Guidelines for regulating digital platforms: A multistakeholder approach to safeguarding freedom of expression and access to information

Draft 2.0[1]

Contents

[1] The original version of this document is in English. The French and Spanish versions will follow.

Introduction

1. In November 1945, UNESCO was created with the mission of "contributing to peace and security by promoting collaboration among nations through education, science and culture in order to further universal respect for justice, for the rule of law and for the human rights and fundamental freedoms which are affirmed for the peoples of the world."[2] UNESCO's global mandate, which includes the promotion of "the free flow of ideas by word and image", has guided the Organization's work for nearly 80 years—as a laboratory of ideas, a clearing house, a standard-setter, a catalyst and motor for international cooperation, and a capacity-builder. This history has shaped our mandate within the United Nations system to protect and promote freedom of expression, access to information, and safety of journalists.

2. Building upon relevant principles, conventions, and declarations over the past decade, the UNESCO Secretariat is now developing, through multistakeholder consultations and a global dialogue, *Guidelines for regulating digital platforms: a multistakeholder approach to safeguarding freedom of expression and access to information* (the Guidelines).

[2] Constitution of the United Nations Educational, Scientific and Cultural Organization, Article 1. https://www.unesco.org/en/legal-affairs/constitution#article-i---purposes-and-functions

3. This endeavour also builds upon UNESCO's work in the domain of broadcast regulation over several decades and furthers the Organization's Medium-Term Strategy for 2022–2029 (41 C/4).[3]

4. In 2015, UNESCO's General Conference endorsed the ROAM principles,[4] which highlight the importance of human rights, openness, accessibility, and multi-stakeholder participation to the development, growth, and evolution of the internet. These principles recognize the fundamental need to ensure that the online space continues to develop and be used in ways that are conducive to achieving the Sustainable Development Goals.

5. UNESCO's 41st General Conference endorsed the principles of the Windhoek+30 Declaration[5] in November 2021, following a multistakeholder process that began at the global celebration of World Press Freedom Day in May of that year. The Declaration recognized information as a public good and set three goals to guarantee that shared resource for the whole of humanity: the transparency of digital platforms, citizens empowered through media and information literacy, and media viability. In speaking about information as a public good, UNESCO recognizes that this universal entitlement is both a means and an end for the fulfilment of collective human aspirations, including the 2030 Agenda for Sustainable Development. Information empowers citizens to exercise their fundamental rights, supports gender equality, and allows for participation and trust in democratic governance and sustainable development, leaving no one behind.

6. The focus of the Guidelines on challenges related to freedom of expression and access to information complement the Organization's work in the areas of education, the sciences, and culture. This includes UNESCO's Recommendation on the Ethics of Artificial Intelligence,[6] the 2005 Convention on the Protection and Promotion of the Diversity of Cultural Expressions,[7] and the MONDIACULT Declaration of 2022.[8]

7. The current version of the Guidelines was produced through a multistakeholder consultation process that began in September 2022. Draft 2.0 will be discussed and consulted during the Internet for Trust Global Conference, to be held at UNESCO Headquarters in Paris from 21 to 23 February 2023. Subsequently,

---

[3] Strategic Objective 3 is to build inclusive, just, and peaceful societies, including by promoting freedom of expression. Strategic Objective 4 is to foster a technological environment in the service of humankind through the development and dissemination of knowledge and skills and ethical standards. https://unesdoc.unesco.org/ark:/48223/pf0000378083

[4] https://www.unesco.org/en/internet-universality-indicators

[5] https://unesdoc.unesco.org/ark:/48223/pf0000378158

[6] https://unesdoc.unesco.org/ark:/48223/pf0000380455

[7] https://en.unesco.org/creativity/convention

[8] https://www.unesco.org/sites/default/files/medias/fichiers/2022/10/6.MONDIACULT_EN_DRAFT%20FINAL%20DECLARATION_FINAL_1.pdf

a revised draft of the Guidelines will be circulated for further consultations with a view towards finalization in the months following the Conference.

The objective of the Guidelines

8. The aim of the Guidelines is to support the development and implementation of regulatory processes that guarantee freedom of expression and access to information while dealing with content that is illegal[9] and content that risks significant harm to democracy and the enjoyment of human rights.[10] They call for States to apply regulation in a manner consistent with international human rights standards and Article 19 of the International Covenant on Civil and Political Rights (ICCPR).[11]

9. The Guidelines may serve as a resource for a range of stakeholders: for policymakers in identifying objectives, principles, and processes that could be considered in policymaking; for regulatory bodies dealing with the implementation of regulation; for digital platforms in their policies and practices; and for other stakeholders, such as civil society, in their advocacy and accountability efforts.

10. The Guidelines will inform regulatory processes under development or review for digital platforms, in a manner that is consistent with international human rights standards. Such regulatory processes should be led through an **open, transparent, multistakeholder, and evidence-based** manner.

    a. The scope of these Guidelines includes digital platforms that allow users to disseminate content to the wider public, including social media networks, messaging apps, search engines, app stores, and content-sharing platforms. Bodies in the regulatory system should define which digital platform services are in scope, and also identify the platforms by their size, reach, and the services they provide, as well as features such as whether they are for-profit or non-profit, and if they are centrally managed or if they are federated or distributed platforms.

