

A FRAMEWORK OF FIRST REFERENCE

Decoding a human rights approach to content moderation in the era of “platformization”

Jacob Mchangama, Natalie Alkiviadou and Raghav Mendiratta



"A framework of first reference": Decoding a human rights approach to content moderation in the era of "platformization"

© Justitia and the authors, November 2021

WHO WE ARE



Justitia

Founded in August 2014, Justitia is Denmark's first judicial think tank. Justitia aims to promote the rule of law and fundamental human rights and freedom rights both within Denmark and abroad by educating and influencing policy experts, decision-makers, and the public. In so doing, Justitia offers legal insight and analysis on a range of contemporary issues.



The Future of Free Speech Project

The Future of Free Speech is a collaboration between Justitia, Columbia University's Global Freedom of Expression and Aarhus University's Department of Political Science. At the Future of Free Speech, we believe that a robust and resilient culture of free speech must be the foundation for the future of any free, democratic society. To understand better and counter the decline of free speech, "The Future of Free Speech" project will seek to answer three big questions: Why is freedom of speech in global decline? How can we better understand and conceptualize the benefits and harms of free speech? And how can we create a resilient global culture of free speech that benefits everyone?

The publications can be freely cited with a clear indication of source.

The project is sponsored by:





Table of Contents

Executive Summary	2
Introduction	4
Contextual Background	4
Why IHRL as a “Framework of First Reference?”	5
Methodology	9
Chapter 1: Freedom of expression as the point of departure for handling content moderation	10
Provided by Law	11
Necessity	12
Proportionality	13
Chapter 2: Hate Speech	16
Article 20(2) ICCPR - Prohibition of Hate Speech	18
Article 20(2) -Threshold	19
Platform regulation of hate speech:	24
Case Studies	28
Uyghyr Muslims in China – Facebook	28
Assam – Facebook	31
Chapter conclusions	34
Chapter 3: Disinformation	35
Introduction	35
The First Step in Effectively Dealing with Disinformation: Defining Disinformation	37
Regulation of Disinformation under IHRL	39
Platform Policies on Disinformation	41
Identifying a Threshold for Disinformation	47
Case Studies	49
Modi and COVID-19	49
Stanford Professor of Disease Prevention, COVID-19 Interview Removed	50
Prescribing plant fertilizer to kill COVID-19	52
Disinformation and violence after the 2020 US Elections	54
Chapter Conclusions	56
Concluding comments	57

Executive Summary

4.66 billion people have Internet access, and 4.20 billion are active social media users. Despite the unprecedented scale and ease with which information and opinions are shared globally, Internet freedom is seen more and more as both a curse and a blessing. On one hand, social media use has empowered previously silenced groups to mobilize and find ways around traditional forms of censorship. On the other, such platforms have become vehicles for phenomena such as hate speech and disinformation.

Authoritarian regimes as well as liberal democracies are placing increasing pressure on social media platforms to deal with allegedly harmful content. For example, the German Network Enforcement Act (NetzDG) imposes a legal obligation on social media companies with more than 2 million users to remove manifestly illegal content, including insult, incitement, and religious defamation, within 24 hours or risk a fine of up to 50 million Euros. The NetzDG blueprint for “intermediary liability” has been followed by over 20 countries around the world, including Belarus, Turkey, Venezuela, and Russia.

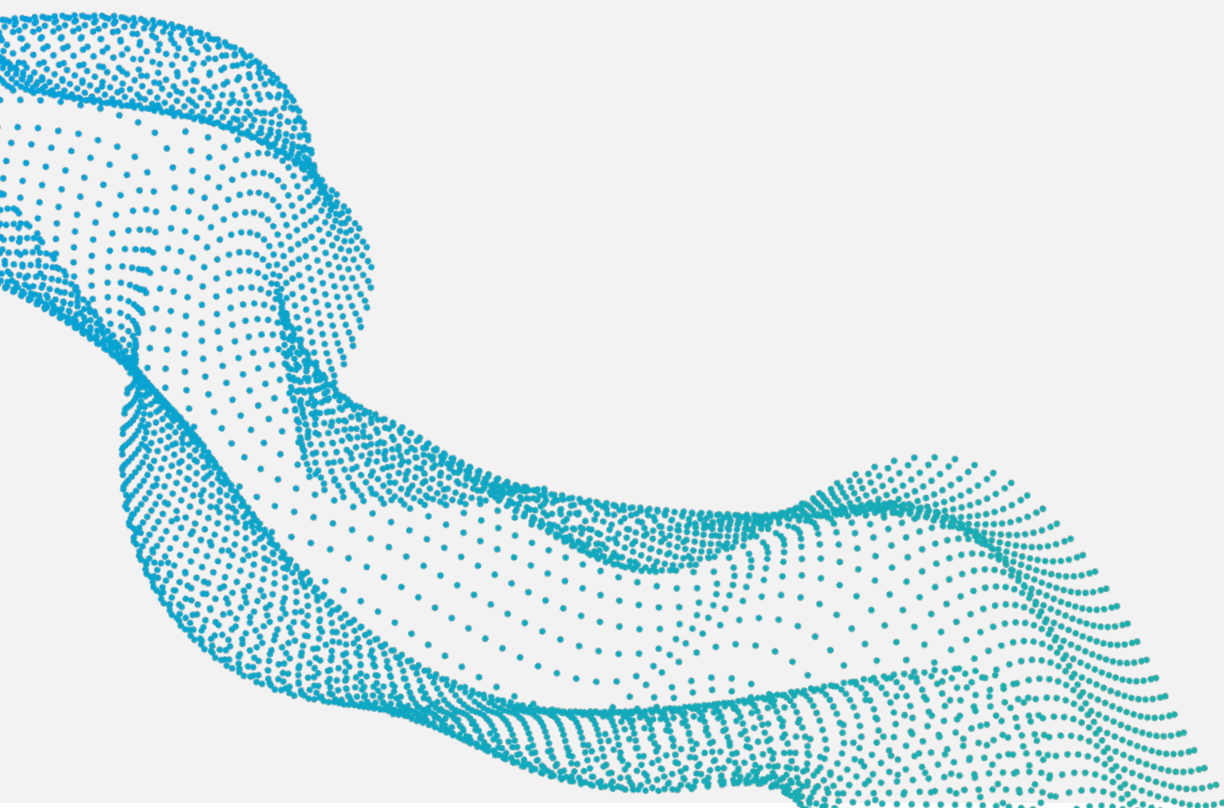
Such legislative measures and the pressure they bring to bear have contributed to a regulatory race to the bottom, and social media platforms have become the ultimate arbiters of harm, truth, and the practical limits of the fundamental right to freedom of expression. This is demonstrated by the drastic increase in content removal over the last few years. For example, Facebook removed 2.5 million pieces of content in Q1 of 2018 for violating its Community Standards on Hate Speech. This rose to 22.3 million pieces of content in Q3 of 2021. With respect to disinformation, Twitter announced that, between March 2020 and July 2020 alone, it took down 14,900 Tweets and “challenged” 4.5 million accounts that regularly posted COVID-19 misinformation. Yet, the standards and practical methods used to regulate content moderation are often vague, conflicting, and nontransparent, which has serious negative consequences for the practical exercise and protection of freedom of expression for users around the world.

As the “great bulwark of liberty”, freedom of expression must be respected and upheld. International Human Rights Law (IHRL) has provided for freedom of expression in both the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights (ICCPR). As private entities, social media platforms are not signatories to or bound by such documents, but, as the former Special Rapporteur for Freedom of Opinion and Expression David Kaye has argued, IHRL is a means to facilitate a more rights-compliant and transparent model of content moderation. At the same time, the global nature of IHRL may also prove useful in dealing with the differences in national perception and legislation that characterize the global ecosystem of online expression. Yet, applying IHRL to private companies is a difficult task involving a plethora of challenges and dilemmas.

In this report, Justitia sets out IHRL as a “framework of first reference” for moderating online hate speech and disinformation. It decodes relevant IHRL principles, applies them to hate speech and disinformation through real-life examples, and offers recommendations on their adoption by social media platforms. The report explains how a human rights approach may be implemented by such platforms to bring about a rights-protective and transparent moderation of online content.

We argue that, to be compliant with IHRL, a platform’s content moderation practices must be legitimate, necessary, and proportional within the framework of Article 19(3) ICCPR (restrictions on freedom of expression), which sets out the grounds for limitation of freedom of expression. For hate speech, platforms should frame terms and conditions based on a threshold established by Article 20(2) ICCPR (prohibition of advocacy of hatred) and take strictly into consideration the Rabat Plan of Action’s six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm before taking any enforcement action. For disinformation, a platform’s terms and conditions should be tailored to protect the grounds in Article 19(3) ICCPR and Article 25 ICCPR (right to participate in voting and elections). In addition, platforms must refrain from adopting vague blanket policies for removal. Only disinformation promoting real and immediate harm should be subject to the most intrusive restrictive measures such as content removal. In determining the limits of disinformation, platforms should focus on the post’s content, its context, its impact, its likelihood of causing imminent harm, and the speaker’s intent.

Justitia recommends that major platforms formally commit to adopting an IHRL approach to content moderation by signing a non-binding Free Speech Framework Agreement (FSFA) administered by the Office of the UN High Commissioner for Human Rights (OHCHR) under the specific auspices of the Special Rapporteur on Freedom of Opinion and Expression.



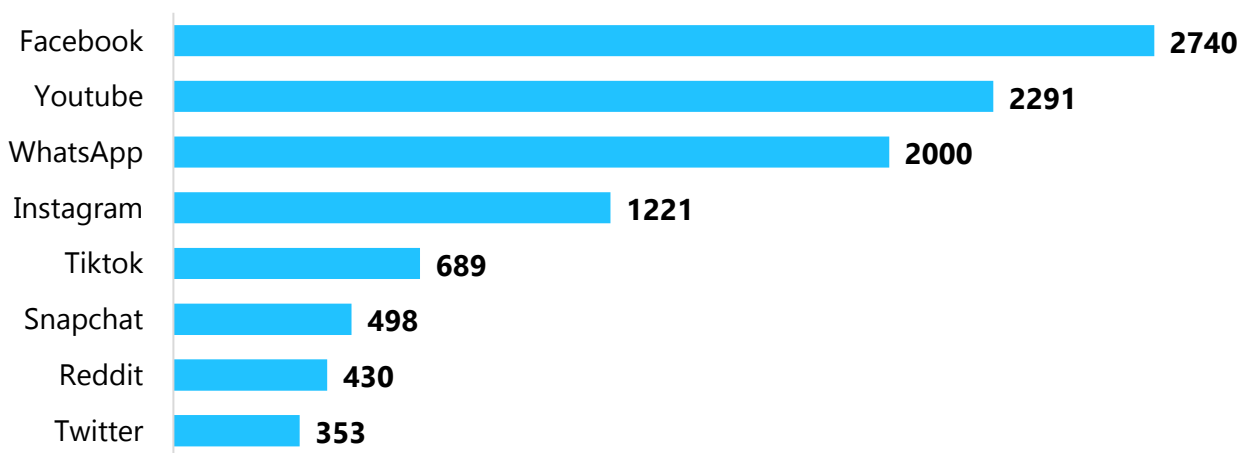


Introduction

Contextual Background

At no other time in history have ordinary people around the world been able to share and access information at the scale, speed and global reach available in the current digital era. Almost 60% of the global population – 4.66 billion people -- are online and 4.20 billion are active social media users. Facebook, YouTube, WhatsApp, Instagram, TikTok, Snapchat, Reddit and Twitter (in that order) are the platforms with the highest number of active monthly users.¹

Global active users (in millions) for some of the world's top social media platforms (Jan. 21)



The digital era has given voice to previously marginalized groups and provided ways to circumvent traditional gatekeepers and censorship. However, as more and more individuals are given space on centralized platforms, the harms of free speech have been amplified since this process of “platformization” provide extremism, hatred, abuse and disinformation with new visibility.

Hate speech has long been a topic of deliberation, legislation and judicial decisions across continents and within International Human Rights Law (IHRL), but its dissemination on private platforms has made the question of its definition and regulation even more acute. Similarly, especially after the 2016 US presidential election and the COVID-19 pandemic, concerns have been expressed about the viral spread of disinformation. These developments have also raised serious questions about the legitimacy of private platforms as the arbiters of truth and deciders of harm as well as practical concerns about the suitability of large-scale content moderation and the dangers of privatized censorship.

¹ Simon Kemp, 'Digital 2021: Global Overview Report' (2021), *Datareportal*, <<https://datareportal.com/reports/digital-2021-global-overview-report#:~:text=Internet%3A%204.66%20billion%20people%20around,now%20stands%20at%2059.5%20percent>>



As a result, more and more countries have passed legislation that restricts online speech and requires social media platforms to remove illegal or “harmful” user content by means of intermediary liability.² The pressure from states, users and civil society to police various forms of harmful content has also caused platforms to remove more and more content with every passing year, as evident by their increased takedown figures (discussed in detail in the chapters on hate speech and disinformation). So, while platforms were once focused on expanding the limits of free speech and access to information, they now increasingly rein in the exercise of these freedoms.

This is in stark contrast to the initial promise of the Internet as an ever-expanding global free speech zone as described in 1999 by Professor Lawrence Lessig:

“Nations wake up to find that their telephone lines are tools of free expression, that e-mail carries news of their repression far beyond their borders, that images are no longer the monopoly of state-run television stations but can be transmitted from a simple modem.”³

Therefore, content moderation by social media companies is one of the key issues affecting the practical exercise of free expression around the world. The global nature of major social media platforms creates significant problems when it comes to determining where to draw the line on various categories of content.

Why IHRL as a “Framework of First Reference?”

