recommends the establishment of a UN-based Fund for Information Integrity. This fund would serve as a crucial mechanism to support independent media, fact-checking organizations and vetted researchers globally. The Fund's support could derive not only from the Member States' contributions but also from voluntary contributions by digital platforms, with the funding decision-making process led by the UN, distributing funds on a proportional geographic basis.

Commitment to increased transparency:

- **Encourage transparency reporting:** Reporting from digital platforms should be encouraged in regular intervals on a per-member-state basis and in all major languages within each Member State, based on the proportionality of user base to a certain percentage of the population in a given country. Digital platforms should be encouraged to adopt similar reporting metrics, where feasible, to ensure comparability of data.

- **Require digital platforms to maintain transparency repository:** Each social media platform should maintain its own online transparency repository, offering citizens clear explanations as to why specific content was taken down or banned. Such transparency centers are instrumental in countering disinformation actors who often claim to be victims of censorship.

- **Require news media to disclose ownership:** Regarding the reporting requirements for news media, we suggest adding a transparency requirement for ownership structure, which would help prevent undue influence from the state, business entities, or foreign actors over specific media outlets.

Commitment to user empowerment:

- **Enhance transparency or reporting:** User empowerment should be enhanced in the context of content reporting and flagging by regular users. Once the content is flagged, there exists a notable lack of transparency in the subsequent actions taken. As the **SafeNet** project illustrated, platforms frequently overlook or fail to remove content reported by regular users, signalling a troubling trend of heightened non-responsiveness. Additionally, it is imperative to discourage platforms from imposing excessive demands on users during the reporting process, such as requesting users' private data.

Commitment to strengthen research and data access:

- **Ensure a fair vetting process:** GLOBSEC appreciates the inclusion of civil society actors into the clause advocating for data access. In many countries, non-governmental

organizations actively participate in researching and monitoring of digital platforms. Ensuring that data access is not restricted to a selected few but available to a diverse range of researchers globally strengthens collective efforts against disinformation and hate speech. While acknowledging the need for a vetting process, it is crucial to avoid leaving this responsibility solely to digital platforms or Member States to prevent politicization or the imposition of unattainable requirements. Instead, GLOBSEC recommends the establishment of an independent UN-led body to conduct and oversee such vetting processes.

- **Ensure free-of-charge data access:** Data access provided by the digital platforms should be free of charge to eliminate financial barriers that might impede research. For example, X (former Twitter) **discontinued** free API access in February 2023, **introducing** a tiered payment system that is not accessible to many non-profit organizations. We also recommend incorporating a reference to providing data in analyzable, and/or machine-readable formats without necessitating additional software investments. Researchers have faced challenges with the CoP reports **submitted** by digital platforms due to formats that are difficult to use for research.

- **Include access to "historical data":** Regarding the collection of data on individuals and groups targeted by harmful content, it is essential to thoroughly study the groups initiating and/or sharing such content to comprehend the dynamics of content sharing and networks involved. For this purpose, including access to "historical data" and data related to the content that has been taken down, banned or demoted would be an added value to the Code. This inclusion would significantly facilitate the work of researchers in countering harmful content.

Commitment to enhanced trust and safety:

- **Encourage platforms to conduct an assessment of AI-based moderation systems:** Evidence gathered globally **indicates** that AI-based content moderation systems are currently rather unreliable. For instance, findings from the SafeNet project, which monitors online illegal hate speech across 18 languages and national contexts **reveal** variations in platform responses to flagged content across platforms and countries. Some platforms, especially in certain languages, do not respond at all. To address this issue, the Code could encourage platforms to collaborate with vetted researchers for an assessment of the efficiency of AI-based moderation systems. This collaboration would aim to enhance human content moderation in languages where AI moderation falls short. Platforms should also ensure that the number of content moderators is sufficient and proportional to the number of users in a given language.

- **Facilitate discussion on what constitutes hate speech:** There is a notable disconnect in the understanding of what constitutes hate speech, incitement to violence, or the violation of human rights. This issue should also be addressed as well to foster a more cohesive approach to content moderation and regulation. Again, the already mentioned project SafeNet **reported** that the user experience in reporting hate speech on social media was unsatisfactory, exemplified by Facebook's failure to remove a comment endorsing the restoration of Auschwitz and the extermination of Roma, despite multiple reports. Similarly, TikTok did not take action against reported posts expressing admiration for Nazism or trivializing the Roma holocaust, highlighting shortcomings in addressing such content. This disparity underscores the importance of fostering an open and constructive dialogue between platforms and civil society to establish clear definitions of what constitutes a breach

of community standards or the dissemination of hate speech. Collaborative efforts in this regard will not only make community standards more publicly accessible but also enhance widespread understanding and effective enforcement. This in turn, will contribute to a more accountable and safer digital space for all users.