11. The Guidelines will:

    a. **Enrich and support a global multistakeholder shared space** to debate and share good practices about digital platform regulation to

[9] Any content which, in itself or in relation to an activity, is illegal in according to international human rights law and corresponding jurisprudence.
[10] Democracy as per UN Human Rights Council resolution 19/36: http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/HRC/RES/19/36&Lang=E. As stated in the Appendix, the definition of this content should be fully aligned with existing provisions in international human rights law.
[11] https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

protect freedom of expression and access to information, while dealing with content that is illegal under international human rights law and content that risks significant harm to democracy and the enjoyment of human rights, gathering different visions and a broad spectrum of perspectives.

b. **Serve as a tool for all relevant stakeholders** to advocate for human rights-respecting regulation and to hold government and digital platforms accountable.

c. **Add to existing evidence-based policy approaches** that respect human rights, ensuring alignment where possible.

12. The Guidelines will **contribute to ongoing UN-wide processes**, such as the implementation of the proposals in "Our Common Agenda," including the development of the Global Digital Compact, the preparation of the UN Summit of the Future to be held in September 2024, and the creation of a Code of Conduct that promotes integrity in public information. The Guidelines will also feed into discussions about the upcoming 20-year review of the World Summit on the Information Society (WSIS) and the Internet Governance Forum (IGF) in 2025.

Structure of the Guidelines

13. The Guidelines start by setting out the overall approach to regulation. They continue by outlining the responsibilities of different stakeholders in fostering an environment for freedom of expression, access to information, and other human rights. This includes:

    a. States' duties to respect, protect, and fulfil human rights.

    b. The responsibilities of digital platforms to respect human rights.

    c. The role of intergovernmental organizations.

    d. The role of civil society, media, academia, the technical community, and other stakeholders in the promotion of human rights.

14. Then the Guidelines propose some preconditions that should be considered in the establishment of an independent regulatory system, regarding its constitution, powers, and external review.

15. Finally, it describes the areas where digital platforms should have structures and processes in place to fulfil the objective of the regulation.

16. It is important to underscore that this document should be considered in its entirety. The adoption or implementation of specific provisions on their own will not be sufficient to achieve the regulatory goals.

Approach to regulation

17. The goal of any regulation of digital platforms that intends to deal with illegal content and content that risks significant harm to democracy and the enjoyment of human rights should include guaranteeing freedom of expression, the right to access information, and other human rights. This goal should be established in law and be drawn up after an open, transparent, multistakeholder, and evidence-based process.

18. Regulation should focus mainly on the systems and processes used by platforms, rather than expecting the regulatory system to judge the appropriateness or legality of single pieces of content. Any specific decisions about the legality of specific pieces of content should follow due process and be open to review by a judicial body, following the three-part test on legitimate restrictions to freedom of expression as laid out in the ICCPR,[12] and where relevant, the six-point threshold for defining criminal hatred that incites to discrimination, hostility, or violence outlined in the Rabat Plan of Action.[13]

19. Within regulation, digital platforms are expected to be transparent about the systems and processes used to moderate and curate content on their platforms and how those systems and processes fulfil the goal of regulation. If the established goal is not being fulfilled, the regulatory system should have the power to require the digital platform to take further action, as described in paragraph 46(f). The regulator will expect digital platforms to adhere to international human rights standards in the way they operate and to be able to demonstrate how they are implementing these standards and other policies contained in their terms of service.

20. Alongside the regulation of digital platforms, it is essential that key media and information literacy skills for users are promoted, including by the platforms themselves. This enables users to engage critically with content and technologies, navigate a rapidly evolving media and information landscape marked by the digital transformation, and build resilience in the face of related challenges.

21. The current approach taken by these Guidelines is one of co-regulation, implying that the State, on the one hand, provides a legal framework that

[12] See the UNESCO explanatory video, “The Legitimate Limits to Freedom of Expression: the Three-Part Test,” at https://www.youtube.com/watch?v=Wg8fVtHPDag.
[13] See the UNESCO explanatory video, “The Rabat Plan of Action on the Prohibition of Incitement to Hatred,” at https://www.youtube.com/watch?v=ADrB32OSe3A&t=8s.

enables the creation, operationalization, and enforcement of rules, and self-governing bodies, on the other hand, create rules and administer them, sometimes through joint structures or mechanisms. This should be done in accordance with international human rights law and under the public scrutiny of civil society organizations, journalists, researchers, and other relevant institutions in a system of checks and balances.

Enabling environment

22. To accomplish the goal of regulation, all stakeholders involved have a role in sustaining an enabling environment for freedom of expression and the right to information, while dealing with content that risks significant harm to democracy and the enjoyment of human rights.

23. Creating a safe and secure internet environment for users while protecting freedom of expression and access to information is not simply an engineering question. It is also a responsibility for societies as a whole and therefore requires whole-of-society solutions.

States' duties to respect, protect, and fulfil human rights

24. States have a particular duty to promote and guarantee freedom of expression and the right to access information, and to refrain from censoring legitimate content.

25. A key element of an enabling environment is the positive obligation to promote universal and meaningful access to the internet. In 2011, in the Joint Declaration on Freedom of Expression and the Internet, the special international mandates on freedom of expression indicated: “Giving effect to the right to freedom of expression imposes an obligation on States to promote universal access to the Internet.”[14]

26. Moreover, it is a responsibility of the State to be transparent and accountable about the requirements they place upon digital platforms.

27. Specifically, States should:

    a. Respect the requirements of Article 19(3) of the ICCPR: any restrictions applied to content should have a basis in law, have a legitimate aim, and be necessary and proportional, ensuring that users’ rights to freedom of expression, access to information, equality and non-discrimination, autonomy, dignity, reputation, privacy, association, and public participation are protected.