This report aims at grounding solutions within IHRL as a “framework of first reference” with a specific focus on freedom of expression. Such a framework promotes making freedom of expression, as envisioned by IHRL, the basis for (large) platforms’ content moderation processes: from developing community guidelines/standards and terms of service/use that are compliant with IHRL standards on freedom of expression to embedding free speech safeguards in content-filtering algorithms and training content moderators in free speech standards.

The report argues that IHRL may constitute a sustainable way forward in promoting the practical exercise of the universal right to freedom of expression and opinion including the right to “seek, receive and impart information and ideas through any media and regardless of frontiers” as guaranteed by the Universal Declaration of Human Rights (UHDR) and to prevent the harms of private censorship of online content, which may be accentuated by government pressure. This approach has also been encouraged by various scholars and experts.

² Jacob Mchangama & Natalie Alkiviadou, ‘Digital Berlin Wall: How Germany (Accidentally Created a Prototype for Global Online Censorship – Act Two’ (2020) *Justitia* <https://justitia-int.org/wp-content/uploads/2020/09/Analyse_Cross-fertilizing-Online-Censorship-The-Global-Impact-of-Germanys-Network-Enforcement-Act-Part-two_Final-1.pdf>

³ Lawrence Lessig, ‘Code Version 2.0’ (2006) *Basic Books*, p. 236.



In his 2018 report on the regulation of user-generated online content, the former UN Special Rapporteur on the Freedom of Opinion and Expression (SRFOE) David Kaye proposed a framework for content moderation that “puts human rights at the very centre”. He stressed that national laws are inappropriate given the geographical and cultural diversity of digital users and that IHRL provides a framework to address this central difficulty because it transcends national boundaries. Kaye stressed that relying on IHRL to determine acceptable and unacceptable speech “enables forceful normative responses against undue State restrictions – provided companies play by similar rules.”⁴ He pointed to the UN Guiding Principles on Business and Human Rights, which provide that private companies must respect IHRL, noting that this obligation “exists independently of States’ abilities and/or willingness to fulfill their own human rights obligations and does not diminish those obligations”.⁵ In her 2021 Report on Disinformation, David Kaye’s successor as SRFOE, Irene Khan, calls for multidimensional responses to disinformation that are grounded in the IHRL framework.⁶

Several leading academics echo this sentiment. Evelyn Aswad argues that international law is the most suitable framework for protecting freedom of expression.⁷ Similarly, Susan Benesch suggests that, even though IHRL cannot be used “right off the shelf”, it can be the framework for content moderation.⁸ Hilary Hurd underlines that, while Article 19 of the International Covenant on Civil and Political Rights (ICCPR) only applies to states, there have been “renewed calls to apply Article 19 to technology companies”.⁹ A team of researchers drafted a set of sixteen recommendations put forth by the Israel Democracy Institute and Yad Vashem, which provide “policy guidelines and benchmarks” for content moderation “anchored in the applicable human rights standards”.¹⁰ Barrie Sander also endorses the adoption of a human rights-based approach to content moderation since this would provide social media platforms with a “common conceptual language to identify the impact of their moderation rules”.¹¹ Thiago Oliva builds on this by arguing that the ICCPR provides a “methodology and vocabulary for platforms to analyze whether their content policies and decisions are reasonable”.¹² In a report by Stanford’s Law and Policy Lab, Sarah Shirazyan and others provide extensive insight into international case law, national legislation and social media content policies in

⁴ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) A/HRC/38/35, para. 42, <<https://www.undocs.org/A/HRC/38/35>>

⁵ UN Guiding Principles on Business and Human Rights (2011), https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

⁶ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2021) A/HRC/47/25, para. 87, <<https://undocs.org/A/HRC/47/25>>

⁷ Evelyn Mary Aswad, ‘The Future of Freedom of Expression Online’ (2018) 17 *Duke Law & Technology Review* 1, 52-53.

⁸ Susan Benesch, ‘But Facebook’s Not a Country: How to Interpret Human Rights Law for Social Media Companies’ (2020) 38 *Yale Journal on Regulation Online Bulletin*, 90.

⁹ Hilary Hurd, ‘How Facebook Can Use International Law in Content Moderation’ (2019) *Lawfare*, <<https://www.lawfareblog.com/how-facebook-can-use-international-law-content-moderation>>

¹⁰ ‘A Proposed Basis for Policy Guidelines for Social Media Companies and Other Internet Intermediaries’ *IDI-Yad Vashem*, <<https://www.idi.org.il/media/13570/recommendations-for-reducing-online-hate-speech.pdf>>

¹¹ Barrie Sander, ‘Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation’ (2020) 43 *Fordham International Law Journal* 939.

¹² Thiago Dias Oliva, ‘Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression’ (2020), 20 *Human Rights Law Review* 617.



relation to violent extremist organizations, misinformation and fake news, online defamation, and cyber harassment and bullying.¹³

The Oversight Board (an independent private body created by Facebook in 2020 to make final decisions regarding content moderation questions, including the evaluation of complaints by users) has also embraced IHRL in judging the appropriateness of Facebook's content moderation decisions. At the time of this writing, the Oversight Board has decided 17 cases involving a variety of issues ranging from hate speech to nudity to dangerous individuals and organizations; and, in all cases, the Board relies on relevant provisions and principles of IHRL.¹⁴

By adopting an IHRL approach to content moderation, private platforms would also accommodate user demand since many remain deeply skeptical about state regulation of social media. Justitia's 2021 global survey on attitudes towards free speech showed that people in two-thirds of the 33 countries surveyed prefer the regulation of social media content to be carried out solely by the companies themselves. While a plurality in the rest prefers the regulation of content to be carried out by social media companies along with national governments.¹⁵

That said, a number of scholars have also expressed important reservations about marrying IHRL and content moderation. Danielle Citron argues that IHRL is just too flexible to provide the level of clarity that social media platforms need.¹⁶ Although Evelyn Douek also underlines the problem of IHRL's flexibility and notes that there is little that actually compels such platforms to adhere to IHRL, she argues that it has the potential to develop "more concrete rules".¹⁷ Brenda Dvoskin takes a more rigid approach, suggesting that adopting IHRL "might not lead to more legitimate content moderation rules" because IHRL is not neutral and "leaves many speech questions unanswered".¹⁸

We recognize that using IHRL will not resolve all the thorny issues and dilemmas related to content moderation by private platforms and that it is unrealistic to expect that all content moderation decisions will be compliant with IHRL, even if all major platforms were to adopt an IHRL approach. We also acknowledge that an IHRL approach to content moderation will result in a significantly more speech-protective social media environment, leaving in place much content that is likely to be false

¹³ Sarah Shirazyan, Allen Weiner, Yvonne Lee & Madeline Magnuson, et al., 'How to Reconcile International Human Rights Law and Criminalization of Online Speech: Violent Extremism, Misinformation, Defamation, and Cyberharassment' (2020), *Stanford Law School Law and Policy Lab*,

<<https://law.stanford.edu/publications/how-to-reconcile-international-human-rights-law-and-criminalization-of-online-speech-violent-extremism-misinformation-defamation-and-cyberharassment/>>

¹⁴ Oversight Board Case decision 2020-006-FB-FBR, <<https://oversightboard.com/decision/FB-XWJQBU9A/>>

¹⁵ Svend-Erik Skaaning & Suthan Krishnarajan, 'Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech (2021) *Justitia*, <https://futurefreespeech.com/wp-content/uploads/2021/06/Report_Who-cares-about-free-speech_21052021.pdf>

¹⁶ Danielle Keats Citron, 'What to Do about the Emerging Threat of Censorship Creep on the Internet' (2017), *Cato Institute*, <<https://www.cato.org/policy-analysis/what-do-about-emerging-threat-censorship-creep-internet>>

¹⁷ Evelyn Douek, 'U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for Human Rights by Default' (2018), <<https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default>>

¹⁸ Brenda Dvoskin, 'International Human Rights Law Is Not Enough to Fix Content Moderation's Legitimacy Crisis' (2020), <<https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd>>



and misleading and/or cause offense and be deemed unacceptable/hateful/harmful by various states and constituencies across the globe.

We also recognize that an IHRL approach to content moderation on private platforms will necessarily have to be adapted to the specific circumstances of social media rather than copied wholesale for entities that are very different from states for which IHRL was developed to constrain and guide. Accordingly, an IHRL approach should be seen as an imperfect improvement rather than a perfect solution. Nevertheless, we believe that the adaption is possible and would be beneficial for moderating the hate speech and disinformation found on platforms in today's centralized social media environment.

Moreover, accepting the current absence of a basic framework setting out global norms for free speech on platforms would be akin to granting absolute discretion to a number of private companies, which constitute the central agora of global and local expression and whose content moderation policies have an enormous impact on the practical limits of freedom of expression around the globe. This is particularly problematic in the many countries in which official censorship and propaganda leave social media as the only way to express and organize dissent. IHRL as a "framework of first reference" will thus provide platforms legitimacy in resisting rising demands by states – as well as non-state actors -- to take down content that they claim is in violation of national laws but which may be protected under IHRL.

Ultimately, the future of free speech online may be best served by a more decentralized media environment and/or through enhanced user control over content. However, until such de-centralization is achieved, we believe that IHRL as a "framework of first reference" for major social media platforms may cultivate a more transparent, legitimate and speech-protective approach to handling online hate speech and disinformation. Moreover, IHRL as a "framework of first reference" should be capable of coexisting with enhanced user control of content, which would offset the very real concerns that IHRL may allow content that some/many users find too offensive, hateful or misleading to tolerate.



Methodology

This report is based on desktop research combined with interviews conducted with platform executives. In relation to the former, we have assessed:

- International/regional conventions/covenants
- Jurisprudence of the UN's Human Rights Committee (HRC)
- Reports of the SRFOE
- National practices (legislation/jurisprudence)
- Community guidelines/standards/terms of service/use of social media platforms and any other policy updates released on their blogs or through other media
- Scholarly input into the subject matter of the report.

In relation to empirical research, we conducted comprehensive interviews with senior policy executives of Facebook, Instagram, YouTube, Reddit and VERO to understand their approach to IHRL (with a focus of freedom of expression) in the ambit of moderating hate speech and disinformation.¹⁹

¹⁹ Please note that Justitia contacted several more social media platforms, including the Global South and East, but only those listed above responded to our request. Moreover, an informal discussion with Snapchat was held but, given its differing infrastructure (in terms of its private rather than public setup and the general ephemerality of its content), its relevance to this analysis is much more limited than social media platforms with, for example, public posts.



Chapter 1: Freedom of expression as the point of departure for handling content moderation

To understand how IHRL can serve as a framework for content moderation, it is necessary to provide a brief overview of the international law principles and standards on the right to freedom of expression – and limitations thereto – with specific emphasis on hate speech and disinformation.

The first step toward recognizing freedom of expression under IHRL was taken by the symbolic, non-binding Universal Declaration of Human Rights (UDHR) and Article 19 therein. The UDHR set the scene for freedom of expression, which was later incorporated within the ICCPR and, unlike other Covenant rights, came with “special duties and responsibilities”. The structure and content of Article 19 may also be compared to Article 10 European Convention on Human Rights (ECHR) that came before the ICCPR. The most significant body in interpreting IHRL is the HRC, which is the supervisory body of the ICCPR. It issues ‘Concluding Observations on States’, considers individual and inter-state complaints, and produces documents such as its ‘General Comments’. Although non-binding, these undertakings provide states (and other relevant stakeholders) with valuable guidance on the meaning and application of the ICCPR. With respect to the right to freedom of expression, the HRC’s General Comment 34 on ‘Article 19 – Freedom of Opinion and Expression’ is the most influential interpretive instrument. The SRFOE reports also provide important observations and interpretations of IHRL. Of particular relevance to this report are the SRFOE Reports on Disinformation and Freedom of Opinion and Expression (2021), Hate Speech (2019), Online Content Regulation (2018) and The Right to Freedom of Opinion and Expression exercised through the Internet (2011).

Article 19 states:

1. Everyone shall have the right to hold opinions without interference.
2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
3. The exercise of the rights provided for in paragraph 2 of this article 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.



Article 19 protects not only freedom of expression, which focuses on the right of the speaker, but also “the freedom to seek, receive and impart information and ideas of all kinds”, focusing on the rights of the reader/viewer/listener in addition to the rights of the speaker. Moreover, Article 19 applies “regardless of frontiers” and to “any..media...of choice”. This is of particular significance in the current digital age and demonstrates how this right can be used to support arguments against blocking and restricting online content. The SRFOE underlined the “unique and transformative nature of the Internet” as a means of receiving (and imparting) such information.²⁰ The relevant report also highlighted the increasing numbers and forms of restrictions on the right to information online, including blocking and content filtering that, when done outside the framework of IHRL, not only inhibit access to information but also the enjoyment of other rights, since freedom of expression is an “enabler” of other civil and political rights.

Although the right to freedom of expression is fundamental, it “does not enjoy such a position of primacy among rights that it trumps equality rights”.²¹ However, restrictions are to be stringently set, must meet the Article 19(3) test and conform to the strict test of proportionality.²²

Provided by Law

IHRL requires any restrictions to freedom of expression to be “provided by law”. When applying an IHRL approach to social media platforms, the terms of service/terms of use contain more general rules for each platform, and community guidelines/standards/policies provide the specific rules for various categories of content. For purposes of clarity, ‘Terms’ will be used in this report unless reference is made to a specific document of a platform. Terms are adopted unilaterally and at the discretion of private companies. If the relevant Terms do not reflect IHRL standards, non-conformity is likely to become systemic when applied by moderators (both human and AI).

For a restriction to be “provided for by law,” it must be formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and must be made accessible to the public.²³ A law may not confer unfettered discretion to restrict freedom of expression upon those charged with its execution,²⁴ and “laws must provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not.”²⁵

²⁰ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2011), A/HRC/17/27, Summary

²¹ Stephanie Farior, ‘Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech.’ (1996), 14 *Berkley Journal of International Law* 1, 3.

