Harmful Content Working Group: Progress Report

A multistakeholder effort to unpack the idea of "harmful" content

WHITE PAPER | November 2022

PARIS PEACE FORUM de PARIS sur la PAIX

Harmful Content Working Group: Progress Report

A multistakeholder effort to unpack the idea of "harmful" content

---

WHITE PAPER | November 2022

About the prospective Harmful Content Working Group

Launched in January 2022, this prospective working group has aimed to unpack the question of “harmful” content and explore possible core principles for a comprehensive qualification of the last, should it be differentiated from illegal content.

It draws on the advantages of the Paris Peace Forum’s multistakeholder platform and community, notably its unique convening power and experience in fostering multi-actor consensus on core global governance issues.

Regular meetings were held under the Chatham House rules and enabled free and agile exchanges between representatives from the public sector, the civil society and from the industry across the world to identify key challenges and discuss the opportunity of common core principles.

About the Paris Peace Forum

In a world requiring more collective action, the Paris Peace Forum is a platform open to all seeking to develop coordination, rules, and capacities that answer global problems. Year-round support activities and an annual event in November help better organize our planet by convening the world, boosting projects, and incubating initiatives.

Composition of the Working Group

This Working Group gathered experts across the world with a sound knowledge of the issues related to the regulation of online content. While this White Paper summarizes the outputs of the year-long reflection, its content does not engage the participants of the group.

Agustina Del Campo

Director, Center for Studies on Freedom of Expression and Access to Information (CELE) at Universidad de Parlermo

Argentina

Mette Finnemann

Head of Tech for Democracy, Ministry of Foreign Affairs of Denmark

Denmark

Chris Gray

Author and Activist, Former Community Operations Analyst at Meta

Ireland

Apar Gupta

Executive Director, Internet Freedom Foundation

India

Rigobert Kenmogne

Project Officer, Paradigm Initiative

Cameroun

Leïla Mörch

Program Manager Europe, Project Liberty

France – USA

Julie Owono

Executive Director, Internet Without Borders

France

Christian Peronne

Head Rights and Technology and GovTech Teams, Institute of Technology and Society of Rio de Janeiro (ITS Rio)

Brazil

David Sullivan

Executive Director, Digital Trust & Safety Partnership

USA

Prateek Waghre

Policy Director, Internet Freedom Foundation

India

Contact

Jérôme Barbier

Head of Outer Space, Digital & Economic Issues
Policy Department | Paris Peace Forum

jerome.barbier@parispeaceforum.org

Pablo Rice

Cyberspace Governance Policy Officer
Policy Department | Paris Peace Forum

pablo.rice@parispeaceforum.org

I) Framing the issue

A) Moderation of online content: a matter of global governance

Since the emergence of Web 2.0, the Internet has radically expanded the scope of the public sphere and the potential for intersubjectivity by providing its users with an unprecedented means for instantly interacting and sharing content across the globe. With the increased uses of such technologies came increased volume of information globally, including both accurate and inaccurate information, mainstream and non-mainstream opinions, child friendly as well as inappropriate content for children. Therefore, contents deemed "harmful" are coexisting with contents considered – in contrast - as tolerable, in an absolute or relative way.

The traditional debate on the extent and limits to the freedom of expression thus soon had to be raised for this new, extended public sphere. The global scale and immediateness of the cyber environment however bring critical tensions in any possible answer. While traditional media are subject to time-tested oversight, the dissemination of content online occurs at a speed that often does not allow for editorial decisions and spans a much broader context of competing values than exists at the national level.

Many actors around the Globe recently tried to solve this issue through a regulative approach towards "harmful' content. But the ambiguous relationship between this freedom and the notion of "harm", as reflected in *article 19 of the International Covenant on Civil and Political Rights (ICCPR)*[1], is particularly at stake today in relation to practices of content moderation by large platforms or internet service providers, and to new expectations from public authorities as well as public opinions across the world in this regard[2].