[14] Adopted 1 June 2011, para. 6(a), http://www.law-democracy.org/wp-content/uploads/2010/07/11.06.Joint-Declaration.Internet.pdf.

b. Provide an effective remedy for breaches of these rights.

c. Ensure that any restrictions imposed upon platforms consistently follow the high threshold set for defining legitimate restrictions on freedom of expression, on the basis of the application of Articles 19 and 20 of the ICCPR.

d. Be open, clear, and specific about the type, number, and legal basis of requests they make to digital platforms to take down, remove, and block content. States should be able to demonstrate how this is consistent with Article 19 of the ICCPR*.*

e. Refrain from disproportionate measures, particularly prior censorship and internet shutdowns, under the guise of combatting disinformation or any other reason inconsistent with the ICCPR.

f. Refrain from imposing a general monitoring obligation or a general obligation for digital platforms to take proactive measures to relation to illegal content. Digital platforms should not be held liable when they act in good faith and with due diligence, carry out voluntary investigations, or take other measures aimed at detecting, identifying, and removing or disabling access to illegal content.

g. Refrain from subjecting staff of digital platforms to criminal penalties for an alleged or potential breach of regulations in relation to their work on content moderation and curation, as this may have a chilling effect on freedom of expression.

h. Promote media and information literacy, including in digital spaces, as a complementary approach to regulation with the aim of empowering users. This should draw upon the expertise of media and information literacy experts, academics, civil society organizations, and access to information institutions.

i. Ensure that the regulatory system with responsibilities in this area is structured as independent and has external review systems in place (see paragraphs 47–49) such as legislative scrutiny, requirements to be transparent and consult with multiple stakeholders, and the production of annual reports and regular audits.

The responsibilities of digital platforms to respect human rights

28. Digital platforms should comply with five key principles:

    a. **Platforms respect human rights in content moderation and curation**. They have content moderation and curation policies and practices consistent with human rights standards, implemented algorithmically and through human means, with adequate protection and support for human moderators.

    b. **Platforms are transparent**, being open about how they operate, with understandable and auditable policies. This includes transparency about the tools, systems, and processes used to moderate and curate content on their platforms, including in regard to automated processes.

    c. **Platforms empower users** to understand and make informed decisions about the digital services they use, including helping them to assess the information on the platform.

    d. **Platforms are accountable to relevant stakeholders**, to users, the public, and the regulatory system in implementing their terms of service and content policies, including giving users rights of redress against content-related decisions.

    e. **Platforms conduct human rights due diligence**, evaluating the risks and impact on human rights of their policies and practices.

29. To follow these principles, there are specific areas on which digital platforms have a responsibility to report to or act before the regulatory system. These areas are described in paragraphs 50–105.

The role of intergovernmental organizations

30. Intergovernmental organizations, in line with their respective mandates, should support relevant stakeholders in guaranteeing that the implementation of these guidelines is in full compliance with international human rights law, including by providing technical assistance, monitoring and reporting human rights violations, developing relevant standards, and facilitating multistakeholder dialogue.

The role of civil society and other stakeholders

31. Every stakeholder engaged with the services of a digital platform as a user, policymaker, watchdog, or by any other means, has an important role to play in supporting freedom of expression, access to information, and other human

rights. Toward this end, the process of developing, implementing, and evaluating every regulation should take a multistakeholder approach; a broad set of stakeholders should also be engaged in oversight.

32. Civil society plays a critical role in understanding the nature of and countering abusive behaviour online, as well as challenging regulation that unduly restricts freedom of expression, access to information, and other human rights.

33. Researchers have a role in identifying patterns of abusive behaviour and where the possible root causes could be addressed; researchers should also be able to provide independent oversight of how the regulatory system is working. Independent institutions and researchers can support risk assessments, audits, investigations, and other types of reports on platforms' practices and activities.

34. Media and fact-checking organizations have a role in promoting information as a public good and dealing with content that risks significant harm to democracy and the enjoyment of human rights on their own platforms.

35. Engineers, data scientists, and all the technical community involved also have a role in understanding the human rights and ethical impacts of the products and services they are developing.

36. All of these stakeholders should have an active role in consultations on the operation of the regulatory system.

The regulatory system

37. There are vastly different types of bodies involved in online regulation throughout the world. They range from existing broadcast and media regulators who may be asked to take on the role of regulating content online, to newly established dedicated internet content regulators or communications regulators given an extended remit. There may also be overlap in some states with advertising or election bodies, or with information commissioners or national human rights institutions. Some regulators exist independently of the government while others are constituted as government agencies.[15] Recognising the complexity of this environment, these Guidelines are meant to be generally applicable to any system of regulation, irrespective of its specific modalities, and accept that local contexts will impact how regulation is enacted and implemented.

[15] It is important to bear in mind how the regulation of online content interacts with and informs other institutions with jurisdiction over issues as diverse as the protection of personal data, consumer protection, access to public information, electoral regulation, telecommunications regulation, antitrust and market regulation authorities, and the protection of human rights. The roles of the legislature and judicial authorities also need to be considered in the structure of the regulatory system.

38. Whatever form it takes, any process that establishes a regulatory system for digital platforms should be open and transparent and include multistakeholder consultation. Additionally, achieving the goal of regulation requires the existence of an independent regulatory system that allows regular multistakeholder consultation on its operation.

39. The World Bank stated that the key characteristic of the independent regulator model is decision-making independence.[16] A guiding document on broadcast regulation commissioned by UNESCO (2006) also highlighted that "an independent authority (that is, one which has its powers and responsibilities set out in an instrument of public law and is empowered to manage its own resources, and whose members are appointed in an independent manner and protected by law against unwarranted dismissal) is better placed to act impartially in the public interest and to avoid undue influence from political or industry interests."[17]

40. The proposal below is divided into three sections: the constitution of an independent regulatory system, its powers, and suggested provisions for review.

