Online Consultation Form for the UN Code of Conduct for Information Integrity on Digital Platforms: Due 1 December

Please review Policy Brief 8: Information Integrity on Digital Platforms as background material for your consultation submission. You are advised to prepare your inputs before using this online submission form.

Online submissions that adhere to the scope of the requested input may be made public, including contact names and organizations (but not specific contact details). Please note that only completed submissions will be considered.

1. Are you submitting on behalf of yourself, a group, or an entity/organization?
Entity/Organization

2. Did you or your entity participate in a UNIC / UN Country Team consultation?
No

3. Contact First Name:
Maria Paz

4. Contact Surname:
Canales

5.Contact E-mail address:
mariapaz@gp-digital.org

6. Contact job title or area of expertise:
Head of Legal, Policy and Research

7. Name of entity/organization (if applicable):
Global Partners Digital

8. Please select the country location of your area of work.
UK

9. Website (if applicable)

https://www.gp-digital.org/

Please provide your inputs on the content of the 9 proposed principles below (as detailed in the section entitled "Towards a United Nations Code of Conduct" in Policy Brief 8: Information Integrity on Digital Platforms) and recommendations to stakeholders (be specific, where relevant). We kindly request that inputs for each principle be limited to 200 words.

10. Commitment to information integrity (200 words maximum)

"(a) All stakeholders should refrain from using, supporting or amplifying disinformation and hate speech for any purpose, including to pursue political, military or other strategic goals, incite violence, undermine democratic processes or target civilian populations, vulnerable groups, communities or individuals;"

We support the commitment to information integrity and particularly welcome the attention paid to how disinformation can undermine democratic processes and target vulnerable groups.

Our recommendations:

1. The commitment could be strengthened by including a commitment to refrain from "creating" disinformation and hate speech in addition to "using, supporting or amplifying". Adding "creating" would more fully capture the dangers of the "disinformation for hire" ecosystem, where actors may create disinformation (through automated means or otherwise) to sell to other actors to use and amplify it (https://www.cima.ned.org/blog/disinformation-for-hire-the-pollution-of-news-ecosystems-and-erosion-of-public-trust/).

2. We recommend that the commitment should include the explicit definitions of disinformation and hate speech adopted by the policy brief (page 5), either in the text of the commitment or in a footnote. As the policy brief says, there is no universally agreed definition of "disinformation", and therefore it is important to clarify within the commitment itself what definition of disinformation is being used, to avoid misuse of this commitment to restrict protected forms of speech which might include false information – such as satire, comedy and fiction – which would not fall under the appropriate definition of disinformation.

11. Respect for human rights (200 words maximum)

"(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 UNGPs;"

Our recommendations:

1. The end of Paragraph b)(i) should include measures taken by states or companies which fall below blanket shutdowns or bans (such as filtering, throttling or downranking". For example, "Ensure that responses.. are not misused to block OR IN ANY WAY CENSOR OR RESTRICT any legitimate expression of views or opinion, including through blanket Internet shutdowns, THROTTLING, BANS OR

ANY OTHER LIMITATIONS on platforms or media outlets."

2. Rather than mandating the undertaking of regulatory measures, Paragraph b)(ii) could be clearer about how regulatory measures – if taken – should align with the protection of fundamental rights. For example: (ii) *ENSURE FULL TRANSPARENCY AS TO THE REQUIREMENTS PLACED ON TECHNOLOGY COMPANIES, ASSESS ANY regulatory measures* FOR POTENTIAL IMPACTS ON FUNDAMENTAL RIGHTS, AND IMPLEMENT *enforcement mechanisms* THAT INCLUDE INDEPENDENT APPEAL AND REMEDY PROCEDURES.

3. Paragraph c) could better explain how the UNGPs should inform stakeholders' response to disinformation. For example:
"All stakeholders should comply with the UNGPs; IN PARTICULAR, POLICIES OR MEASURES TO ADDRESS DISINFORMATION MUST ENABLE BUSINESS RESPECT FOR HUMAN RIGHTS, BE ASSESSED FOR POTENTIAL IMPACTS ON HUMAN RIGHTS THROUGH APPROPRIATE DUE DILIGENCE MECHANISMS, AND BE ACCOMPANIED BY APPROPRIATE APPEAL AND REMEDY PROCEDURES."

12. Support for independent media (200 words maximum)

"(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;"

We welcome the focus on support for and funding independent media as crucial in the fight against disinformation.

Recommendations:

1. An additional paragraph should be added under this principle to address the responsibility of technology companies and online platforms to support independent media sources, including by equitable monetisation schemes for news content hosted on the platforms services and by prioritising rigorous independent journalism in users' feeds over clickbait articles designed to maximise user engagement.
2. Technology companies and member states should also support the work of independent media with civil society organizations and academia in developing effective indicators of trustworthiness of information sources.