²² HRC General Comment 34: ‘Article 19 – Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34, para. 34.

²³ United Nations, Economic and Social Council, Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights, E/CN.4/1985/4, Annex (1985).

²⁴ HRC Concluding Observations: Lesotho (1999) CCPR/C/79/Add.106, para. 23.

²⁵ HRC General Comment 34: ‘Article 19 – Freedom of Opinion and Expression’ (2011), CCPR/C/GC/34, para. 24.

The requirement of precision and foreseeability has been eloquently expressed by the Norwegian Supreme Court in a hate speech case from 2002:

“The rule of law, and especially the consideration of foreseeability, dictates restraint when it comes to an expansive interpretation based on context. When it comes to punishable expressions, the point must be that you can only be punished for what you have said, not what you might have said. From this it follows, in my opinion, that, in the interest of freedom expression, no one should risk criminal liability by attributing to a statement a viewpoint that has not been expressly made and which cannot with a reasonably high degree of certainty be inferred from the context”.²⁶

With respect to the precision and accessibility of content moderation practices, we note that the publicly available Terms are often only the tip of the iceberg. For example, the confidential implementation standards developed to guide Facebook’s human content moderators are a constantly growing index of prohibited content with many moderators concerned about the “inconsistency and peculiar nature of some of the policies”.²⁷ There are at least two issues with the existing system: many content moderators struggle to understand fully and apply this index consistently, and there is a lack of transparency and public availability of the regulatory framework that binds users. Without this transparency, one of the central aims of the “provided by law” requirement is undermined, and the risk of arbitrariness and opacity is heightened.

Necessity

The requirement that a restriction be “necessary” means that the restriction pursues a legitimate aim, based on one of the grounds set forth in Article 19(3) which includes, amongst others that restrictions occur “for the respect of the rights and reputations of others”. General Comment 34 notes that the term “rights” refers to human rights, as recognized in the Covenant and, more generally, in IHRL, and notes that the term “others” may, for example, “refer to individual members of a community defined by its religious faith or ethnicity.”²⁸ In *Ross v Canada* (2000), the HRC held that the term “others” may relate to other persons or to a community as a whole.²⁹ In relation to the concept of morals as a limitation ground in Article 19, the HRC reiterated in *Fedotova v Russia* (2012) what it had stated in General Comment 34, namely, that:

²⁶ HR- 2001-01428 – Rt-2002-1618, (Saks nr. 361-2002) 17 December 2002 (our translation from Norwegian).

²⁷ Nick Hopkins, ‘Revealed: Facebook’s internal rulebook on sex, terrorism and violence’ (2017), *The Guardian*, <<https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence>>

²⁸ United Nations, Economic and Social Council, Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights, E/CN.4/1985/4, Annex (1985).

²⁹ *Ross v Canada*, Communication no. 736/1997 (18 October 2000) CCPR/C/70/D/736/1997, para. 11.5.

“the concept of morals derives from many social, philosophical and religious traditions; consequently, limitations ... for the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition. Any such limitations must be understood in the light of universality of human rights and the principle of non-discrimination”.³⁰

Fedotova v Russia

Proportionality

There must be an assessment of whether the restrictions are proportional to the legitimate aim pursued. The HRC has underlined that such restrictions must not be too broad and that proportionality must be “respected not only in the law that frames the restrictions but also by the administrative and judicial authorities in applying the law.” The HRC also notes that restrictions must be “appropriate to achieve their protective function” and “must be the least intrusive instrument amongst those which might achieve their protective function”.³¹ For example, it may be reasonable to restrict freedom of expression to protect the right to vote under Article 25 or the right to privacy under Article 17,³² but these restrictions must be carefully tailored. So, although it may be permissible to restrict speech that constitutes intimidation or coercion of voters, any such restriction must not impede political debate.³³ Public and political speech enjoys a particularly high degree of protection in the HRC’s jurisprudence. General Comment 34 states that “the free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential.”

As recognized in General Comment 34, Article 19 protects “even expression that may be regarded as deeply offensive”.³⁴ For example, “prohibition of displays of lack of respect for a religion... including blasphemy laws, are incompatible with the Covenant” and penalizing the expression of opinions about historical facts is incompatible with the ICCPR, since Article 19 “does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events.” The South African Constitutional Court provides a good example of how the HRC standard on “memory laws” can be protected in practice. In a case involving Holocaust denial, the Court did not accept the restriction of the impugned speech but, instead, underlined the importance of freedom of expression, since it “lies at the heart of a democracy. It is valuable for many reasons, including its instrumental functions as a guarantor of democracy, its implicit recognition and protection of the moral agency

³⁰ Fedotova v Russia, Communication no. 1932/2010 CCPR/C/106/D/1932/2010, para. 10.5, General Comment No. 34, para. 32.

³¹ Marques v Angola, Communications no. 1128/2002 (18 April 2005) CCPR/C/83/D/1128/2002 and Coleman v Australia, No. 1157/2003 (10 August 2006) CCPR/C/87/D/1157/2003.

³² HRC General Comment 34: ‘Article 19 – Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34; UN Special Rapporteur, Research Paper 1/2019 ‘Freedom of Expression and Elections in the Digital Age’ (2019).

<<https://www.ohchr.org/Documents/Issues/Opinion/ElectionsReportDigitalAge.pdf>>

³³ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) A/67/357, para. 12.

³⁴ HRC General Comment 34, ‘Article 19 – Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34, para. 11.



of individuals in our society and its facilitation of the search for truth by individuals and society generally.”³⁵

According to the then-UN High Commissioner for Human Rights, “defining the line that separates protected from unprotected speech is ultimately a decision that is best made after a thorough assessment of the circumstances of each case”.³⁶ In *Faurisson v France* (1996), the HRC underlined that the limitation grounds of Article 19 are not to be viewed as a “license to prohibit unpopular speech”.³⁷ In *Fedotova v Russia* (2012), the HRC noted that any limitations to Article 19 must meet the “strict test of necessity and proportionality”;³⁸ and, in *Kirill Nepomnyashchiy v Russia* (2019), the HRC underlined that legislation restricting the freedom of expression must be “strictly necessary and proportional to a legitimate aim set forth in that article and directly related to the specific need”.³⁹

Careful tailoring of speech restrictions, as required by the HRC, is rendered particularly complex online, given the sheer scale and speed of content creation. One of the issues here is timing. Laws such as the German Network Enforcement Act (NetzDG) compel platforms to make a rushed evaluation about the legality of content. This might lead to a situation in which the threat of heavy regulatory fines renders platforms prone to err on the side of caution and take down legitimate speech to protect themselves from potential liability.⁴⁰ Thus, such legislation creates an incentive to adopt overly speech-restrictive Terms, leading to too much removal or a “better safe than sorry” approach that might lead to a disproportionate removal of legal content.

Moreover, it is unrealistic to expect thousands of complex speech cases to be processed within hours while simultaneously attaching the proper weight to due process and freedom of expression. This might cause systemic collateral damage to the online ecosystem of information and opinion. Justitia’s 2021 report on takedown limits found that authorities in various EU jurisdictions took significantly longer than the time mandated for social media platforms to decide the legality of content. As compared to the short time limit (often 24 hours) granted by national laws to social media platforms to determine the legality of content, authorities in the Council of Europe states studied in the report took on average 778.47 days.⁴¹

Linked to the pressure for removal, speed and quantity of content is the use of artificial intelligence (AI). Every minute, Facebook users upload over 147,000 photos and YouTube users post 500 hours

³⁵ *Islamic Unity Convention v Independent Broadcasting Authority and Others*, Case CCT36/01 (11 April 2002) para. 26.

³⁶ Opening Remarks by Navanethem Pillay, UN High Commissioner for Human Rights, 2 October 2008, Expert Seminar on the Links between Articles 19 and 20 of the International Covenant on Civil and Political Rights; HRC General Comment 34, ‘Article 19 – Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34, para. 35.

³⁷ *Faurisson v France*, Communication no. 550/1993 (8 November 1996) CCPR/C/58/D/550/1993, Individual opinion of Elizabeth Evatt and David Kretzmer (Concurring Opinion).

³⁸ *Fedotova v Russia*, Communication No. 1932/2010 CCPR/C/106/D/1932/2010, para 10.3.

³⁹ *Kirill Nepomnyashchiy v Russia*, Communication no.2318/2013, para.7.8.

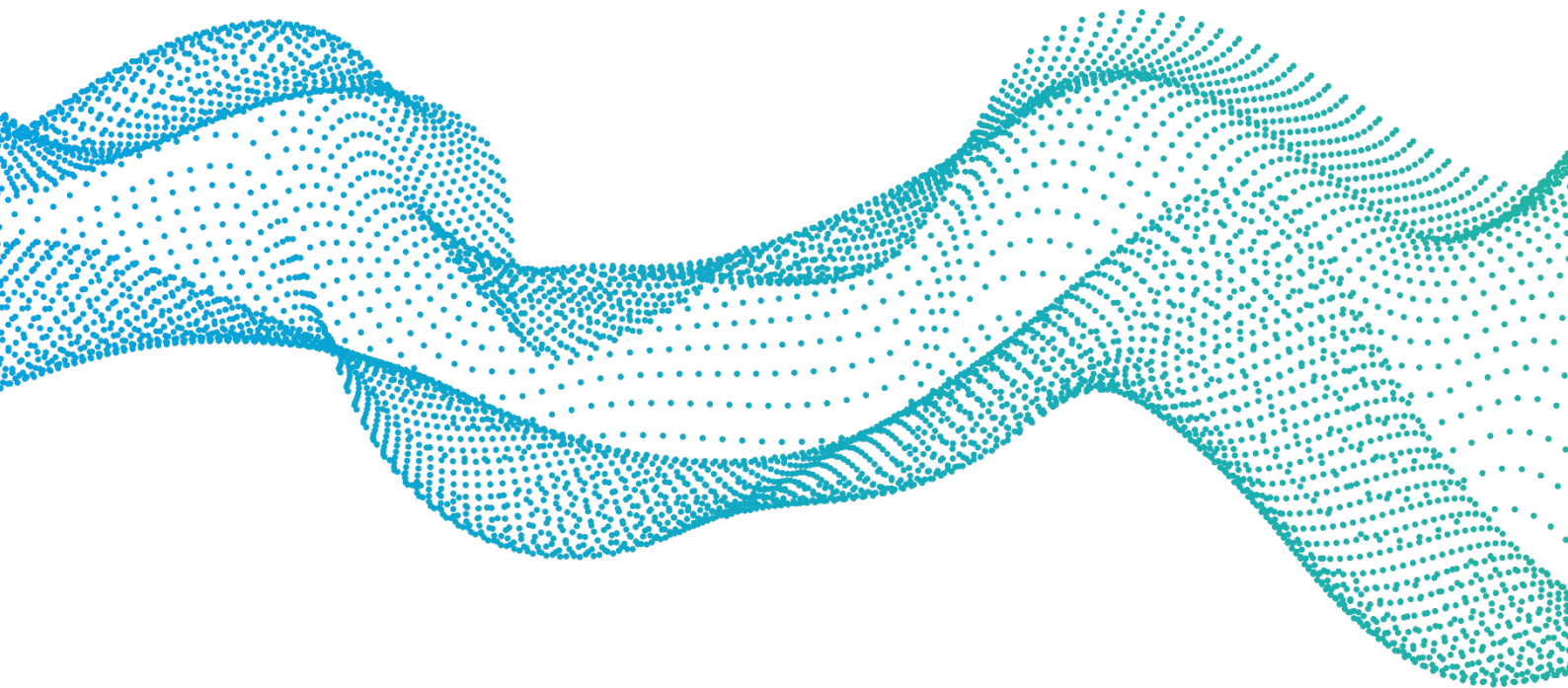
⁴⁰ Kaye, D. (2018). Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/HRC/38/35). United Nations Human Rights Council.

<https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx>

⁴¹ Jacob Mchangama, Natalie Alkiviadou & Raghav Mendiratta ‘Rushing to Judgment,’ <https://futurefreespeech.com/wp-content/uploads/2021/01/FFS_Rushing-to-Judgment-3.pdf>



of videos.⁴² Accordingly, platforms increasingly rely on automated content moderation to flag and remove violative content. In its latest enforcement report, Facebook noted that advancements in AI have allowed it to remove more hate speech, with its proactive rate being over 90% in 12 out of 13 policies (including hate speech). Its latest enforcement report notes that “our investments in AI enable us to detect more kinds of hate speech violations...Steady and continuous AI improvements and advancements...enable our AI models to spot hate speech”.⁴³ However, AI cannot realistically be subjected to a thorough human review that mirrors the standards of independent courts or tribunals, as required by IHRL.⁴⁴ AI-based filters might not judge content, tone and context in the same way that a human does and thus might remove too much content such as satire or humor. As such, the former SRFOE noted that calls to expand the use of automated tools for terrorist or other areas of content “threaten to establish comprehensive and disproportionate regimes of pre-publication censorship”.⁴⁵



⁴² Andrew Hutchinson, 'What Happens on the Internet Every Minute (2020 Version)' (2020) *Social Media Today*, <<https://www.socialmediatoday.com/news/what-happens-on-the-internet-every-minute-2020-version-infographic/583340/>>

⁴³ Facebook Community Standards Report, Second Quarter 2020: < <https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/>>

⁴⁴ Alexandra Severson, 'Facebook Admits It Was Used to Incite Violence in Myanmar' (2018), *New York Times*, <<https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html>>

⁴⁵ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 'Content Regulation' (2018) A/HRC/28/25 12



Chapter 2: Hate Speech

There are compelling reasons as to why open and democratic societies committed to values such as freedom, dignity and equality wish to counter incitement to hatred, since hate speech may cause harm to individuals and communities. In October 2016, Myanmar's military took to Facebook to incite large-scale violence against the Rohingya Muslim minority. Facebook overlooked several warnings and let the campaign continue before belatedly taking action.⁴⁶ Moreover, research shows that hate speech (online and offline) may lead to negative consequences for targeted individuals and groups.⁴⁷ For example, they may experience psychological trauma and fear,⁴⁸ and this may lead to self-censorship.⁴⁹ Studies have also shown a link between online hate speech and hate crimes.⁵⁰

However, it does not necessarily follow that banning hate speech is an efficient remedy that may be implemented without serious risks to freedom of expression. In particular, broadly-worded bans of hate speech may be used to target dissenting views and the very groups such speech restrictions are supposed to protect.