Constitution

41. Any regulatory system—whether comprised of a single body or multiple overlapping bodies—which **assesses applications or performs inspectorial, investigative, or other compliance functions** over how digital platforms conduct content moderation and curation, needs to be independent and free from economic or political pressures.

42. The regulatory system must have sufficient funding to carry out its responsibilities effectively. The sources of funding must also be clear, transparent, and accessible to all and not subject to the decisions of the regulator(s).

43. Officials or members of the regulatory system should:

    a. Be appointed through a participatory and independent merit-based process.

[16] This means that the regulator's decisions are made without the prior approval of any other government entity, and no entity other than a court or a pre-established appellate panel can overrule the regulator's decisions. The institutional building blocks for decision-making independence are organizational independence (organizationally separate from existing ministries and departments), financial independence (an earmarked, secure, and adequate source of funding), and management independence (autonomy over internal administration and protection from dismissal without due cause). See *Handbook for Evaluating Infrastructure Regulatory Systems,* p.50 http://elibrary.worldbank.org/doi/book/10.1596/978-0-8213-6579-3

[17] https://unesdoc.unesco.org/ark:/48223/pf0000144292

b. Be accountable to an independent body (which could be the legislature, an external council, or an independent board/boards).

c. Have relevant expertise in international human rights law.

d. Deliver a regular public report to an independent body (ideally the legislature) and be held accountable to it, including by informing the body about their reasoned opinion.

e. Make public any possible conflict of interest and declare any gifts or incentives.

f. After completing the mandate, not be hired or provide paid services to those who have been subject to their regulation, and this for a reasonable period, in order to avoid the risk known as “revolving doors.”

Powers

44. The regulatory system should primarily focus on the systems and processes used by digital platforms to moderate and curate content, rather than making judgements about individual pieces of content. The system should also look at how digital platforms promote freedom of expression and access to information and the measures it has established to deal with illegal content and content that risks significant harm to democracy and the enjoyment of human rights.

45. The regulatory system should have the power to assess applications or perform inspectorial, investigative, or other compliance functions over digital platforms to fulfil the overarching goals to protect freedom of expression and access to information, while moderating illegal content and content that risks significant harm to democracy and the enjoyment of human rights**,** in a way consistent with the provisions of Article 19 of the ICCPR.

46. To fulfil the goal of regulation, the regulatory system should have the following powers:

    a. Establish standardized reporting mechanisms and formats. Ideally, reports should be made annually in a machine-readable format.

    b. Commission off-cycle reports if there are exigent emergencies, such as a sudden information crisis (such as that brought about by the COVID-19 pandemic) or a specific event which creates vulnerabilities (for example, elections or protests).

    c. Summon any digital platform deemed non-compliant with its own policies or failing to protect users. Any decision by the regulator should be

evidence-based, the platform should have an opportunity to make representations and/or appeal against a decision of non-compliance, and the regulatory system should be required to publish and consult on enforcement guidelines and follow due process before directing a platform to implement specific measures.

d. Commission a special investigation or review by an independent third party if there are serious concerns about the operation or approach of any platform or an emerging technology when dealing with illegal content or content that risks significant harm to democracy and the enjoyment of human rights.

e. Establish a complaints process that offers users redress should a platform not deal with their complaint fairly, based on the needs of the public they serve, the enforcement powers they have in law, their resources, and their local legal context.

f. Oversee the fulfilment by the digital platforms of the five principles detailed in these guidelines, taking necessary and proportional enforcement measures, in line with international human rights law, when platforms consistently fail to implement these principles.

Review of the regulatory system

47. There should be a provision for a periodic independent review of the regulatory system, conducted by a respected third party, reporting directly to the legislature.

48. Any part of the regulatory system should act only within the law in respect of these powers, respecting fundamental human rights—including the rights to privacy and to freedom of expression. It should be subject to review in the courts if it is believed to have exceeded its powers or acted in a biased or disproportionate manner.

49. Decisions on eventual limitations of specific types of content must be allowed to be reviewed by an independent judicial system, following a due process of law.

Responsibilities of digital platforms

50. Digital platforms should respect human rights and adhere to international human rights standards in accordance with the UN Guiding Principles on Business and Human Rights.[18]

[18] https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

51. According to the five principles set above, digital platforms are expected to have structures and processes in place and should be accountable to the regulatory systems, in line with the powers described above, in the following areas:

Principle 1. Platforms respect human rights in content moderation and curation

Content moderation and curation policies and practices[19]

52. Digital platforms should ensure that human rights and due process considerations are integrated into all stages of the content moderation and curation policies and practices.

53. The content moderation and curation policies of digital platforms should be consistent with the obligations of corporations to respect human rights, as set out in the UN Guiding Principles on Business and Human Rights and other established international human rights standards.

54. Content moderation and curation structures and processes should be applied consistently and fairly across all regions and languages.

55. No distinction should be made between content that is similar or between users. However, content moderation decisions should, in a transparent manner, take into account the context, the wide variation of language nuances, and the meaning and linguistic and cultural particularities of the content.

56. Digital platforms should—in policy and practice—ensure whenever they become aware of the availability of illegal content that they act with due diligence and in accordance with international human rights standards. At a minimum, they should ensure that there is quick and decisive action to remove known child sexual abuse materials or other explicit and severe illegal content which is not contextually dependent.

57. It would be expected that illegal content be made unavailable solely in the geographical jurisdiction where it is illegal.[20] Identification of illegal content should be interpreted consistently with international human rights law to avoid unjustified restrictions on freedom of expression.

[19] Given the importance and complexity of this issue, UNESCO particularly welcomes further contributions on how the spread of content that risks significant harm to democracy and the enjoyment of human rights can best be addressed through automated means, while preserving freedom of expression and access to information.