While the very core parameters of this debate at the crossroad of politics, law and moral have not been swept away by the advent of internet, the policy concerns raised by the actual weight of social media (4.95 billions of users estimated for 2022[3]) as well as the amount of moderation decisions to be made by platforms must now be addressed on a different scale.

Unlike traditional areas where freedom of expression is at stake, moderation and regulation of online content is an issue that cannot be dealt with within national or

[1] United Nations General Assembly Resolution 2200A (XXI), 16 December 1966, Article 19: *"(1) Everyone shall have the right to freedom of expression; (…) (2) The exercise of the rights (…) carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order, or of public health or morals."*

[2] For an overview, see: Herbert Smith Freehills, "Race to regulate: online harms" (2021)

[3] We Are Social & Hootsuite, "Digital 2022: Global Overview Report" (2022)

regional borders, or through usual regulatory schemes[4], but is a matter of global governance. This change of dimension and perspective proceeds from 3 major characteristics:

1- **Regulation of online content is a transnational process.** In line with the broader dynamics of cyberspace, the dissemination of online content is cross-border phenomenon addressed by a range of geographically situated actors. While potentially problematic content does not stop at the borders of a State, the capacity of stakeholders to act may be limited to the boundaries of a jurisdiction. At the same time, each jurisdiction retains its own understanding of the issue, which does not necessarily converge with the approach of another jurisdiction. This jurisdictional "patchwork" therefore adds an additional layer of complexity for platforms, which a lack of cooperation and interoperability is likely to further accentuate. In addition, the adoption of a regulation in a given jurisdiction may have direct or indirect extraterritorial effects. This may be due to, among other things, the size of the market to which the regulation applies, or the location of service providers' headquarters. Finally, the concentration of most of the biggest platforms in the United States may have influenced the way global content moderation policies are shaped due to an over-reference to national concepts, and may have increased the gaps with foreign frameworks.

2- **Global private actors are the key enforcers of an increasingly fragmented normative regime.** Service providers operate in a complex environment composed of norms, principles and standards that are either part of the platforms' internal policies, enacted in an ad hoc fashion, or are derived from national or regional law requirements. But in the end, and apart from cases of direct intervention by public authorities, private companies - and ultimately their front-line moderators and software - are the crucial enforcement agents. Even when it comes to implementing state-originated norms, public authorities do not have the material means, access and skills to police thoroughly the content shared on platforms. It should be noted, however, that the platforms themselves, although in the best position to monitor the content they host and act when necessary, operate with limited means considering the multiple and sometimes contradictory requirements against which they must arbitrate in specific situations.

3- **Potential consequences of content moderation for public interest are unprecedented.** The suspension of the accounts of U.S. President Donald Trump by major social networks in the wake of the January 6, 2021, insurrection at the

[4] For an overview of "conventional" regulatory schemes, as distinct from the notion of governance, see: Hans J. Kleinsteuber, "The internet between Regulation and Governance" in in: Möller/ Amouroux (eds), *OSCE Representative on Freedom of the Media, The Media Freedom Internet Cookbook* (2004), 61–75

Capitol Hill was probably the most striking example of the far-reaching implications of platform decisions in the public sphere. At the same time, service providers' inaction has often been blamed for the spread of extreme ideologies in society – such as terrorist propaganda (e.g. during the rise of the Islamic State) or incitement to hatred (e.g. incitements to genocide in Myanmar) – as well as for the success of electoral manipulations. This has further revived the debate on the democratic legitimacy as well as on accountability of platforms which have become, despite their traditional positioning as mere intermediaries, key players in public life.

B) Tackling "Legal but Harmful" Content: a renewed challenge for content governance

Content moderation is currently mainly carried out with regard to user-generated content deemed "harmful" by a variety of stakeholders but overwhelmingly legal speech. Until recently, harmful contents were primarily defined by national laws as corresponding to categories of illegal content (e.g. child pornography, terrorist propaganda, and all forms of content deemed to fall under criminal law), and then by the internal and discretionary rules of each platform (through their "terms of service" or "community guidelines") for its own services and the content they host.