This risk is heightened by the fact that the removal rate of hate speech online is rising. For example, in the first quarter of 2018, Facebook removed 2.5 million pieces of content for violating its community standards on hate speech. This increased to 4.1 million pieces of content in the first quarter of 2019 and 9.6 million in the first quarter of 2020. In the second quarter of 2020, more than 20 million pieces of content were deleted for violating Facebook's hate speech ban, 22.1 million pieces in the third quarter of 2020 and 26.9 million in the fourth quarter. There was a slight drop in the first quarter of 2021 – 25.2 million pieces of content were removed with a notable increase of 31.5 million pieces in Q2.⁵¹ Due to its broad policies on removing hate speech and the use of algorithms (which cannot pick up on the nuance of language⁵²) to enforce these policies, the result is often the silencing of communities which hate speech bans purport to protect. For example, LGBT groups and anti-racism activists have reported that Facebook's algorithms have flagged words such as 'tramp' and 'fag' that activists have reclaimed to 'cope with hostility'⁵³ and have censored posts

⁴⁶ Evelyn Doyek, 'Facebook's Role in the Genocide in Myanmar: New Reporting Complicates the Narrative' (2018) *Lawfare*

⁴⁷ Alexandra A. Siegel, 'Online Hate Speech' in Nathaniel Persily & Joshua A. Tucker *Social Media and Democracy - The State of the Field, Prospects for Reform* (2020),

<https://alexandra-siegel.com/wp-content/uploads/2018/09/Siegel_Online_Hate_Speech.pdf>

⁴⁸ Phyllis B. Gerstenfeld, 'Hate Crimes: Causes, Controls, and Controversies' (1st ed. 2017 Sage).

⁴⁹ Billy Henson, Bradford W. Reynolds, Bonnie S. Fisher, 'Fear of Crime Online? Examining the Effect of Risk, Previous Victimization, and Exposure on Fear of Online Interpersonal Victimization' (2013) *Journal of Contemporary Criminal Justice*,

<<https://journals.sagepub.com/doi/abs/10.1177/1043986213507403>>

⁵⁰ Karsten Muller & Carlo Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime' (2017),

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972>; Karsten Muller & Carlo Schwarz, 'Making America Hate again? Twitter and Hate Crime under Trump' <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103>

⁵¹ Please refer to respective enforcement reports for more details.

⁵² Thiago Dias Oliva, Dennys Marcelo Antonialli & Allesandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 *Sexuality & Culture*, 701

⁵³ Thiago Dias Oliva, Dennys Marcelo Antonialli & Allesandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) 25 *Sexuality & Culture*, 702



talking about racial discrimination.⁵⁴ This issue is not particular to Facebook, research found that African American English tweets are twice as likely to be considered offensive compared to others, reflecting the infiltration of racial biases in technology.⁵⁵ YouTube removed 6,000 videos documenting the Syrian conflict. For example, it shut down Qasioun News Agency, an independent media group reporting on war crimes in Syria.⁵⁶ Another problematic possibility with algorithm use must also be taken into account. Facebook whistleblower Frances Haugen has recently alleged that the platform encourages online hate and extremism, fails to protect children from harmful content and lacks any incentive to tackle the problem. She said that engagement algorithms take people with mainstream interests to the extremes.⁵⁷

It could be argued that the increase in removal denoted above with the example of Facebook is positively correlated – at least, in part -- to national developments such as the NetzDG but also developments at regional levels. For example, in 2016, the EU Code of Conduct on Countering Illegal Hate Speech Online, a voluntary agreement, was signed as between the Commission and several social media platforms. The proposed Digital Services Act (DSA) aims at creating a coordinated response to such issues as the removal of illegal online content, which includes “hate speech ... and unlawful discriminatory content.”

The enhanced focus on regulating hate speech does not necessarily reflect the pervasiveness of hate speech. In fact, contrary to popular conception, hate speech seems to constitute a relatively small share of the content on social media platforms. For example, a 2019 study led by Siegel looked at whether Trump’s 2016 election campaign and its immediate aftermath (6 months) contributed to the rise in hate speech or white nationalist language. The study analyzed 1.2 billion tweets, 750 million of which were election-related, and nearly 400 million were random samples. It found that, on any given day, between 0.001% and 0.003% of the tweets contained hate speech, “a tiny fraction of both political language and general content produced by American Twitter users”. There is also an extensive Danish study that shows consistent data. This study looked at deleted comments on the Facebook pages of five Danish media outlets and found that 6.2% of the comments posted were deleted, but only 1.1% of these comments violated provisions in the Danish Criminal Code on hate speech. So, for each time a comment was removed for being in violation of the Criminal Code, 36 non-hateful and non-offensive comments on issues such as politics were removed on other grounds.⁵⁸

⁵⁴ Jessica Guynn, ‘Facebook While Black: Users call it getting ‘Zucked,’ Say Talking about Racism is Censored as Hate Speech,’ *USA Today*, <<https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/>>; ‘Facebook’s Hate Speech Policies Censor Marginalized Users’ *Wired*, <<https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/>>

⁵⁵ <https://www.semanticscholar.org/paper/Fighting-Hate-Speech%2C-Silencing-Drag-Queens-in-and-Oliva-Antoniali/954f95b6cc8b447a6bd9c42183e689f65a85897b>

⁵⁶ <https://globalfreedomofexpression.columbia.edu/publications/new-report-from-justitia-digital-freedom-of-expression-and-social-media/>

⁵⁷ Kevin Chan, ‘Facebook Whistleblower says Platforms Amplifies Onlien Hate, Extremism.’ (25 October 2021) <<https://globalnews.ca/news/8322712/facebook-whistleblower-frances-haugen-uk-testimony/>>

⁵⁸ Jacob Mchangama, Eske Vinther-Jensen & Brandt Taarnborg, ‘Digital Freedom of Speech and Social Media’ (2020) *Justitia*, <<https://justitia-int.org/en/new-report-digital-freedom-of-speech-and-social-media/>>



Article 20(2) ICCPR - Prohibition of Hate Speech

Article 20(2) of the ICCPR states that “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.” General Comment 34 notes that “Articles 19 and 20 are compatible with and complement each other. The acts that are addressed in Article 20 are all subject to restriction pursuant to Article 19, paragraph 3”. Problematic is the variation in thresholds for the relationship between protected and prohibited speech -- particularly, when it comes to hate speech. Specifically, while Article 20(2) prohibits advocacy of hatred that constitutes incitement, Article 4 broadly prohibits racist expression with no requirement that this speech reach any particular threshold (such as incitement). Given the premise of this report (namely, that it is imperative to safeguard the right to freedom of expression online), the enabling role this right has with respect to other rights and given the latest developments in the field of content moderation field (such as pressure by the state to remove content and a significant rise in content purging), we argue that the better approach to hate speech and disinformation online should emanate from Articles 19 and/or Article 20(2) of the ICCPR (depending on the situation) since the Covenant’s relevant provisions endorse a speech protective approach to expression in an already vulnerable environment. Thus, references to IHRL in this report refer to Article 19 and/or Article 20(2) ICCPR.

The Rabat Plan of Action (RPA), which assists in the interpretation and application of Article 20(2) and is discussed in more detail below, stresses that Article 20(2) ICCPR needs to be read in light of Article 19.⁵⁹ The compatibility between Articles 19 and 20(2) has also been emphasized in reports of the SRFOE.⁶⁰

The HRC’s General Comments 11 and 34 and the RPA are particularly instructive for interpreting Article 20(2). The SRFOE’s report entitled “Online Hate Speech”, and the 2020 UN Strategy and Plan of Action on Hate Speech also provide important content.

The 2012 report of the SRFOE⁶¹ defined advocacy as “explicit, intentional, public and active support and promotion of hatred towards the target group”; hatred is defined as “a state of mind characterized as intense and irrational emotions of opprobrium, enmity and detestation towards the target group”; and incitement is defined as “statements about national, racial or religious groups that create an imminent risk of discrimination, hostility or violence against persons belonging to those groups”. The notion of “discrimination” is given a particularly broad meaning as:

⁵⁹ HRC General Comment 34: ‘Article 19 – ‘Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34, para. 11.

⁶⁰ See, for example, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, ‘Online Hate Speech’ (9 October 2019) A/74/486, para.12.

⁶¹ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2012) A/67/357, para. 44.

“any distinction, exclusion or restriction made on the basis of race, colour, descent, national or ethnic origin, nationality, gender, sexual orientation, language, religion, political or other opinion, age, economic position, property, marital status, disability, or any other status that has the effect or purpose of impairing or nullifying the recognition, enjoyment or exercise, on an equal footing, of all human rights and fundamental freedoms in the political, economic, social, cultural, civil or any other field of public life”.

Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2012)

Hostility is considered to be a “manifestation of hatred beyond a mere state of mind”, but the report recognizes that this theme has received little attention in relevant case law and needs to be considered further. Lastly, “violence” is defined as the use of physical force or power against another person or against a group or community, which either results or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation.

The HRC noted that “it is with regard to the specific forms of expression indicated in Article 20 that States Parties are obliged to have legal prohibitions”.⁶² In *Ross v Canada* (2000), the HRC said that “restrictions on expression which may fall within the scope of Article 20 must also be permissible under Article 19, paragraph 3”.⁶³ This line of reasoning was followed in the more recent case of *Mohammed Rabbae, A.B.S and N.A v. The Netherlands* (2016) in which the HRC underscored that Article 20(2) should be read in light of Article 19.⁶⁴

Article 20(2) - Threshold

The 2012 SRFOE report held that “the threshold of the types of expression that would fall under the provisions of Article 20(2) should be high and solid”.⁶⁵ The HRC has discussed the threshold of restrictions of expression under Article 20(2) in two significant cases. In *Ross v Canada* (2000), the HRC noted that “restrictions on expression which may fall within the scope of Article 20 must also be permissible under Article 19, paragraph 3”.⁶⁶ Nevertheless, the HRC found that the applicant’s transfer to a non-teaching post because of his anti-Semitic publications did not violate Article 19.

In *Mohamed Rabbae, A.B.S and N.A v The Netherlands* (2016), the applicants claimed to be victims of a violation of their rights under Article 20(2). They argued that statements made by Geert Wilders, leader of the anti-Islamic Dutch Freedom Party and, specifically, his acquittal by the domestic court,

⁶² HRC General Comment 34, ‘Article 19 – Freedom of Opinion and Expression’ (2011) CCPR/C/GC/34, para.55.

⁶³ *Ross v Canada* Communication no 736/1997 (18 October 2000) CCPR/C/70/D/736/1997, para. 10.6.

⁶⁴ *Mohamed Rabbae, A.B.S and N.A v The Netherlands* (18 November 2016) CCPR/C/117/D/2124/2011, para. 9.7.

⁶⁵ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/67/357 (7 September 2012), para. 45.

⁶⁶ *Ross v Canada* Communication no 736/1997 (18 October 2000) CCPR/C/70/D/736/1997, para. 10.6.



was in contravention of Article 20(2). This was the first time that the HRC engaged in a relatively extensive analysis of Article 20(2). It held that this article secures the right of persons to be free from hatred and discrimination but that it is “crafted narrowly” to ensure protection of free speech. It recalled that free speech may incorporate “deeply offensive” speech and speech that is disrespectful to a religion unless the strict threshold of Article 20(2) is met.⁶⁷ In their concurring individual opinion, Cleveland and Politi discussed the structure of Article 20(2), which prohibits only advocacy of hatred that actually constitutes incitement to discrimination, hostility and violence. They also noted, among other issues, that overbroad restrictions may be abused to censor and silence dissenting voices and suppress the very minorities they are designed to protect.⁶⁸ The importance of allowing offensive speech, as set out in the above case, is of particular relevance to the online environment. What is (or is not) offensive is deeply subjective. This was reflected in Justitia’s 2021 global survey on attitudes towards free speech, which found that there is no universal agreement on whether statements offensive to religions or minorities should be tolerated. In Scandinavia and the US, around 65% of the population believe that free speech should extend to statements offensive to minority groups while around 80% think that statements offensive to religion should be allowed; whereas, in Kenya, Indonesia, Turkey and Tunisia, only between 18% and 27% of the population favor tolerating statements offensive to minorities, and 26-29% of the population favor tolerating expressions offensive to religion.⁶⁹ As such, for global platforms with users from a multitude of ethnic, religious and other backgrounds, it has become impossible to ensure that no one is offended while at the same time protecting the core of free speech. Thus, prohibitable speech online must meet a certain threshold of harm.

It is noteworthy that the RPA (2013) and General Comment 34 (2011) seem to have strengthened Article 19’s protection of speech vis-à-vis Article 20(2). This was manifested in the different outcomes of the two cases above, with *Rabbae* being more speech protective than *Ross*.