[20] However, it is important to recognise that no systems and processes will be 100% precise in identifying illegal content (at least not without disproportionate intrusion and monitoring). Therefore, it should not automatically be a breach of the regulations if illegal content is found on the service, unless it can be shown that the platform knew of it and failed to report it, or if the relevant systems and processes can be shown to be inadequate.

58. Platforms should be able to demonstrate to the regulatory system about the measures they carry out to detect, identify, or remove Illegal content.

59. In the case of other content that risks significant harm to democracy and the enjoyment of human rights, digital platforms should systematically assess the potential human rights impact of such content and take action to reduce vulnerabilities and increase their capacities to deal with it. For instance, companies should be able to demonstrate to the regulatory system the measures that they have in place if such risk is identified. These could be by, for example, providing alternative reliable information,[21] indicating concerns about the origin of the content to users, limiting or eliminating the algorithmic amplification of such content, or de-monetizing from advertising revenue.

Human content moderation

60. Human content moderators should be adequately trained, sufficiently staffed, fluent in the language concerned, vetted, and psychologically supported. Platforms should further put in place well-funded and -staffed support programmes for content moderators to minimize harm caused to them through their reoccurring exposure to violent or disturbing content while at work. Where possible and when it would not negatively impact human rights or undermine adherence to international norms for freedom of expression, human moderation of content should take place in the country or region where it is published to ensure close awareness of local or national events and contexts, as well as fluency in the language concerned.

61. The platform should also be explicit about whether it partners with outside organizations or experts to help it make decisions, particularly in countries or regions where the platform itself has little local knowledge. In so doing, they should always follow the “do no harm principle” and refrain from revealing partners in situations in which revealing these partners may present risks for their safety.

Use of automated systems for content moderation and curation

62. Digital platforms should commission regular external audits of machine learning tools utilised for content moderation for their precision, accuracy, and for possible bias or discrimination across different content types, languages, and contexts. They should also commission regular independent assessments of the impacts of automated content moderation tools on human rights. The results of these reviews should be reported to the regulatory system.

[21] For instance, several digital platforms have instituted “disputed news” tags that warn readers and viewers about contentious content.

63. Digital platforms should commission regular external audits of machine learning tools utilised for automated curation and recommender mechanisms – designed to enhance user engagement – for their precision, accuracy, and for possible bias or discrimination across different content types, languages, and contexts. They should also commission regular independent assessments of the impacts of these mechanisms on human rights. The results of these reviews should be reported to the regulatory system.

64. Digital platforms should have in place systems and processes to identify and take necessary action, in line with the provisions of these guidelines, when automated curation and recommender mechanisms – designed to enhance user engagement – result in the amplification of content that risks significant harm to democracy and human rights.

65. Users should be given the ability to control the algorithmic curation and recommender mechanisms used to suggest content to them. Content curation and recommendation systems that provide different sources and include different viewpoints around trending topics should be made clearly available to users.

66. Finally, digital platforms should notify users when their content is removed or subject to content moderation and why. This would allow users to understand why that action on their content was taken, the method used (algorithmic or after human review), and under which platform rules action was taken. Digital platforms should also have processes in place that permit users to appeal such decisions (see paragraphs 89-91).

Principle 2. Platforms are transparent

67. Digital platforms should report to the regulatory system on how they fulfil the principles of transparency, explicability, and reporting against what they say they do in their terms of services and community standards.[22] The meaning of transparency depends upon the audience. For users, it can mean, for example, understanding how the platform finds and presents information and collects their data; for regulators, it can mean information needed to verify the way in which digital platforms' business operations may impact democracy and human rights, and if terms of service and community standards are consistently and fairly applied; and for researchers, it can mean understanding the impact of the services on society in general.

68. The regulatory system and digital platforms should understand transparency as *meaningful* transparency. Transparency is not simply the provision of legal texts

[22] Guidance on transparency for digital platforms can be found in the 26 high-level principles set forth by UNESCO in *Letting the Sun Shine In: Transparency and Accountability in the Digital Age*. https://unesdoc.unesco.org/ark:/48223/pf0000377231

or a data dump—it should be understood as providing stakeholders with the information they need to make informed decisions.

Meaningful transparency

69. The effectiveness of digital platforms' transparency mechanisms should be independently evaluated through qualitative and empirical quantitative assessments to determine whether the information provided for meaningful transparency has served its purpose. Reports should be made available to users on a regular basis.

70. Digital platforms should publish information outlining how they ensure that human rights and due process considerations are integrated into all stages of the content moderation and curation policies and practices. This publicly available information should include:

Transparency in relation to terms of service

a. Any measures used to moderate and curate content, set out in platforms' terms of services.

b. Any information about processes used to enforce their terms of service and to sanction users, as well as government demands/requests for content removal, restriction, or promotion.

c. Information about the reasons behind any restrictions imposed in relation to the use of their service, publicly available in an easily accessible format in their terms of service.

Transparency in relation to content moderation and curation policies and practices

d. How content is moderated and curated, including through algorithmic (automated) means and human review, as well as content that is being removed or blocked under either terms of service or pursuant to government demands/requests.

e. Any change in content moderation and curation policies should be communicated to users periodically.

f. Any use made of automated means for the purpose of content moderation and curation, including a specification of the role of the automated means in the review process, and any indicators of the benefits and limitations of the automated means in fulfilling those purposes.

g. Any safeguards applied in relation to any content moderation and curation that are put in place to protect freedom of expression and the

right to information, including in response to government requests, particularly in relation to matters of public interest, including journalistic content.

h. Information about the number of human moderators employed and the nature of their expertise in local language and local context, as well as whether they are in-house staff or contractors.

i. How personal data is used and what treatment is made of users' personal data, including personal and sensitive data, to make algorithmic decisions for purposes of content moderation and curation.