This state of the art, whose implementation was already challenging since it placed the burden of assessment *in concreto* on the service providers themselves in the first place, was recently disrupted by several innovative regulatory intervention across the world - among which the European Union's Digital Service Act is the most advanced example - prescribing restrictive measures that target "harmful" content as distinct from illegal content. These new regulatory frameworks establish behavioral obligations for internet intermediaries ranging from risk analysis to rapid withdrawal of reported content, the non-performance of which engages their responsibility in a very concrete manner.

Authority	Title of the regulation	State of progress (November 2022)	Reference to "harmful" content
European Union	Digital Service Act	Adopted by European Parliament in July 2022, to take full effect in 2024.	Recital n°5: "This *Regulation should apply to providers of intermediary services, and in particular intermediary services consisting of services known as 'mere conduit', 'caching' and 'hosting' services, given that the exponential growth of the use made of those services, mainly for legitimate and socially beneficial purposes of all kinds, has also increased their role in the intermediation and spread of unlawful* ***or otherwise harmful*** *information and activities*"[5]
United Kingdom	Online Safety Bill	Discussed in UK parliament.	Part 3, Chapter 7, 54, (3): *"Content that is harmful to adults" means (a) priority content that is harmful to adults, or (b) content (…) of a kind* ***which presents a material risk of significant harm to an appreciable number of adults*** *in the United Kingdom"*[6]
California, United States	California Age-Appropriate Design Code	Adopted by Californian parliament, to take full effect in 2024.	1798.99.31, 1, (B) "*The Data Protection Impact Assessment shall address, to the extent applicable, all of the following: (i) Whether the design of the online product, service, or feature could harm children, including by* ***exposing children to harmful, or potentially harmful, content*** *on the online product, service, or feature.*"[7]

[5] European Parliament legislative resolution on the proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, 5 July 2022

[6] United Kingdom Parliament, Online Safety Bill (121 2022-23), as amended in the Public Bill Committee, 28 June 2022

[7] California State Assembly Bill n°AB-2273, California Age-Appropriate Design Code, Act, Chapter 320, 2022 Statutes, 15 September 2022

Singapore	Online Safety (Miscellaneous Amendments) Bill	Discussed in Singaporean Parliament	*Section 45L, Subsection 4 "An online Code of Practice issued or amended under this section applicable to providers of any regulated online communication service or specified types of such providers may provide for all or any of the following:* *(a) requirements that a provider of the regulated online communication service must, by establishing and applying appropriate systems or processes, provide the service in a way that —* *(i) prevents Singapore end-users of its service (particularly children of different age groups) from accessing content that presents a* ***material risk of significant harm to them****; and* *(ii) mitigates and manages the risks of danger to Singapore end-users of its service (particularly children of different age groups) from content provided or that may be provided on its service;* *(b) any matter so as to provide practical guidance or certainty in respect of what content* ***does or does not present a material risk of significant harm*** *to Singapore end-users generally or certain types of Singapore end-users of the service"*[8]

Non-exhaustive summary of the main regulatory efforts creating a category of harmful content distinct from illegal content.

However, these legislations rarely define the kind of content which should be qualified as "harmful", even though a set of new obligations depends on it, relying on judicial appreciation. This lack of a clear definition *a priori* makes it harder for platforms to act appropriately within the global ecosystem of content governance. Service providers are no longer mere enforcers of content-specific restrictions provided by law, in the spirit of article 19 ICCPR for instance, but are required to proceed themselves to the determination of what harmful means under the law(s) to which they are subject. Such a function is traditionally performed by courts as part of their role in interpreting the law in relation to cases brought before them. But judicial means do not match the time frame of dissemination of potentially harmful content online. Therefore, they may not be the most appropriate way to effectively prevent or respond to online harms while ensuring that the rights of all stakeholders involved are best preserved.