The RPA aims at clarifying the threshold of Article 20(2) ICCPR and sets a high bar for legitimate restriction of expression. The RPA understands that to assess the severity of the hatred and, therefore, determine whether the high threshold is met, potential issues to be considered are “the cruelty of what is said or of the harm advocated and the frequency, amount and extent of the communications”. The RPA⁷⁰ includes a six-part threshold test to be used when applying Article 20(2), which incorporates a consideration of:

⁶⁷ Mohamed Rabbae, A.B.S and N.A v The Netherlands, Communication no. 2124/2011 (14 July 2016) CCPR/C/117/D/2124/2011, para. 10(4).

⁶⁸ Individual Opinion (concurring) of Committee Members Sarah Cleveland and Mauro Politi in Mohamed Rabbae, A.B.S and N.A v The Netherlands, Communication no. 2124/2011 (14 July 2016) CCPR/C/117/D/2124/2011, para. 8.

⁶⁹ Svend-Erik Skaaning & Suthan Krishnarajan, ‘Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech (2021) *Justitia*, <https://futurefreespeech.com/wp-content/uploads/2021/06/Report_Who-cares-about-free-speech_21052021.pdf>

⁷⁰ Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22.



The social and political context



Status of the speaker



Intent to incite the audience against a target group



Content and form of the speech



Extent of its dissemination and



Likelihood of harm, including imminence

South African courts provide good examples of how freedom of expression and hate speech limitations can be reconciled in a manner that limits the risk of abuse while they explicitly reference IHRL more generally and the ICCPR in particular.⁷¹ These lines of reasoning might be helpful for stakeholders, including social media platforms, in deciphering the limits of free speech.

Section 16(2) of the South African Constitution stipulates that freedom of expression does not extend to “advocacy of hatred that is based on race, ethnicity, gender or religion, and that constitutes incitement to cause harm.” Its structure and requirements are thus relatively similar to Article 20(2) ICCPR, and this is also reflected in case law.

In *Islamic Unity Convention v Independent Broadcasting Authority and Others* (2002), which involved Holocaust denial, the Constitutional Court found that an administrative clause prohibiting prejudice (and used against the impugned speech) did not meet the constitutional threshold. The Court underlined that “individuals in our society need to be able to hear, form and express opinions and views freely on a wide range of matters....” The Court also underlined that “not every expression of speech that is likely to prejudice relations between sections of the population would constitute ‘advocacy of hatred’ which also constitutes ‘incitement to cause harm’”.⁷²

In *Economic Freedom Fighters* (2020), a case involving incitement, the Constitutional Court demonstrated its extensive understanding of freedom of expression, highlighting that it is “the lifeblood of constitutional democracy” and that “[w]hen citizens are very angry or frustrated, it serves

⁷¹ *Qwelane v South African Human Rights Commission and Another* (CCT 13/20) [2021] ZACC 22 (31 July 2021) para. 87, *Islamic Unity Convention v Independent Broadcasting Authority and Others* (CCT36/01) [2002] ZACC 3; 2002 (4) SA 294; 2002 (5) BCLR 433 (11 April 2002) para. 28.

⁷² *Islamic Unity Convention v Independent Broadcasting Authority and Others* (CCT36/01) [2002] ZACC 3; 2002 (4) SA 294; 2002 (5) BCLR 433 (11 April 2002).



as the virtual exhaust pipe through which even the most venomous of toxicities within may be let out to help them calm down, heal, focus and move on".⁷³

In *Moyo v Minister of Justice and Constitutional Development and Sonti v Minister of Justice and Correctional Services and Others* (2018), which involved the use of allegedly threatening and violent language, the Supreme Court of Appeal noted that, "unless hate speech, incitement to imminent violence or propaganda for war as proscribed in...the constitution are involved, no one is entitled to be insulated from opinions and ideas that they do not like even if those ideas are expressed in ways that place them in fear..."⁷⁴

This delineation was followed by the Supreme Court of Appeal in *Qwelane v South African Human Rights Commission* (2019), which involved homophobic speech.⁷⁵ The Court found that the Equality Act, which prohibited hurtful, hateful and harmful speech, was not in line with Section 16(2) of the Constitution. The Constitutional Court reiterated the Supreme Court of Appeal position that "[t]he fact that political expression may be hurtful of people's feelings or wounding, distasteful, politically inflammatory or downright offensive, does not exclude it from protection." However, the Constitutional Court reversed the previous judgment by finding that only the inclusion of the term "hurtful" was unconstitutional.⁷⁶ Despite this deviation from the tendency to err on the side of speech protectiveness, South African top courts still stand out for their willingness to protect even deeply offensive and hurtful speech.

⁷³ Economic Freedom Fighters, Julios Selo Malema v Minister of Justice and Correctional Services, National Director of Public Prosecutions, Case CCT 201/19, para.1.

⁷⁴ *Moyo v Minister of Justice and Constitutional Development and Others; Sonti v Minister of Justice and Correctional Services and Others*, Cases 287/2017; 286/2017, para. 31.

⁷⁵ *Qwelane v South African Human Rights Commission and Another* (686/2018) [2019] ZASCA 167; [2020] 1 All SA 325 (SCA); 2020 (2) SA 124 (SCA); 2020 (3) BCLR 334 (SCA) (29 November 2019).

⁷⁶ *Qwelane v South African Human Rights Commission and Another* (CCT 13/20) [2021] ZACC 22 (31 July 2021).

Summary: Key points regarding restriction of freedom of expression on the grounds of hate speech under Article 19:

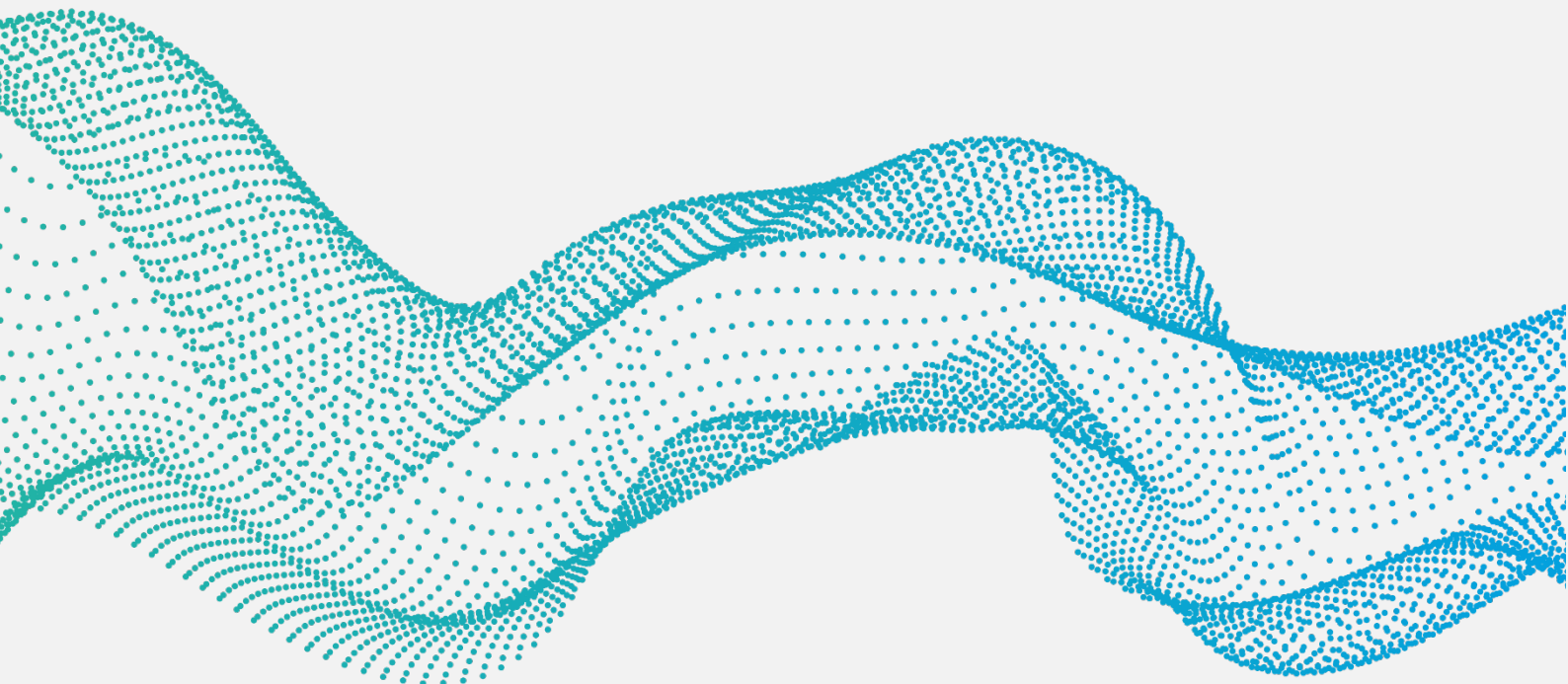
- The restrictions must be necessary, legitimate, and proportional and must occur within the framework of one of the grounds set forth in Article 19.
- The restriction must be the least intrusive measure.
- The right to freedom of expression allows “even expression that may be regarded as deeply offensive” as well as hate speech-adjacent categories such as denial of historical facts and religious insult/blasphemy.

Key points to note regarding Article 20(2):

- The threshold must be high.
- Article 20(2) is compatible with Article 19.
- Reference to good practices such as South African case law can assist in decoding the threshold element

Important documents:

- HRC General Comment 11: ‘Article 20 Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred’ (1983).
- HRC General Comment 34: ‘Article 19 – Freedom of Opinion and Expression’.
- Report of the SRFOE, ‘Online Hate Speech’.
- The RPA.
- The 2020 UN Strategy and Plan of Action on Hate Speech.





Platform regulation of hate speech:

From platform policies and/or our interviews with representatives, it is clear that several companies such as Facebook,⁷⁷ Instagram,⁷⁸ YouTube, Reddit and Twitter⁷⁹ take IHRL into account – at least, theoretically. Other platforms such as VERO said that, although they did not rely on IHRL initially when they devised their community standards and based their policies on national laws in the US, they might be open to revising their standards in accordance with IHRL in the future.

However, IHRL may not be suitable to all platforms. For instance, LinkedIn caters mostly to professionals for career development. General debate and politics have played a comparatively smaller role on this platform. So, IHRL may be less meaningful for determining the appropriate limits of speech. Likewise, IHRL is less relevant to platforms such as Snapchat or WhatsApp, which are primarily messaging applications between private parties without a public news feed on which photos that have been end-to-end encrypted are shared between private users. Furthermore, an IHRL approach might be burdensome for smaller platforms whose business models rely on user-generated content since they do not have the resources needed to mainstream IHRL systematically in their automated and human content moderation. The EU's DSA, for example, differentiates between two types of online platforms – online platforms and very large online platforms (over 45 million users) with higher standards of transparency and accountability on the latter.

Determining the exact criterion to determine which platforms an IHRL approach is relevant is thus a crucial question to be answered, though not one that will be pursued further in this report.



Facebook⁸⁰ and **Instagram**⁸¹ (both owned by Facebook, Inc.) recognize three levels of hate speech. They formulate their understanding of hate speech around the notion of a 'direct attack' against people based on a broad list of protected characteristics such as race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. Beyond those characteristics, age and occupation are deemed to be protected, not as standalone, but when accompanied by another of the above characteristics. Their 'Community Standards' note that "we also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies." In an interview we conducted with Facebook representatives for purposes of this report, we

⁷⁷ Facebook: Commitment to Human Rights, <<https://about.fb.com/news/2021/03/our-commitment-to-human-rights/>>

⁷⁸ Instagram Community Guidelines, <<https://help.instagram.com/477434105621119>>

⁷⁹ Twitter: Defending and respecting the rights of people using our service, <<https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice>>

⁸⁰ Facebook Community Standards: Hate Speech, <https://www.facebook.com/communitystandards/hate_speech>

⁸¹ Instagram: An update on our work to tackle abuse, <<https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram>>

were told that “there is an effort to make sure that the human rights voice is there, that international human rights standards, the Rabat principles are all part of defining and enforcing our content policies.”⁸²

The Community Standards define a direct attack broadly as “violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes...” While the manner in which Facebook conceptualizes a direct attack is at least partially in line with Article 20(2) in that it incorporates violent speech, the fact that it extends hate speech to lower thresholds such as harmful stereotypes, fall outside the spectrum of IHRL.

Another issue of potential concern is the expansion of protected characteristics - most recently, to include occupation – and the incorporation of additional grounds such as age. These are much more fluid categories than race and ethnicity and do not, to the same extent, reflect underlying historical and systemic harms such as genocide, crimes against humanity or violent hate crimes that have frequently been committed against racial, ethnic, religious and sexual minorities, with hate speech fanning the flames. In addition, the inclusion of more categories increases the risk of arbitrary content moderation. Moreover, there is the question of whether the newly-included grounds may be said to function in a non-discriminatory and non-hierarchical manner.

Facebook and Instagram argue that they allow hate speech to be shared in order to condemn it or to raise awareness. This is a safety net for those who wish to be part of a public discourse on tackling hate speech without or with less fear of restriction, an element that is not recognized on all platforms and which is welcomed since it allows for crucial context that is part and parcel of the determination of whether speech falls within or outside the scope of hate speech under IHRL. YouTube, for example, only extends this safety net to educational material such as videos on hate groups.



Reddit⁸³ prohibits incitement to violence and the promotion of hatred based on protected characteristics, demonstrating a comparable approach and threshold to Article 20(2). It also stipulates that “it does not protect those who try to hide their hate in bad faith claims of discrimination”. This prong of the policy is rather abstract and unclear and does not meet the threshold of incitement found in Article 20(2). The characteristics covered by its ‘Content Policy’ are relatively broad and include (but, as stipulated in the policy, are not limited to) “actual or perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy or disability. These include victims of a major violent event and their family”.