13. Increased transparency (200 words maximum)

"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;"

We welcome the commitment to increased transparency, which will be crucial for developing more effective policies and measures on disinformation in the future; yet we note that there is no universal understanding of what "meaningful transparency" is, nor is this concept defined in the policy brief accompanying the Code of Conduct. To address this, we recommend:

1. Clarifying what is meant by "meaningful transparency", including guidelines on how to determine the scope of disclosure, degree of disclosure required for different stakeholders (including policymakers, regulators, users, researchers and the general public) and degree of granularity or technical detail.
2. Ensuring consistency of this Code of Conduct with existing initiatives on transparency from digital platforms, including by referencing the work of the Action Coalition on Meaningful Transparency and the UNESCO Guidelines for the Governance of Digital Platforms
3. Inserting an additional paragraph requiring transparency from states on specific orders, measures, enforcement actions or content takedown requests taken in relation to concern over disinformation (building on the requirement for states to be transparent about regulatory requirements on tech companies in commitment b)i)).

14. User empowerment (200 words maximum)

"(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;"

Recommendations:

1. Principle (h) should focus on Digital Platforms and address the accessibility of information. It could be amended as follows: "Member states should ensure public access to accessible, in relevant languages, accurate, transparent and

credibly-sourced DIGITAL government information, particularly information that serves the public interest, IN ORDER TO ENSURE THAT USERS CAN VERIFY INFORMATION SHARED ONLINE AGAINST OFFICIAL SOURCES AND IN ORDER TO SUPPORT FACT-CHECKING AND FACT-VERIFICATION INITIATIVES WITH TIMELY AND ACCURATE CONTENT...”

2. We would intercalate the following wording from the Government of the Netherlands’ Global Declaration on Information Integrity Online within principle (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; INCLUDING THROUGH SUPPORT FOR INTEROPERABILITY AND THE INCORPORATION OF THIRD-PARTY APPLICATIONS THAT PROVIDE USERS WITH FEATURES, FUNCTIONS AND TOOLS TO PROMOTE INFORMATION INTEGRITY AND ADDRESS MISINFORMATION AND DISINFORMATION.”
3. We suggest adding within principle (j) reference to media literacy to help citizens consume digital information in a more critical manner. We would add: *“All stakeholders should invest in robust MEDIA AND digital literacy drives to empower users of all ages* TO INTERACT WITH CONTENT CRITICALLY*, to better understand how digital platforms work…”*

15. Strengthen research and data access (200 words maximum)

“(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. Researchers should be enabled to gather examples and qualitative data on individuals and groups targeted by mis- and disinformation and hate speech to better understand the scope and nature of harms, while respecting data protection and human rights;*
> *(ii) Ensure the full participation of civil society in efforts to address mis- and disinformation and hate speech;”*

We support the commitment to research and data access, and the call for investment in and support of independent research by Member states.

1. We recommend that “quantitative” data should be included as well as “qualitative” in principle l)i), because both are required for analysing trends on mis- and disinformation and hate speech.
2. We recommend clarifying who falls under the scope of “researchers” in l)i), including those working with civil society organisations as well as academics. Principle l)i) could also include baseline requirements about the indepence of the work of said researchers and limitations on how platform data could be used, to avoid repeats of previous situations where platform data accessed for “research purposes” has been used maliciously.
3. We recommend greater clarity around how civil society should be included in efforts to address mis- and disinformation and hate speech. See our additional principle proposal (Q19) for further details.

4. We also recommend, either here or elsewhere in the principles, a recommendation for research and interventions specifically to address misinformation and disinformation targeted at women, LGBTIQ+ persons, persons with disabilities and Indigenous Peoples, acknowledging that such groups are disproportionately targeted and impacted by mis-/disinformation and hate speech online.

16. Scaled up responses (200 words maximum)

"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;"*

Our recommendations:

1. Principle m)i) is similar to previous principle k). To avoid repetition, m)i) could instead focus specifically on the topic of scaled emergency responses to crisis situations. For example: "ALL STAKEHOLDERS SHOULD INVEST IN SUITABLE MECHANISMS FOR HANDLING EMERGENCY SITUATIONS, TO ENSURE THAT THEY ARE PREPARED TO PROMPTLY DETECT AND MANAGE INCREASED RISKS AND HARMS OF MIS- AND DISINFORMATION AND HATE SPEECH DURING TIMES OF CRISIS OR CONFLICT."
2. Principle m)ii) to "form broad coalitions on information integrity" could refer more directly to the need for civil society and local expertise to be incorporated into member states' and technology companies' responses to disinformation. We have suggested dedicating a specific principle to this multistakeholder engagement (see Q19). We recommend, therefore, rewording m)ii) as "FORM TARGETED COALITIONS TO ADDRESS MIS- AND DISINFORMATION AND HATE SPEECH IN SPECIFIC TOPICAL AND GEOGRAPHICAL CONTEXTS, INCORPORATING RELEVANT LOCAL, POLITICAL, SOCIAL AND LINGUISTIC EXPERTISE TO FORM EFFECTIVE INTERVENTIONS".
3. We recommend that m)iii) could be more specific about the intended audiences of the training and capacity-building envisaged in this principle: journalists, policymakers, law enforcement officials, courts, user groups. If all, suggest including "promote training and capacity-building AMONGST ALL RELEVANT STAKEHOLDERS to develop understanding…"