[8] Parliament of Singapore, Bill No. 28/2022, Online Safety (Miscellaneous Amendments) Bill, 3 October 2022

Platforms are therefore faced with an unprecedented responsibility, far from the regulatory environment in which they developed – characterized by the concept of intermediaries' immunity from liability for third-parties content as enshrined in the Section 230 of Title 47 of the United States Code. This growing legal uncertainty places service providers in a dilemma. On the one hand, they can choose to adopt a restrictive conception of the notion of "harm", at the risk of incurring liability for non-performance of their obligations. On the other hand, they might be tempted to over-compliance as a precaution, to the potential detriment of users' rights - in particular freedom of expression. It also raises critical questions when it comes to legitimacy of such intervention on material content, performed by private operators rather than a public authority, while the qualification of the "harm" and subsequent actions taken on this basis can assume a character of public interest.

Moreover, the challenge here is not only to articulate the different national requirements within a platform-specific content policy, but to achieve a common understanding among the service providers themselves. Recent situations (such as the Buffalo shooting in May 2022 that was live-streamed on Twitch before being spread to other platforms) demonstrates that the dissemination of content that appears to be problematic is not confined to a single platform for long.

The current content governance structure therefore lacks a common and clearly established process for qualifying and, if necessary, to guide the response against "harmful" content online while taking due account of the diverse expectations, requirements and rights of all stakeholders.

II) Methodology

Following a discussion during the 4th edition of the Paris Peace Forum (*To harm or not to harm: defining "“harmful content"*, Friday 12 November), a prospective working group was launched in 2022, aiming to explore possible core principles and processes for a comprehensive qualification of harmful content, should it be differentiated from illegal content. It draws on the advantages of the Paris Peace Forum's multistakeholder platform and community, notably its unique convening power and experience in fostering multi-actor consensus on core global governance issues.

From January 2022, regular meetings were held every two months under the Chatham House rules to enable free, agile and inclusive exchanges between public authorities, representatives from the civil society and key platforms across the world to identify key challenges and design common core principles. During these sessions, key observations, ideas and lessons learned were collected from a diverse group of experts, practitioners and stakeholders. This progress report seeks to capture the key findings from these discussions, and to set out a path forward for the governance of "harmful"

content. The 5th edition of the Paris Peace Forum (November 11-12, 2022) was also an opportunity for the working group to meet in person and share its conclusions with a range of interested actors.

III) A path toward an operational "harmful" content governance framework: key recommendations

Although considered in the early stages, the prospective working group did not attempt to determine concretely the harmfulness of online content. Rather, it has attempted to take a step back and to outline a method for designing a due diligence process that can achieve such a goal, by identifying key challenges and gaps, as well as by providing some caveats while replacing the issue in the framework of a larger international, political debate in which the role and legitimacy of each actor to act should be taken into account.

This first year of reflection led to identify the following critical issues that should be addressed when attempting to build an effective content governance architecture:

a) **Determining the right scale is key**. The existence of a "harm" is always a function of the context of value in which the fact at stake occurs. The perception of damage done to persons, to public or private interests, to public order or even morality will vary from one environment to another[9]. Norms, in particular legal norms, then strive to transcribe this relationship. When applied to online content, this means that the same content is likely to be considered harmful in one place of dissemination, while it will be tolerated and considered as legitimate in another. This age-old tension between seeking common ground and considering specific contexts is thus one of the biggest challenges when it comes to building a shared process for the governance of harmful content. In all societies, the rule of law makes it possible to overcome this complexity in part because its enactment results of a choice between divergent interests and expectations that is biding to the entire community. Private persons then define for themselves and their relationships the principles that govern them, in accordance with the law. An example is the platforms' community guidelines that enshrines the relationship between it and the user, in relation to the services it provides, consistently with the definition of "harm" defined in law. An additional complexity is added, however, when it comes to a global issue for which a