⁸² Interview conducted with Facebook representatives for purposes of this report.

⁸³ Reddit: Promoting Hate based on Identity or Vulnerability, <<https://www.reddithelp.com/hc/en-us/articles/360045715951>>



YouTube⁸⁴ prohibits content promoting violence or hatred against individuals or groups based on age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin and veteran status. This aligns with Article 20(2)

in that the element of the promotion of violence or hatred is incorporated. However, YouTube's Hate Speech Policy also refers to "other types of content that violates this policy", and this includes content such as slurs and stereotypes, adopting a lower threshold than the standard required by Article 20(2). Yet, during our interview with a YouTube representative conducted for purpose of this report, we were informed that IHRL is "at the center of how we deal with hate speech, we discuss Article 19 and Article 20. Guidelines of the UN guide our methodology when we review and update policies."⁸⁵



Twitter⁸⁶ prohibits the promotion of violence or direct attacks or threats against others based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease. Its policy outlines the type of content on which it will take action, which includes incitement of violence, fear, harassment or discrimination against certain characteristics as well as

repeated and/or non-consensual slurs, epithets, racist and sexist tropes or other content that degrades a person. Twitter's threshold is arguably lower than that provided for in Article 20(2) since, in addition to prohibiting the promotion of violence, it also prohibits slurs and epithets.



TikTok⁸⁷ prohibits content that contains hate speech or involves "hateful behavior" and removes such content from the platform. TikTok defines hate speech or hateful behavior as content that attacks, threatens, incites violence against or otherwise dehumanizes an individual or a group on the basis of a list of protected attributes.

While the references to attacks and incitement reflect the Article 20(2) threshold, dehumanization may fall within a framework less protective of speech. Moreover, TikTok's Community Guidelines refer to "slurs" as terms that are intended to disparage an ethnicity, race, or any other protected attribute. TikTok states that "to minimize the spread of egregiously offensive terms", it removes all slurs (unless used self-referentially or do not disparage). Extending removal to merely offensive terms reflects a lower level of speech protection than required under IHRL.

The overview of platforms' Terms on hate speech reflects a need for them to conform to IHRL. While some elements meet IHRL thresholds, other elements, such as the low threshold of harm associated with speech (e.g., stereotypes and slurs) and the expansion of protected characteristics to include more fluid notions, fall outside the scope of Article 19(2) and are in contravention of Article 20(2) as elaborated by the RPA.

⁸⁴ YouTube Hate Speech Policy, <<https://support.google.com/youtube/answer/2801939?hl=en>>

⁸⁵ Interview with Google conducted for purposes of this report.

⁸⁶ Twitter Hateful Conduct Policy, <<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>>

⁸⁷ TikTok Community Guidelines, <<https://www.tiktok.com/community-guidelines?lang=en#38>>

Hate speech policies – Thresholds of prohibited speech

This is a non-exhaustive list of the different thresholds of prohibited speech across company policies going from the most severe (violence) to the least severe (stereotypes). The words chosen are umbrella terms to cover the broad range of terms used across platforms.

	 Facebook	 Instagram	 Reddit	 TikTok	 Twitter	 VERO	 YouTube
Violence⁸⁸	✓	✓	✓	✓	✓	✓	✓
Attack	✓	✓	✓	✓	✓		
Threat				✓	✓		✓
Abuse/harm					✓	✓	
Intimidation/ fear	* 89	* 90			✓		
Dehumanization	✓	✓		✓	✓		✓
Disgust	✓	✓					
Contempt	✓	✓					✓
Exclusion/ segregation/ discrimination	✓	✓	✓	✓	✓	✓	✓
Humiliation/ shaming						✓	
Cursing/slurs	✓	✓		✓	✓		✓
Stereotypes	✓	✓			✓		✓

⁸⁸ Note: The word 'hate' is used interchangeably across platforms either in headings or in the text of hate speech terms.

⁸⁹ Used only to define harmful stereotypes: "dehumanizing comparisons that have historically been used to attack, intimidate or exclude specific groups."

⁹⁰ Used only to define harmful stereotypes: "dehumanizing comparisons that have historically been used to attack, intimidate or exclude specific groups."

Case Studies

In this section, we apply relevant IHRL norms and principles to actual content that has appeared on and, in some cases, been removed by social media platforms. The public does not have direct access to purged content, so we relied on content reviewed by the Facebook Oversight Board and cases that have been reported by the media. We recommend that in order to promote transparency about the decisions made by social media companies on such content, the decisions (and the impugned content in anonymized format) should be made available to the public (or, at least, to relevant stakeholders on request). Without this information, discussion and additional investigation for the improvement of the current system of content moderation become impossible. How legitimate restrictions to Article 19 can be operationalized is demonstrated in Chapter 1. All case studies involve an application of the RPA test, which examines the context of the post, the status of the speaker, the speaker's intent, the content and form of the post, the extent of the speech act and the likelihood of harm.

Uyghur Muslims in China – Facebook

On 29 October 2020, a user in Myanmar posted in a Facebook group in Burmese. The post included two widely-shared photographs of a Syrian toddler of Kurdish ethnicity who drowned attempting to reach Europe in September 2015. The accompanying text inserted by the user stated that “there is something wrong with Muslims (or Muslim men) psychologically or with their mindset”. It questioned the lack of response by Muslims to the treatment of Uyghur Muslims in China compared to killings in response to cartoon depictions of the Prophet Muhammad in France. The user concluded that recent events in France reduce his/her sympathies for the child and implied that the child may have grown up to be an extremist.

Facebook removed this content because it contained the phrase “[there is] something wrong with Muslims psychologically” as in contravention of its Community Standards on hate speech, which prohibit generalized statements of inferiority about the mental deficiencies of a group on the basis of their religion.

Analysis of this decision under IHRL

The RPA test:



CONTEXT

The context of the post shows that the user was not advocating discrimination, hate or violence but, instead, sought to show the (alleged) indifference of Muslims towards the situation in China. This can be seen by:

- The use of the viral photo of the dead child on the beach, which evoked emotions of empathy and anger at systemic disfunction;
- The reference to the refugee crisis, which links the post to humanitarian issues; and
- The comparison between the killings in France following the publication of the cartoons and the Uyghur Muslims' situation in China.



SPEAKER

In this case, the user was not a public figure and his/her readership had no significant impact.



INTENT

Given the context, it can be reasonably assumed that the user's intent was to raise awareness of the plight of Uyghur Muslims in China and the alleged disinterest of the Muslim community, using strong pictures and words to enhance the user's argument.



CONTENT AND FORM

While provocative and offensive (to some), the content of the post was political speech and its aim was not to promote hostility but, instead, to raise awareness of the Uyghur Muslims' current plight in China through a comparison with the responses to cartoons of the Prophet Muhammad in France. Political speech enjoys a particularly high degree of protection under IHRL because of its importance to public debate and may only be limited in strict circumstances.



EXTENT OF THE SPEECH ACT

The post did not incite any discrimination, hatred or violence for the reasons explained above; and, given the low profile of the user, the post would have obtained a wave of reception analogous to the number of friends that the user had. (The exact data is not available).



LIKELIHOOD

The post did not aim to incite any particular harm such as violence, nor could it reasonably have led to any tangible harm.

This analysis demonstrates that the post does not fall within the ambit of Article 20(2), nor does it meet the threshold test laid out by the RPA. Such analysis is in line with the decision of Facebook's Oversight Board, which relied on IHRL (including the RPA) in holding that the post did not fall within the framework of Article 20(2) ICCPR. The Oversight Board also applied the test under Article 19(3), concluding that the post's removal was not necessary to protect the rights of others since it did not entail a threat or identifiable individuals. Overall, it noted that:

“considering international human rights standards on limiting freedom of expression, the Board found that, while the post might be considered pejorative or offensive towards Muslims, it did not advocate hatred or intentionally incite any form of imminent harm. As such, the Board does not consider its removal to be necessary to protect the rights of others.”⁹¹

The use of IHRL by the Oversight Board in this case is to be welcomed. However, in a previous case involving the Dutch “Black Pete” tradition, the Board adopted a lower threshold of harm than in the Uyghur Muslim case, while still relying on IHRL. In Dutch tradition, “Black Pete” is Saint Nicholas's helper. He is depicted as a blackface character wearing a wig, Renaissance clothing, large gold earrings and red lipstick. It is a controversial topic due to the racial connotations and links with the country's history with the transatlantic slave trade.⁹² The case referred to the Oversight Board involved the posting of a video (17 seconds long) which showed a young child meeting three adults – one dressed as Saint Nicholas and two dressed as Black Pete. The faces of those dressed as Black Pete were in blackface and they wore the traditional attire. Festive music was playing and one of the Black Petes said to the child, “[l]ook here, and I found your hat. Do you want to put it on? You'll look like an actual Pete!” Facebook removed the video for violating its hate speech standards.

In its decision, the Oversight Board referred to issues of stereotypes and structural racism and relied on findings such as that of the Committee on the Elimination of All Forms of Racial Discrimination (CERD) that Black Pete “is experienced by many people of African descent as a vestige of slavery”.⁹³ The Oversight Board found that “allowing such posts to accumulate on Facebook would help create a discriminatory environment for Black people that would be degrading and harassing ... [and] that the impacts of blackface justified Facebook's policy and that removing the content was consistent with the company's human rights responsibilities.”

We disagree with this interpretation and argue that “Black Pete” (and his depiction in the post), however offensive and insensitive to black people in the Netherlands (and elsewhere), meets neither

⁹¹ Oversight Board Case Decision 2020-002-FB-UA, <<https://oversightboard.com/news/773985406543781-oversight-board-overturns-facebook-decision-case-2020-002-fb-ua/>>

⁹² Becky Little, 'This Notorious Christmas Character is Dividing a Country' (2018), <<https://www.nationalgeographic.com/history/article/black-pete-christmas-zwarte-piet-dutch>>

⁹³ CERD Concluding Observations (Netherlands) (28 August 2015) CERD/C/NLD/CO/19-21 para. 17.

the requirement of incitement to tangible harm in the form of discrimination, hostility or violence nor the requirement of imminence.⁹⁴

This decision highlights the problem with implementing competing IHRL standards, given that Article 4(a) of the ICERD adopts a much lower threshold of protection than Article 20(2) by requiring states parties to “declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred [and] incitement to racial discrimination....” The Oversight Board relied heavily on the ICERD. We maintain that the HRC has reiterated the acceptability of “deeply offensive speech”, which seems difficult to reconcile with the Board’s decision in the Black Pete case. In this context, it may be relevant to reiterate the position of the South African Constitutional Court that “not every expression of speech that is likely to prejudice relations between sections of the population would constitute advocacy of hatred which also constitutes ‘incitement to cause harm’”.⁹⁵

Assam – Facebook

In 2018, Jitten Dutta, a leader of the United Liberation Front of Asom (Pro Talks division) - a separatist organization in Assam, a state in Northeast India, shared a post on his Facebook account. In Assam, Muslims have been at risk of discrimination, violence and hatred due to the increase in its migration population from neighboring Bangladesh.

Dutta wrote that “strict laws” are needed to stop the “population explosion happening in sandbar areas” because “otherwise our nationality will not survive. Along with Hindu Bangladeshis, this aspect also needs to be monitored.”

The posts triggered comments such as:

“Sir, you break the pause regarding war. Will kill or die; 1000 youths are ready. Please, Sir. Do something”.

“There is no alternative but to take arms. Within night, it has to be over. Assamese people can be sold with some money; that is the fear, Sir”.

“It will be okay to burn the houses of the people near the river even now”.⁹⁶

Facebook removed Dutta’s primary and duplicate profiles after being alerted to the content above by Avaaz, an International NGO that is also active in Assam.⁹⁷

Analysis of this decision under IHRL

We can apply the threshold test under the RPA to this case:

⁹⁴ Oversight Board: Case Decision 2021-002-FB-UA, <<https://oversightboard.com/decision/FB-S6NRTDAJ/>>

⁹⁵ *Qwelane v South African Human Rights Commission and Another* (686/2018) [2019] ZASCA 167; [2020] 1 All SA 325 (SCA); 2020 (2) SA 124 (SCA); 2020 (3) BCLR 334 (SCA) (29 November 2019).

⁹⁶ Megaphone for Hate, Disinformation and Hate Speech on Facebook during Assam Citizenship Count (2019) *Avaaz*, <[https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20\(1\).pdf](https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf)>

⁹⁷ *Avaaz*, <<https://secure.avaaz.org/page/en/about/>>



CONTEXT

The post was made against the backdrop of a process updating the National Register of Citizens in Assam through which the Indian government sought to identify and expel illegal immigrants. The UN has warned that, within this context, there was a “rise of hate speech directed against these minorities in social media” in Assam.⁹⁸ A report issued by the International NGO Avaaz noted that Muslims are “facing an extraordinary chorus of abuse and hate in Assam on Facebook”.⁹⁹ Dutta’s post must be considered within that framework as well as in the context of physical violence perpetrated against Muslims in Assam. Although exact numbers about violence are disputed, violent attacks have occurred over the years, such as the infamous Nellie massacre in 1983 in which 1,800 Muslims were killed in six hours and the 2014 violence in which homes of Muslim families were burned and over 30 persons killed.¹⁰⁰



SPEAKER

Dutta was a public figure with a strong Facebook presence at the time (including multiple profiles) in a region where Muslims are vulnerable to exclusion and abuse.