- **Define AI:** According to **EU DisinfoLab**, current inconsistencies in defining AI create varied mitigation and resolution measures for users and regulators. For example, only Facebook and TikTok explicitly mention the term "artificial intelligence" in their policies aimed at countering disinformation, while TikTok and X use the term "synthetic media". Additionally, platforms often prioritize addressing on images and videos, neglecting the same level of attention to AI-generated text. Therefore, the Code should include a clear and uniform definition of AI to guide platforms consistently.

- **Require all AI content to be visibly marked:** The Code should include a minimum requirement of visibly marking all AI-generated content. At the same time, it should encourage investments in technology capable of recognizing and labelling instances of misinformation, disinformation, or hate speech.

- **Ensure safety of children:** As **research** consistently underscores the adverse effects of social media on the mental health of youth, GLOBSEC recommends incorporating an additional section focused on the safety and protection of children. Specifically, platforms designed for children, such as YouTube Kids, should be prohibited from implementing recommendation systems. Moreover, social media profiles of users below the age of 18 should include parental controls and daily usage time limits to enhance their safety and well-being.

Assessment of existing instruments and their implications for the UN Code of Conduct

EU's Strengthened Code of Practice on Disinformation

The EU's Code of Practice on Disinformation ("**CoP**") serves as a self-regulatory initiative, reflecting a collective effort among major online platforms, emerging and specialised platforms, players in the advertising industry, fact-checkers, research, and civil society organisations to address the evolving challenges posed by disinformation. The CoP encompasses a wide array of commitments and measures, with signatories voluntarily committing to actions such as demonetizing the dissemination of disinformation, ensuring transparency in political advertising, empowering users through enhanced tools and media literacy initiatives, and providing increased support to researchers and the fact-checking community.

Recognizing the dynamic nature of the digital landscape, signatories established a permanent task force to ensure ongoing collaboration and adaptability. The CoP incorporates a monitoring framework, involving regular reporting on the implementation of its commitments. Notably, the CoP is poised to transition into the Code of Conduct under the Digital Service Act ("DSA") and being a signatory to the Code will be one of the mitigating measures against disinformation under the DSA.

So far, platform signatories have submitted two sets of reports, published online in February and September 2024, through the **Transparency Centre**, a requirement established by the CoP. Initial reports, however, appear inadequate, with seemingly incomplete data lacking context and sufficient detail. For instance, listing the number of removed posts in a country without a context is insufficient, as is documenting media literacy campaigns without information on their impact or reach. While these issues are not unresolvable, the introduction of structural indicators should address

them, providing researchers with a clearer and more comparable understanding of the data.

The CoP reporting highlighted that digital platforms, given their varied functionalities, may employ different methodologies for data collection. If the UN Code of Conduct mandates reporting from digital platforms, it should consider this diversity and strive for basic harmonization of data where feasible, enhancing comparability.

The co-regulatory nature of the CoP renders it a valuable source of data, fortifying a network of stakeholders. The UN Code of Conduct offers an advantage by potentially introducing consistent commitments globally, establishing a unified and comprehensive approach to addressing challenges posed by digital platforms on an international scale.

Guidelines for the governance of digital platforms (UNESCO)

In October 2023, UNESCO published *Guidelines for the Governance of Digital Platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach*. This document can serve as a valuable reference point for adopted solutions, outlining a set of duties, responsibilities and roles for a diverse range of stakeholders, including states, digital platforms, intergovernmental organizations, civil society, media, academia, the technical community, and other stakeholders. These roles aim to foster an environment where freedom of expression and information are central to the governance processes of digital platforms.

The UN Code of Conduct could benefit from using these Guidelines as a reference in various instances. For example, in defining stakeholders, the Guidelines provide detailed insights into the roles stakeholders play in the governance of digital platforms. Similarly, in section addressing the respect of human rights, the UN Code may refer to the Guidelines for potential approaches to the governance of digital platforms based on context.

Conclusion

In response to the escalating challenges of harmful content and disinformation on digital platforms, the United Nations Code of Conduct for Information Integrity is an important initiative for creating a safer online environment globally. GLOBSEC, with its expertise in disinformation research and regulatory frameworks, applauds the endeavor and provides specific recommendations for its enhancement.

Our reflections on the Code's commitments emphasize the need for alignment with international human rights laws, demonetization of disinformation, stakeholder definition, equal enforcement of terms, and support for independent media through a UN-based Fund for Information Integrity. Additionally, transparency, user empowerment, and research and data access are crucial, with encouragement for transparency reporting, maintenance of transparency repositories, and accessible data.

The recommendations also stress the importance of AI-based moderation system assessments, clear definitions for hate speech, and promoting discussions on content moderation. Drawing lessons from existing instruments like the EU's Strengthened Code of Practice on Disinformation and UNESCO's Guidelines for the Governance of Digital Platforms, the UN Code holds promise for establishing a unified global framework. GLOBSEC eagerly anticipates contributing to the Code's ongoing refinement to fortify its effectiveness in countering disinformation and promoting global information integrity.

GLOBSEC
IDEAS SHAPING THE WORLD

Vajnorská 100/B
831 04 Bratislava
Slovak Republic

+421 2 321 378 00
info@globsec.org

www.globsec.org