[9] In this sense, see: Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler & Jed R. Brubaker, "Understanding international perceptions of the severity of harmful content online", *PLoS One* (2021). The study's findings "*show significant differences in perceptions of harmful content across different countries in the world. The cross-country disagreement widely existed in different facets of the analysis—in different countries' rankings of content, in individual types of harmful, and in the higher-level topics (the authors) used to categorize individual content types*."

global response is sought. As mentioned earlier, platforms face a "patchwork" of rules and rights across the world that do not all converge. One solution could be to refer to internationally recognized rules and rights as a baseline when qualifying the harm – at least in the cases of cross-border dissemination.

b) **Multi-actor as an answer to the quest for legitimacy.** Whether or not it is established by law in the first place, a governance process for qualifying "harmful" content should be based on a broad and inclusive mode of participation among all stakeholders – public authorities, enterprises, civil society - that can best represent divergent interests and values and make the issue of content regulation subject to a genuine public debate. This could lead to a better understanding of the expectations and constraints of each actor involved at different stages of content regulation process, and improve coordination to bridge the gaps between norm setting and actual implementation. All actors would also be able to have access to more transparent information on content moderation processes, whose current opacity is often questioned. In addition, Global South should be fairly represented in such a process, considering the gaps that exist today between the North-based platforms and the true consideration of the values and contexts in other regions. The idea emphasized here is therefore that not all the weight of the decision should necessarily rest on the platforms alone. As with every representative decision-making format, and since elections on a global scale seem more than unlikely, one challenge that remains to be overcome, however, is the selection of representatives who enjoy sufficient legitimacy.

c) **Considering a taxonomy of harmful online content based on existing classifications for the regulation of other mediums of public expression.** The classification of harmful content, similar to what is done for illegal content, helps to make content policies truly operational and proportionate. Although the singularities of online platforms compared to traditional media have been highlighted previously, an attempt to classify online content could be made by taking as a reference the regulatory frameworks proven for curbing harmful content in traditional fields of information - such as radio, television, or print media.

d) **The qualification process should be evidence-based.** The process of qualification of the harm should not be done exclusively by reference to the applicable law(s), but should also be rely on regular, comprehensive and transparent investigations of the actual impacts of the diffusion of certain types of contents on a set of social groups over time. Although this field of

research is currently very dynamic, there is a lack of evidence from rigorous studies on the harmful consequences of platform's use, whether on the individual, the group, the political body, etc. In order to avoid bias and methodological discrepancies, the terms of the studies, such as the level of granularity or the type of effect investigated (e.g. psychological, social, political), shall be subject to prior agreement.

e) **Taking due account of the human factor in the moderation process.** Content moderation still relies on human agents, whether they perform their task in full autonomy or with the assistance of automated systems[10]. In the end, it is therefore individuals who are the key enforcers of content policies. Any harmful content governance process should therefore take full account of this "human-in-the-moderation-loop". This should be done both by including content moderators at the deliberative stage, and by ensuring that the policies and/or decisions formulated stick to the realities on the ground so that they can be effectively implementable in daily work. It is not only a question of effectiveness, but also of the well-being and social rights of the moderators.

[10] In this sense, see : The Internet Commission, "Accountability Report 2.0: An independent evaluation of online trust and safety practice" (2022). The report concludes the leading organizations "seek an appropriate synergy of human and automated systems", with platforms incorporating human moderators into their processes in different ways.

Harmful Content Working Group: Progress Report

A multistakeholder effort to unpack the idea of "harmful" content

Contact

Jérôme Barbier

Head of Outer Space, Digital & Economic Issues
Policy Department | Paris Peace Forum

jerome.barbier@parispeaceforum.org

Pablo Rice

Cyberspace Governance Policy Officer
Policy Department | Paris Peace Forum

pablo.rice@parispeaceforum.org