INTENT

Since 2007, Dutta had been calling for actions against “foreigners” and was vehemently against the naturalization of migrants from Bangladesh. In May 2018, he had been charged with sedition for inflammatory statements – some of which seemed to be aimed at inciting violence. For example, he said that “the youths of Assam would take to arms and join the ULFA to realize the objective even at the cost of bloodshed”.¹⁰¹ He had also spoken out against Muslims from the neighboring Indian state of West Bengal.¹⁰²



The post is direct and highly inflammatory since it uses a hyperbolic euphemism such as “population explosion” – cultivating the sentiment of fear in the other communities that they may be “overtaken” by Bengali Muslims and Bangladeshi

⁹⁸ ECOI.NET, ‘UN experts: Risk of Statelessness for Millions and Instability in Assam, India’ (2019), <<https://www.ecoi.net/en/document/2012494.html>>

⁹⁹ Megaphone for Hate, Disinformation and Hate Speech on Facebook during Assam Citizenship Count (2019) *Avaaz*, <[https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20\(1\).pdf](https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf)>

¹⁰⁰ ‘Dozens of Muslims Killed in Ethnic Violence in North-East India’ (2014) *The Guardian*, <<https://www.theguardian.com/world/2014/may/03/dozens-muslims-killed-ethnic-violence-north-east-india-assam>>

¹⁰¹ Laxman Sharma, ‘Sedition Charges Against Pro-Talk ULFA Leader Jiten Dutta’ (2018), <<https://nenow.in/north-east-news/sedition-charges-against-pro-talk-ulfa-leader-jiten-dutta.html>> *NorthEast Now*.

¹⁰² Shoaib Daniyal & Arunabh Saikia, ‘Assam Killings: Bengali Groups Blame BJP Government for Escalating Ethnic Tensions over NRC’ (2018), *Scroll.In* <<https://scroll.in/article/900707/assam-killings-bengali-groups-blame-bjp-government-for-escalating-ethnic-tensions-over-nrc>>

CONTENT AND FORM

migrants. The post constitutes incitement to hostility against Muslims, even if the statements did not incite violence.



EXTENT OF THE SPEECH ACT

Dutta's public post had a significant reach. He is a public figure with at least 4 pages and over 12,900 followers and at least 7 profiles on Facebook.¹⁰³ As reflected in subsequent atrocities committed against Bengalis (which are by no means solely linked to this post), the community that Dutta aimed to incite did have the means to act against the targeted community.



LIKELIHOOD

Given the context of the volatile situation in Assam, previous instances of mass killings, and particularly the vulnerable position of Bengalis during the National Register of Citizens which aimed at determining "illegal" Bangladeshi migrants, the likelihood and imminence of harm emanating from Dutta's speech were real, even in the absence of explicit calls for violence.

In light of the above, the user called for stopping a "population explosion" which, taking into account the particular context of the Assam region and the vulnerability of Muslims, amounted to advocacy of hatred and an incitement to (at the very least) discrimination or hostility. The user did not directly call for violence, but he did refer to "strict laws" to halt a "population explosion". It is relevant that the post subsequently triggered comments that promoted hatred and violence. This case, therefore, meets the thresholds set out in Article 20(2).

¹⁰³ Megaphone for Hate, Disinformation and Hate Speech on Facebook during Assam Citizenship Count (2019), *Avaaz*, <[https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20\(1\).pdf](https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf)>



Chapter conclusions

The lack of a universally accepted definition of hate speech provides space in which the content moderation process may address the issue of context, as set forth in the RPA and as underlined by the HRC. This is significant in terms of content that, for example, may be posted during or leading up to armed conflicts, particularly, in volatile political times or times during which targeted minorities are at risk of imminent harm. When dealing with hate speech, social media companies should look at the nature of Articles 19 and 20(2) as conceptualized in HRC documents, reports of the SRFOE and the RPA. The use of the RPA allows for a practical examination of the disputed content by looking at the six-part test on issues such as context and likelihood of harm to ensure that the determination that content should be removed matches the benchmarks laid out by IHRL. Accordingly, platforms should make sure that their Terms and content moderation guidelines properly reflect the standards set out by Articles 19 and 20 and train their AI and human moderators accordingly. Adopting an IHRL approach to tackling online hate will significantly narrow the applicable definition of hate speech on platforms such as Facebook and YouTube and raise the threshold for determining when speech – even speech that is deeply offensive and hurtful to various groups – can be removed. The negative consequences of strengthening the protection of controversial speech could be mitigated if platforms provide users with better means to adopting their own filters or use filters developed by third parties so that users can protect themselves from content they may deem offensive but falls short of the high threshold required to constitute hate speech under ICCPR Article 20(2).



Chapter 3: Disinformation

Introduction

In 2016, the Oxford English Dictionary named “post-truth” as its word of the year. This was after the US presidential election and the Brexit referendum in 2016, when misinformation and disinformation and their alleged disruptive effects on political debate had come under the microscope as a focal point in public discourse.¹⁰⁴ A study commissioned by the EU Parliament in 2019 noted that disinformation erodes faith in institutions, weakens democratic processes and distorts the information on which persons make crucial decisions during elections, health crises, and public emergencies and about important issues of governance.¹⁰⁵ Events such as the attack on the US Capitol by Trump supporters on 6 January 2021 demonstrated that disinformation can fan the flames of insurrection and contribute to large-scale deadly violence. Yet, the available evidence suggests that like hate speech, the amount of online disinformation has been significantly exaggerated by dominant narratives in traditional media and among many politicians about the “flood” of disinformation “drowning” social media. In fact, recent scientific evidence has shown that sharing articles from fake news domains was much rarer and less prevalent than perceived. A comprehensive study by Guess, Nagler and Tucker in the context of the 2016 US presidential election found that, while some groups of people (such as persons above the age of 65) were much more likely to share fake news, the vast majority of social media users – (90%) in the study – did not share any articles from fake news domains at all.¹⁰⁶ Another 2020 study found that, in contrast to popular perception, “fake news” comprised only 0.15% of Americans’ daily media diet.¹⁰⁷ In France, during the 2017 presidential election, a mere 4,888 out of sixty million tweets (less than 0.01 percent) were deemed to contain “fake news”.¹⁰⁸ Moreover, some experts warn that relying on censorship and restrictions on free speech to counter disinformation is both dangerous and counterproductive.

The report of the European Commission’s independent High-Level Group on fake news and online disinformation (“EU HLEG”) concluded “that the best responses to disinformation are multi-dimensional, with stakeholders collaborating in a manner that protects and promotes freedom of expression, media freedom, and media pluralism”.¹⁰⁹ The recommendation of the report was “to disregard simplistic solutions” and that “any form of censorship either public or private should clearly be avoided”. Several studies support this concern. For instance, there is evidence that improper and overbroad removals make some users suspicious and may counteractively reinforce conspiracy

¹⁰⁴ Hunt Allcott & Matthew Gentzkow, ‘Social Media and Fake News in the 2016 Election’ (2017) 31 *Journal of Economic Perspectives* 2.

¹⁰⁵ European Parliament (LIBE Committee) ‘Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and its Member States’ (2019), <[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU\(2019\)608864_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf)>

¹⁰⁶ Andrew Guess, Jonathan Nagler & Joshua Tucker, ‘Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook’ (2019), 5 *ScienceAdvances* 1.

¹⁰⁷ Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild & Duncan J. Watts ‘Evaluating the Fake News Problem at the Scale of the Information Ecosystem’ (2020) 6 *ScienceAdvances* 14.

¹⁰⁸ Andrew M. Guess & Benjamin A. Lyons, ‘Misinformation, Disinformation, and Online Propaganda’ in Persily & Tucker in Nathaniel Persily & Joshua A. Tucker (eds.), *Social Media and Democracy: Social Media and Democracy* (1st ed. Cambridge 2020), 26.

¹⁰⁹ European Commission, ‘Final Report of the High Level Expert Group on Fake News and Online Disinformation’ (2018), <<https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation>>



theories.¹¹⁰ A recent study published in Harvard's Misinformation Review documented how, once President Trump's election fraud Tweets were labelled as misinformation on Twitter, they gained more traction on other platforms.¹¹¹ The study argued that Twitter's labelling of certain Tweets was not only ineffective at preventing the spread of Trump's claims, it might have even backfired at an ecosystem level by drawing additional attention to messages that Twitter deemed problematic. Nevertheless, the pressure on platforms to remove disinformation is steadily increasing. The pressure to act decisively on misinformation and disinformation and the resultant calls to restrict more expression online may be better understood in the context of "elite panic", which sociologists Lee Clarke and Caron Chess explained as a phenomenon resulting when decision-makers are under intense media scrutiny to act decisively. They argue that, in such situations of crisis and upheaval, social elites might make rash decisions that could potentially make things worse than the very real problems against which these actions were aimed.¹¹²

Perhaps as a result of 'elite panic', the regulatory approach to disinformation in regions around the world has been dominated by the criminalization of speech.¹¹³ Criminal provisions are considered the most intrusive and speech-restrictive in nature because they cause a severe chilling effect on free speech. Moreover, much national legislation dealing with misinformation and disinformation contains overbroad provisions, vaguely-worded definitions and no objective threshold for narrowing the contours of mis/disinformation. After the onset of the COVID-19 pandemic in 2020, at least 24 countries around the world passed vague laws to punish the spread of misinformation. Such legislation was misused extensively to silence legitimate criticism of governmental responses to the pandemic, media reporting, and social media engagement by users on government responses.¹¹⁴ In Wuhan, when Dr. Li Wenliang warned about a novel infection in January 2020, he was questioned for eight hours by China's Public Security Bureau and threatened with prosecution for making "false comments" and spreading rumors.¹¹⁵ These criminal laws chill speech in two ways: they target individual users who spread disinformation and, secondly, they pressure social media platforms, which might tend excessively to remove legitimate content. This could ultimately lead to inappropriate censorship and stifle discussion of issues of public interest as well as criticism of the government.

¹¹⁰ Or Levi, Pedram Hosseini, Mona Diab, David A. Broniatowski, 'Identifying Nuances in Fake News vs Satire: Using Semantic and Linguistic Cues' *Association for Computational Linguistics* (2019),

¹¹¹ 'Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both On and Off the Platform' (2029), 2 *Harvard Kennedy School Misinformation Review* 4.

¹¹² Lee Clarke & Caron Chess, 'Elites and Panic: More to Fear than Fear Itself' (2008), 87 *Social Forces* 2; Jacob Mchangama, 'Free Speech: A History from Socrates to Social Media' (1st ed. Basic Books forthcoming 2022)

¹¹³ Sarah Shirazy, Allen Weiner, Yvonne Lee & Madeline Magnuson et al., 'How to Reconcile International Human Rights Law and Criminalization of Online Speech: Violent Extremism, Misinformation, Defamation, and Cyberharassment' (2020), *Stanford Law School Law and Policy Lab*,

<<https://law.stanford.edu/publications/how-to-reconcile-international-human-rights-law-and-criminalization-of-online-speech-violent-extremism-misinformation-defamation-and-cyberharassment/>>

¹¹⁴ Human Rights Watch, 'Covid-19 Triggers Wave of Free Speech Abuse' (2021), <<https://www.hrw.org/news/2021/02/11/covid-19-triggers-wave-free-speech-abuse>>

¹¹⁵ Stephanie Hegarty, 'The Chinese Doctor Who Tried to Warn Others about Coronavirus' (2020), *BBC News* <https://www.bbc.com/news/world-asia-china-51364382>.



In addition, countries have also sought to tackle disinformation and to outsource censorship through the lens of intermediary liability obligations for online platforms.¹¹⁶ Such legislation generally allows for the imposition of heavy regulatory fines and even criminal sanctions on company executives in the event of the failure to remove disinformation and other content such as hate speech in an expeditious manner.¹¹⁷ Under these pressures, platforms have responded by increasingly relying on automated content-filtering algorithms and moving towards the removal of more and more content (as evident from platforms' increased content removal figures discussed later in the chapter).

There is also the fundamental question of whether *anybody* should have the authority to determine definitively whether content is true or not. It is dangerous to vest this authority in the hands of governments because this means that governments become the arbiters of truth for online speech and governments often get it wrong. For example, although scientists and many governments and intelligence agencies (in the UK, Australia, New Zealand and Canada) initially ridiculed the theory that COVID-19 had entered the population through a laboratory leak, some of the same countries (UK, Australia, and other G7 countries) are now exploring it seriously.¹¹⁸ In April 2020, social media platforms such as Facebook announced that they were following the then-prevailing opinion and removing lab leak theories from their platforms as COVID-10 misinformation;¹¹⁹ but after removing claims on the issue for over a year, platforms have reversed their policies on removing discussions of lab leaks.¹²⁰

The First Step in Effectively Dealing with Disinformation: Defining Disinformation

In theory, platforms would need to set forth a clear taxonomy of terms with respect to disinformation to be able to formulate speech-protective policies in these areas. In practice, however, it appears highly challenging for social media platforms to do so. This is because the lines between misinformation, malinformation and disinformation are often blurred, and it is difficult to assess motivation from content alone since members of the public often engage in highly ambiguous practices online with indiscernible motivations.¹²¹ If the classification underpinning the search for false information is unscientific or vague, errors will be common, and their regulation will be less effective.¹²² A definition that is too broad, vague, or ambiguous is open to abuse and overzealous

¹¹⁶ Kalina Bontcheva & Julie Posetti (eds.) 'Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression' (2020), *ITU & UNESCO*, <https://www.broadbandcommission.org/Documents/working-groups/FoE_Disinfo_Report.pdf>

¹¹⁷ For example, legislation such as the German NetzDG 2018, France's law against information legislation, as well as pending proposals such as Ireland's Proposal to Regulate Transparency of Online Political Advertising.

¹¹⁸ Dan Sabbagh, 'Five Eyes Network Contradicts Theory Covid-19 Leaked from Lab' (2020), *The Guardian*.