Transparency in relation to user complaints mechanisms

j. Information relevant to complaints about the removal, blocking, or refusal to block content and how users can access the complaints process.

Transparency and commercial dimensions

k. Information about political advertisements, including the author and those paying for the ads; these advertisements should be retained in a publicly accessible library online.

l. Practices of advertising and data collection.

m. Information which allows individuals to understand the basis on which they are being targeted for advertising.

71. Many regulatory regimes require broader and more granular transparency standards than those outlined here. The standards presented in these Guidelines can be considered as a baseline from which regulatory regimes can elaborate further.

Data access for research purposes

72. Digital platforms should provide access to non-personal data and anonymised data for vetted researchers that is necessary for them to undertake research on content to understand the impact of digital platforms. This data should be made available through automated means, such as application programming interfaces (APIs), or other open and accessible technical solutions allowing the analysis of said data.

73. They should provide access to data to undertake research on illegal and harmful content such as hate speech, disinformation, misinformation, and content which incites or portrays gender-based violence; such data should be

disaggregated for the purpose of investigating impacts on specific populations. There need to be additional safeguards to protect the privacy and personal data of users, as well as businesses' proprietary information, trade secrets, and respect of commercial confidentiality.

74. Platforms should build reliable interfaces for data access. The independent regulatory system should determine what is useful, proportionate, and reasonable for research purposes.

Principle 3. Platforms empower users

User reporting

75. It is critical to empower users of digital platforms. In addition to the digital platform making information about its policies accessible in a digestible format and in all relevant languages, it should demonstrate how users can report potential abuses of the policies, whether that be the unnecessary removal of content or the presence of allegedly illegal content or content that risks significant harm to democracy and the enjoyment of human rights, or of any other content which is in breach of its policies. Digital platforms should also have the means to understand local context and local conditions when responding to user complaints and ensure that their systems are designed in a culturally sensitive way.

76. The user reporting system should give high priority to concerns regarding content that threatens users, ensuring a rapid response, and, if necessary, by providing a specific escalation channel or means of filing the report. This is particularly important when it to comes to gender-based violence and harassment.

Media and information literacy

77. When reporting to the regulatory system, platforms should demonstrate their overall strategy related to media and information literacy and the actions they have taken to advance on it. There should be a specific focus inside the digital platform on how to improve the digital literacy of its users, with thought given to this in all product development teams. The digital platform should consider how any product or service impacts user behaviour beyond the aim of user acquisition or engagement.

78. Platforms should train their product development teams on media and information literacy from a user empowerment perspective, based on international standards, and put in place both internal and independent monitoring and evaluation mechanisms. They should inform the regulatory system about any relevant result of these evaluations.

79. Digital platforms should implement such measures in close collaboration with organizations and experts independent of the platforms, such as public authorities responsible for media and information literacy, academia, civil society organizations, researchers, teachers, specialized educators, youth organizations, and children's rights organizations. Specific measures should be taken for users and audiences in social or cultural vulnerability and/or with specific needs.

80. Digital platforms should be explicit about the resources they make available to improve media and information literacy, including digital literacy about the platform's own products and services, as well as relevant processes, for their users.

81. Digital platforms should also ensure that users understand their rights online and offline, including the role of media and information literacy in the enjoyment of the rights to freedom of expression and access to information. Toward this end, they could partner with independent media and information literacy experts or organizations that have relevant expertise in the thematic area, including academic and civil society organizations.

Language and accessibility

82. Major platforms should have their full terms of service available in the primary languages of every country where they operate, ensure that they are able to respond to users in their own language and process their complaints equally, and have the capacity to moderate and curate content in the user's language. Automated language translators, while they have their limitations, can be deployed to provide greater language accessibility.

83. Platforms should also ensure that content that risks significant harm for democracy and human rights is not amplified by automated curation or recommender mechanisms simply due to a lack of linguistic capacity of those mechanisms.

84. The rights of persons with disabilities should always be taken into account, with particular attention to the ways in which they can interact with and make complaints in relation to the platform.

Children's rights

85. Children have a special status given their unique stage of development, limited or lack of political voice, and the fact that negative experiences in childhood can

result in lifelong or transgenerational consequences.[23] Digital platforms should therefore also recognise their specific responsibilities toward children.

86. Where digital platforms allow use of their services by children, they should provide all children with equal and effective access to age-appropriate information, including information about their rights to freedom of expression, access to information, and other human rights. Terms of services and community standards should be made available in age-appropriate language for children and, as appropriate, be co-created with a diverse group of children; special attention should be paid to the needs of children with disabilities to ensure they enjoy equal levels of transparency as set out in the previous section.

Principle 4. Platforms are accountable to relevant stakeholders

87. Digital platforms should be able to demonstrate that any action taken when moderating and curating content has been conducted in accordance with their terms of services and community standards and should report fairly and accurately to the regulatory system on performance vis-à-vis their responsibilities and/or plans. In case of failure to comply with this provision, the regulatory system should act in accordance with the powers outlined in these Guidelines.

Use of automated tools

88. Digital platforms should be able to explain to the regulatory system about the use and impact of the automated systems, including the extent to which such tools affect the data collection, targeted advertising, and the disclosure, classification, and/or removal of content, including election-related content. In case of failure to comply with this provision, the regulatory system should act in accordance with the powers outlined in these Guidelines (see paragraph 46(f)).

User appeal and redress

89. There should be an effective user complaints mechanism to allow users (and non-users if impacted by specific content) meaningful opportunities to raise their concerns. This should include a clear, easily accessible, and understandable reporting channel for complaints, and users should be notified about the result of their appeal.