17. Stronger disincentives (200 words maximum)

"(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;"

1. Commitment (n) is too vague to be meaningfully tracked or implemented by signatories.We suggest that the commitment be rephrased as follows: "DIGITAL PLATFORMS SHOULD PRIORITISE THEIR ESSENTIAL OBLIGATIONS TOWARDS HUMAN RIGHTS, PRIVACY AND SAFETY UNDER THE UNGPS ABOVE PROFIT MARGINS AND USER ENGAGEMENT. THIS MAY REQUIRE ADJUSTMENTS TO BUSINESS MODELS WHICH ARE BASED ON TARGETED ADVERTISING OR COMMITMENT OF ADDITIONAL RESOURCES TO MODERATION AND RISK ASSESSMENT. DIGITAL PLATFORMS SHOULD HAVE IN PLACE PROCESSES TO IDENTIFY AND TAKE NECESSARY ACTION WHEN THEIR SYSTEMS' ARCHITECTURE MIGHT RESULT IN THE AMPLIFICATION OF CONTENT THAT COULD BE RESTRICTED UNDER INTERNATIONAL HUMAN RIGHTS LAW."
2. Advertisers may not have granular control over where their advertisements are placed on a digital platform, and it is the responsibility of the platform to assess whether advertisements comply with their terms and conditions and to ensure that they are not placed near mis- or disinformation or hate speech. We recommend that the language of commitment (o) therefore be made specific to digital platforms. A separate commitment could be included for advertisers to ensure that the content of their adverts is not misleading and complies with platform trust and safety requirements prior to submission.

18. Enhanced trust and safety (200 words maximum)

"(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."

We support the proposed commitment for digital platforms to invest more in safety and privacy by design. Our recommendations:

1. We suggest the wording in q)ii) should reflect the need for content moderation systems to be fit-for-purpose and accompanied by well-staffed and responsive appeals processes for erroneous takedowns of content; for example,

"Invest in ROBUST, RELIABLE AND RESPONSIVE human and AI content moderation systems AND ACCOMPANYING APPEALS SYSTEMS in all languages used in countries of operation…"

2. There is a clear need for more research on the impact of artificial intelligence on mis-/disinformation and hate speech, but the wording in r) around taking "urgent and immediate measures" is vague, does not specifically focused on information integrity on digital platforms, and does not provide a concrete commitment for signatories to implement. We would recommend, therefore, amending r) as follows: "All stakeholders should ensure the safe, secure, responsible, ethical and human rights-compliant use of artificial intelligence. THIS SHOULD BE ACCOMPLISHED BY MONITORING AND RESPONDING TO: i) THE RISKS THAT ADVANCES IN ARTIFICIAL INTELLIGENCE POSE TO INFORMATION INTEGRITY ON DIGITAL PLATFORMS AND ii) THE USE OF ARTIFICIAL INTELLIGENCE FOR SPREADING MIS- AND DISINFORMATION AND HATE SPEECH ONLINE."

19. Other (proposal for a new principle not already addressed) (200 words maximum)

Multi-stakeholder and civil society engagement is highlighted in measures to address mis-/disinformation and hate speech briefly in l)i) and m)ii). However, we believe that given that this is a fundamental aspect of designing effective responses to mis- and disinformation and hate speech, there should be a standalone principle on effective multi-stakeholder engagement by Member states and platforms seeking to develop policies or interventions on disinformation. Example wording:

Principle X:

- "Member states and digital platforms should seek input from a wide range of stakeholders at every stage of drafting and implementing policies or interventions on mis- and disinformation and hate speech. It is particularly important to seek input from any groups or individuals who might be adversely affected by the proposed intervention, as well as from experts in human rights, media, education and other relevant sectors.
- Member states and digital platforms seeking multi-stakeholder input should ensure that participants are given sufficient time, insight and resources in order to meaningfully evaluate and input into proposals on mis- and disinformation and hate speech, and that there are transparent and robust mechanisms for communicating with participants how and why their input has or has not been taken on board."

20. In no more than 200 words, please provide any additional suggestions for methodologies of implementation.

The Policy brief notes that it can be difficult to distinguish between mis- and disinformation in practice (see page 5), but the commitments do not provide any guidance or methodologies that states and digital platforms can use for assessing this. Facts can also be disputed or subjected, and in some cases governments will demand that platforms remove critical or political content on grounds that it is "disinformation" when others may consider it to be true information. There may be a role for sector-specific UN bodies such as the WHO or IPCC to provide expertise, clarifications or evidence to help address specific narratives or claims reliant on health-based or climate-based mis or disinformation, as demonstrated with the Verified tool for health-based misinformation.

It will also be important to establish regular review processes for signatories to ensure that the commitments are actually being implemented in practice and reflected in national legislation; for example, by incorporating compliance with the Code of Conduct as a focus of the Universal Periodic Review. Otherwise the code of conduct risks simply contributing to the proliferation of soft guidance and voluntary codes of conduct without meaningful accountability or implementation.