[23] See United Nations Committee on the Rights of the Child (2013), "General comment No. 16 (2013) on State obligations regarding the impact of the business sector on children's rights," para. 4. See also General comment No. 25 (2021) on children's rights in relation to the digital environment.

90. The appeals mechanism should follow the seven principles outlined in the UN Guiding Principles on Business and Human Rights for effective complaints mechanisms: legitimacy, accessibility, predictability, equitability, transparency, rights-compatibility, and continuous learning.

91. Digital platforms should notify users and explain processes for appeal when their content is removed or expressly labelled, restricted in terms of comments or re-sharing or advertising association, given special limits in terms of amplification or recommendation (as distinct from "organic/algorithmic" amplification and recommendation), and why. This would allow users to understand the reasons that action on their content was taken, the method used (algorithmic or after human review), and under which platform rules action was taken. Also, they should have processes in place that permit users to appeal such decisions.

Principle 5. Platforms conduct human rights due diligence

Human rights safeguards and risk assessments

92. Digital platforms should be able to demonstrate to the regulatory system the systems or processes they have established to ensure user safety while also respecting freedom of expression, access to information, and other human rights.

93. Platforms should conduct periodic risk assessments to identify and address any actual or potential harm or human rights impact of their operations, based on the provisions of Article 19 of the ICCPR and drawing on the principles set out in the UN Guiding Principles on Business and Human Rights.

94. Apart from periodic assessments, risk assessments should also be undertaken:

    a. Prior to any significant design changes, major decisions, changes in operations, or new activity or relationships;

    b. To protect the exercise of speech by minority users and for the protection of journalists and human rights defenders;[24]

    c. To help protect the integrity of electoral processes;[25]

    d. In response to emergencies, crises, or conflict or significant change in the operating environment.[26]

[24] See paragraphs 85-86 and 97-98 Gender disinformation and online gender-based violence).
[25] See paragraphs 99-103 on election integrity.
[26] See paragraphs 104-105 on emergencies, crisis, or conflict.

95. Digital platforms should be open to expert and independent input on how these assessments are structured.

96. Platforms can create spaces to listen, engage, and involve victims, their representatives, and users from minorities to identify and counter illegal content and content that risks significant harm to democracy and the enjoyment of human rights, to identify opportunities and systemic risks in order to then promote solutions and improve their policies. Consideration should be given to the creation of specific products that enable all relevant groups to actively participate in the strengthening of counter-narratives against hate speech.

Specific measures to fight gendered disinformation and online gender-based violence

97. There is considerable evidence that women in public life—including politicians, journalists, and public figures—are targeted by disinformation, fake stories, sexual harassment and threats, and incitement to violence. While some of these instances may be the result of individuals, others are the result of deliberate campaigns designed to undermine women's participation in civil and political life, to undermine their trustworthiness, or simply drive them off the digital platform and deny their right to freedom of expression. This phenomenon is even more marked for women from racial or other minority groups. Such disinformation can all too often lead to gender-based violence. This represents a significant erosion of women's human rights.

98. To fight gendered disinformation and online gender-based violence, digital platforms should:

    a. Conduct annual human rights and gender impact assessments, including algorithmic approaches to gender-specific risk assessment, with a view to identify the systemic risks to women and girls and to adjust regulations and practises to mitigate such risks more effectively.

    b. Use privacy-enhancing technology to provide external researchers access to internal data of platforms to help identify algorithmic amplification of gendered disinformation, gender-based harassment, hate speech, and toxic speech.

    c. Create dedicated engineering teams that are made up of both men and women who are specifically trained to develop algorithmic solutions to different forms of gendered disinformation, including violent and other forms of toxic speech and harmful, stereotypical content.

    d. Develop and launch inclusive structured community feedback mechanisms to eliminate gender bias in generative AI and generative

algorithms producing content that perpetuates or creates gendered disinformation or harmful or stereotypical content.

Specific measures for the integrity of elections

99. While electoral bodies and administrators need to ensure that the integrity of the electoral process is not affected or undermined by disinformation and other harmful practices, digital platforms should have a specific risk assessment process for any election event. Such risk assessments should also consider the users, the level of influence that advertisement messages may have on them, and the potential harm that may come out of such messages if used against specific groups, such as minorities or other vulnerable groups.

100. Within the assessment, digital platforms should review whether political advertising products, policies, or practices arbitrarily limit access to information for citizens, voters, or the media, or the ability of candidates or parties to deliver their messages.

101. Digital platforms should also engage with the election's administrator/regulator (and relevant civil society groups), if one exists, prior to and during an election to establish a means of communication if concerns are raised by the administrator or by users/voters. Engagement with other relevant independent regulators may be necessary according to the particular circumstances of each jurisdiction.

102. Digital platforms that accept political advertising should clearly distinguish such content as advertisements and should ensure in their terms of service that to accept the advertisement, the funding and the political entity are identified by those that place them.

103. The platform should retain these advertisements and all the relevant information on funding in a publicly accessible library online.

Specific measures in emergencies, conflict, and crisis

104. As a human rights safeguard, digital platforms should have risk assessment and mitigation policies in place for emergencies, crises, and conflict, and other sudden world events where content that risks significant harm to democracy and the enjoyment of human rights is likely to increase and where its impact is likely to be rapid and severe. In the case of emerging conflicts, digital platforms should be alert to this type of content, which has in many instances fuelled or even driven conflict. Measures such as fact-checking content related to the crisis should be considered.

105. Risk assessments may require digital platforms to have processes in place for cases in which a large number of simultaneous requests for action by users are made, as sometimes happens in the context of social unrest or massive violations of human rights.

Conclusion

106. Digital platforms have empowered societies with enormous opportunities for people to communicate, engage, and learn. They offer great potential for communities in social or cultural vulnerability and/or with specific needs, democratizing spaces for communication and opportunities to have diverse voices engage with one another, be heard, and be seen. But due to the fact that key risks were not taken into account earlier, this potential has been gradually eroded over recent decades.

107. The goal of these Guidelines is to support the development and implementation of regulatory processes that guarantee freedom of expression and access to information while dealing with illegal content and content that risks significant harm to democracy and the enjoyment of human rights. They aim to enrich and support a global multistakeholder shared space to debate and share good practices about digital platform regulation; serve as a tool for all relevant stakeholders to advocate for human rights-respecting regulation and to hold government and digital platforms accountable; add to existing evidence-based policy approaches that respect human rights, ensuring alignment where possible; and contribute to ongoing UN-wide processes.

108. The Guidelines were produced through a multistakeholder consultation process that began in September 2022. The present draft Guidelines will be the basis for the dialogue taking place during the Internet for Trust Global Conference.

109. Consultations will continue in the following months to seek a wide diversity of voices and positions to be heard around this complex issue that requires immediate action to protect freedom of expression, access to information, and all other human rights in the digital environment.

Appendix

Resources

United Nations

The Rabat plan of action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence
https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action

United Nations Guiding Principles on Business and Human Rights
https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

United Nations Secretary General report - Countering disinformation for the promotion and protection of human rights and fundamental freedoms
https://www.ohchr.org/sites/default/files/2022-03/NV-disinformation.pdf

UN Special Rapporteur on freedom opinion and expression - A human rights approach to online content moderation
https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/Factsheet_2.pdf

UNESCO

Letting the sun shine in: transparency and accountability in the digital age
https://unesdoc.unesco.org/ark:/48223/pf0000377231

"The Legitimate Limits to Freedom of Expression: the Three-Part Test" - UNESCO [video]
https://www.youtube.com/watch?v=Wg8fVtHPDag

References on terminology

Regarding illegal content
Any content which, in itself or in relation to an activity, is illegal in line with international human rights law and corresponding jurisprudence.

Regarding content that risks significant harm to democracy and the enjoyment of human rights
For the purposes of these Guidelines, this term refers to different types of content that have been broadly discussed by the UN System, as follows:

Hate speech
United Nations Strategy and Plan of Action on Hate Speech
https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf

Disinformation and misinformation

Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/Factsheet_2.pdf

Content which incites or portrays gender-based violence

Special Rapporteur on violence against women, its causes and consequences
https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/HRC/38/47&Lang=E
Statement by Irene Khan, Special Rapporteur on the promotion and protection of freedom of opinion and expression
https://www.ohchr.org/en/statements/2022/02/statement-irene-khan-special-rapporteur-promotion-and-protection-freedom-opinion

On digital platforms

For the purposes of these Guidelines, the relevant digital platforms are those that allow users to disseminate content to the wider public. Such platforms include social media networks, search engines, app stores, and content-sharing platforms.

On regulation

For the purposes of these Guidelines, regulation is a process where a set of rules for private actors is set out in law, usually overseen by a body, usually a public agency, that is established to monitor and enforce compliance with these rules. Regulation can be understood as being based upon “hard” law, where statutory requirements are made of private actors. This is distinct from “soft law,” which takes the form of guidelines, recommendations, or codes of practice which are not legally binding, but which may be followed by private actors, and which may have a moral force.

Regulatory system

The regulatory system is the group of institutions designated for supervising and monitoring digital platforms. A system for supervision and monitoring of an actor or industry, potentially composed of multiple bodies.

Regulator

A body that supervises, monitors, and holds to account a private actor.

Independent regulator

An independent regulator has its powers and responsibilities set out in an instrument of public law and is empowered to manage its own resources, and whose members are appointed in an independent manner and protected by law against unwarranted dismissal. In this case, the regulator’s decisions are made without the prior approval of any other government entity, and no entity other than a court or a pre-established appellate panel can overrule the regulator’s decisions. The institutional building blocks for decision-making independence are organizational independence (organizationally separate from existing ministries and departments), financial independence (an earmarked, secure, and adequate source of funding), and management independence (autonomy over internal administration and protection from dismissal without due cause).

Sources:
UNESCO. Guidelines for Broadcasting Regulation.
https://unesdoc.unesco.org/ark:/48223/pf0000144292

World Bank. World Bank Handbook for Evaluating Infrastructure Regulatory Systems.
http://elibrary.worldbank.org/doi/book/10.1596/978-0-8213-6579-3

Co-regulation
The term “co-regulation” covers a wide range of different regulatory approaches that involve cooperation between State regulation and self-regulation. Co-regulation implies that State, on the one hand, provides a legal framework that enables the creation, operationalization, and enforcement of rules; self-governing bodies, on the other hand, create rules and administering them, sometimes through joint structures or mechanisms.

Source: UNESCO. Privacy, free expression and transparency: redefining their new boundaries in the digital age
https://unesdoc.unesco.org/ark:/48223/pf0000246610.locale=en

Self-regulation and codes of practice
Self-regulation refers to situations when a non-State group engages in a rule-making process, by developing a set of rules, such as codes of conduct, a process of enforcement of the rules, or a comprehensive regulatory system altogether. It is supposed to replace the procedural, substantive, and implementation functions that might otherwise be included in State legislation/regulation.

Source: UNESCO. Privacy, free expression and transparency: redefining their new boundaries in the digital age
https://unesdoc.unesco.org/ark:/48223/pf0000246610.locale=en