<<https://www.theguardian.com/world/2020/may/04/five-eyes-network-contradicts-theory-covid-19-leaked-from-lab>>; Glenn Kessler, 'Timeline: How the Wuhan Lab-Leak Theory Suddenly Became Credible' (2021), *The Washington Post*, <<https://www.washingtonpost.com/politics/2021/05/25/timeline-how-wuhan-lab-leak-theory-suddenly-became-credible/>>

¹¹⁹ Facebook Update on Misinformation (April 2020), <<https://about.fb.com/news/2020/04/covid-19-misinfo-update/>>

¹²⁰ Taylor Hatmaker, 'Facebook Changes Misinfo Rules to Allow Posts Claiming Covid-19 is man-made' (2021), *The Crunch*, <<https://techcrunch.com/2021/05/28/facebook-covid-man-made-lab-theory/>>

¹²¹ Whitney Phillips & Ruan M. Milner, 'The Ambivalent Internet. Mischief, Oddity and Antagonism Online' (1st ed. Polity Press 2017)

¹²² Darrin Baines & Robert J R Elliot, 'Defining Misinformation, Disinformation and Malinformation: An Urgent Need for Clarity during the Covid-19 Infodemic' (2020), *Discussion Papers 20-26, University of Birmingham*.



implementation by both governments and social media platforms, and this could jeopardize freedom of expression.¹²³

The HLEG defines disinformation as “all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.” In 2017, a Joint Declaration (Joint Declaration) on “Fake News”, Disinformation and Propaganda was adopted by the SRFOE, the Organization for Security and Co-Operation in Europe (OSCE) Representative on Freedom of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples’ Rights (ACHPR).¹²⁴ This Joint Statement suggested the following definitions:

- **Disinformation** consists of statements that are known or reasonably should be known to be false. It misleads the population, and as a side effect it interferes with the public’s right to know and the right of individuals to seek, receive, and impart information.
- **Misinformation** is false information, but the person who is disseminating it believes it to be true.
- **State-sponsored propaganda** consists of statements by state actors that demonstrate a reckless disregard for verifiable information.

Other concepts and definitions related to disinformation are:

- **Fake news**, a phrase popularized by former President Trump, is unhelpful from a legal and definitional perspective. It blurs crucial elements such as the information’s content, speaker’s intent and impact of the information. It is, rather, a catch-all phrase that could potentially include anything from deepfakes to hate speech against a particular community and even legitimate criticism from unfavorable news outlets. Thus, the use of fake news should be avoided in discourse surrounding misinformation as it has the potential to erode the protection of freedom of expression¹²⁵ – for example, by mixing up legitimate criticism stemming from journalists and civil society with deliberately fabricated information disseminated to spread falsehood and undermine verifiable facts, providing politicians a pretext to censor both.
- **Malinformation** is “reconfigured true information” that requires both intention and equivalence and often involves a repurposing of the truth value of information for deceptive ends. A 2020 report on misinformation by the Reuters Institute at the University of Oxford explains that this is unlike disinformation since it intentionally repurposes truth to deceive its

¹²³ Alessio Sardo, ‘Categories, Balancing and Fake News’ The Jurisprudence of the European Court of Human Rights’ (2020), 33 *Canadian Journal of Law & Jurisprudence* 2.

¹²⁴ Joint Declaration on Freedom of Expression and Fake News, Disinformation and Propaganda’ (2017). <<https://www.osce.org/files/f/documents/6/8/302796.pdf>>

¹²⁵ Tarlach McGonagle, ‘Fake News: False Fears or Real Concerns?’ (2017), 35 *Netherlands Quarterly of Human Rights* 4.



reader/viewer unlike disinformation, which intentionally uses false information to deceive its user.¹²⁶

Thus, platforms' definitions of disinformation must be narrowly tailored and consider some of the concepts discussed above such as the content of the information, the intent of the speaker, and the identity of the speaker (in the context of state-sponsored propaganda). Later in this chapter, we use the umbrella term 'disinformation' to refer broadly to the moderation difficulties around misinformation, disinformation, propaganda and malinformation.¹²⁷ If we wish to refer specifically to misinformation, malinformation or propaganda, we will do so explicitly at that point.

Regulation of Disinformation under IHRL

Disinformation impacts a variety of rights under IHRL, for example, as articulated in Article 19 ICCPR and Article 25 ICCPR. However, unlike hate speech, disinformation does not constitute a specific category of speech exempted from the protection of Article 19 or subject to specific prohibition such as, for example, under Article 20(2). As a result, there is scope for discussion about what kinds of disinformation are permitted under IHRL. That said, it appears amply clear that IHRL does not generally justify the "dissemination of knowingly or recklessly false statements, especially by official or State Actors."¹²⁸ It also does not protect disinformation that rises to the level of incitement to violence, hate speech or fraud.¹²⁹

Article 25's protection of participation in public affairs is relevant for platforms and content moderation because disinformation might be targeted to manipulate the effective right to participate in a democracy. Disinformation could impede the flow of information and facts, which might affect how voters participate in the electoral process.¹³⁰ That said, overbroad removal of unverified political information in the guise of dealing with misinformation could also affect participation in public affairs by thwarting democratic debate and slowing down the flow of discourse.

On disinformation and IHRL, the 2017 Joint Declaration is the most explicit instrument to underscore the adoption of IHRL for regulating disinformation online. Some of the key points noted in the 2017 Joint Declaration are:

¹²⁶ Scott Brennen, Felix Simon, Philip Howard & Rasmus Kleis Nielsen, 'Types, Sources and Claims of Covid-19 Misinformation' (2020), *Reuters Institute & University of Oxford*, <<https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation#conclusions>>

¹²⁷ Darrin Baines & Robert J R Elliot, 'Defining Misinformation, Disinformation and Malinformation: An Urgent Need for Clarity during the Covid-19 Infodemic' (2020), *Discussion Papers 20-26, University of Birmingham*.

¹²⁸ Joint Declaration on Freedom of Expression and Fake News, Disinformation and Propaganda' (2017), <<https://www.osce.org/files/f/documents/6/8/302796.pdf>>

¹²⁹ Sarah Shirazyan, Allen Weiner, Yvonne Lee & Madeline Magnuson et al., 'How to Reconcile International Human Rights Law and Criminalization of Online Speech: Violent Extremism, Misinformation, Defamation, and Cyberharassment' (2020), *Stanford Law School Law and Policy Lab*,

<<https://law.stanford.edu/publications/how-to-reconcile-international-human-rights-law-and-criminalization-of-online-speech-violent-extremism-misinformation-defamation-and-cyberharassment/>>

¹³⁰ European Parliament (LIBE Committee), 'Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and its Member States' (2019), <[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU\(2019\)608864_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/608864/IPOL_STU(2019)608864_EN.pdf)>



- The right to impart information and ideas is not limited to “correct” statements, but, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or state actors;
- General prohibitions on the dissemination of information based on vague and ambiguous ideas of “non-objective information” or “false news” are incompatible with international standards as set forth in Article 19.
- Intermediaries should never be liable for any third-party content relating to those services unless they specifically intervene in that content or refuse to obey an order by an independent and authoritative oversight body (such as a court) to remove it and they have the technical capacity to do that.
- Where intermediaries intend to take action to restrict third-party content that goes beyond legal requirements, they should adopt clear, predetermined policies governing those actions. Those policies should be based on objectively justifiable criteria rather than ideological or political goals and should, where possible, be adopted after consultation with their users.
- Intermediaries should respect minimum due process guarantees including the prompt notification of users when content that they created, uploaded or host may be subject to a content action and giving the user an opportunity to contest that action. This requirement is also applicable when this determination occurs through an algorithmic process.
- Intermediaries should support the research and development of appropriate technological solutions for disinformation and propaganda, which users may apply on a voluntary basis. They should cooperate with initiatives that offer fact-checking services to users and review their advertising models to ensure that they do not adversely impact the diversity of opinions.¹³¹

¹³¹ Joint Declaration on Freedom of Expression and “Fake News”, Disinformation and Propaganda, <<https://www.osce.org/files/f/documents/6/8/302796.pdf>>

Platform Policies on Disinformation

In her 2021 Report on Disinformation, SRFOE Irene Khan called for multidimensional responses to disinformation that are well grounded in an IHRL framework. She urged platforms to adopt clear, narrowly defined content and advertising policies on disinformation and misinformation with a special emphasis on adopting clear policies relating to public figures that are consistent with IHRL standards, applying them consistently across geographical areas.

Social media platforms have adopted an approach to disinformation that involves wide-scale removals, and this has become particularly pronounced during the COVID-19 pandemic. Facebook announced on its company blog, Newsroom, that, between the start of the pandemic in March 2020 and April 2021, it took down 18 million pieces of content from Facebook and Instagram for violating its COVID-10 misinformation policies.¹³³ This trend was also visible for Twitter and YouTube: after the onset of COVID, between 18 March 2020 and 14 July 2020, Twitter took down 14,900 Tweets and “challenged” 4.5 million accounts that regularly posted COVID-19 misinformation. YouTube’s removal figures showed a 93% rise between Q4 2019 (5.8 million removals) and Q2 2020 (11.4 million removals). YouTube’s removals decreased slightly in 2021, and it took down 9.5 million pieces of content in Q1 2021.¹³⁴ This broad increase in removal is also accentuated and partly attributable to the use of AI.

The first step to determine whether platform regulation conforms to IHRL standards is to assess the relevant Terms and compare them with obligations under Article 19 ICCPR.



Facebook’s Terms contain a section on ‘False News’, and its policy is not to remove ‘false news’ but, instead, show it lower on news feeds.¹³⁵ In this policy, it does not define ‘false news’ and does not provide much direction about what it would remove under this policy. A vague policy that does not provide adequate clarity to users about what content is prohibited might not comply with ‘precision’ and ‘foreseeability’ requirements Article 19(3).

Facebook also has a policy on ‘Manipulated Media’ where it provides for the removal of images and videos. It considers media manipulated if it, in ways that are not apparent to an average person, is likely mislead an average person to believe that subjects of the video said words they did not say. The policy appears to allow a blanket removal of synthetic media even if it does not lead to imminent harm.

¹³² Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018), A/HRC/47/25 <<https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Report-on-disinformation.aspx>>

¹³³ Facebook Community Standards Enforcement Report, First Quarter 2020: <<https://about.fb.com/news/2021/05/community-standards-enforcement-report-q1-2021/>>

¹³⁴ YouTube Community Guidelines Enforcement, <https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_channels_removed=period:2020Q1&lu=total_removed_videos&total_removed_videos=period:2021Q1;exclude_automated:all>

¹³⁵ Facebook Community Standards: False News, <https://www.facebook.com/communitystandards/recentupdates/false_news>

Facebook introduced a detailed section on COVID-19 misinformation in 2020.¹³⁶ In an apparent departure from its 'downranking' policy, it said that it would remove COVID-19 content that contributes to a risk of real-world harm. In an interview for the Clear and Present Danger podcast, Monica Bickert (Facebook's Head of Global Policy Management) discussed with Jacob Mchangama that statements such as "the working class is immune to COVID" regardless of whether the person making the statement had 100 or 5000 friends would be seen as fomenting imminent physical harm. Since 2020, this policy has evolved and expanded. Now Facebook removes many additional claims such as claims linking the cause of COVID-19 to 5G communication technologies.

In addition to its Community Standards on 'false news', Facebook explains on its Newsroom blog that its policy to counter misinformation has three parts: *first*, to remove accounts and content that violate its Community Standards or ad policies;¹³⁷ *second*, to reduce the distribution of false news and inauthentic content such as clickbait; and, *third*, to inform people by giving them more context on the posts they see. However, even in this blog, Facebook does not provide a definition of misinformation or disinformation.



YouTube does not provide a definition of disinformation but prohibits categories of disinformation through different policies. In a non-exhaustive list in its misinformation policy, it forbids users from posting "content that has been technically manipulated or doctored in a way that misleads users".¹³⁸ This type of manipulated

media goes beyond clips taken out of context and may pose a serious risk of egregious harm. It also prohibits users from posting "content aimed to mislead voters about the time, place, means, or eligibility requirements for voting, or false claims that could materially discourage voting."

After the Capitol Hill riot on 6 January 2021, YouTube inserted a provision on presidential election integrity. This section prohibited any content that claimed a "candidate only won a swing state in the U.S. 2020 presidential election due to voting machine glitches that changed votes" or that "dead people voted in numbers that changed the outcome" of the election.¹³⁹ In an important disclaimer on the same page, the company said that it may allow content that violates its U.S. presidential election integrity policy if the content includes additional context in the video, audio, title, or description. It said that "this is not a free pass to promote misinformation. Additional context may include countervailing views, or if the purpose of the content is to condemn, dispute, or satirize misinformation that violates our policies."

¹³⁶ Facebook, COVID-19 and Vaccine Policy Updates and Protections, <<https://www.facebook.com/help/230764881494641>>

¹³⁷ Facebook, 'Hard Questions: What's Facebook Strategy for Stopping False News?', <<https://about.fb.com/news/2018/05/hard-questions-false-news/>>

¹³⁸ YouTube, Spam, Deceptive Practices & Scams Policies, <https://support.google.com/youtube/answer/10834785>

¹³⁹ YouTube, Election Misinformation Policies, <<https://support.google.com/youtube/answer/10835034#zippy=%2Cus-presidential-election-integrity>>

YouTube introduced a separate section for 'COVID-19 misinformation' in which it said it did not allow content about COVID-19 that posed a 'serious risk of egregious harm' and does not allow content that contradicts WHO or local health authorities' guidance on the treatment, prevention, diagnostics or transmission of COVID-19.¹⁴⁰ The policy provides a range of examples of treatment misinformation, prevention misinformation, diagnostic misinformation, transmission misinformation, vaccines and other things. This policy is notable for its extensive use of examples to illustrate the threshold of what constitutes medical misinformation. While these examples are consonant with the need for clarity as envisaged by the 'legality' requirement in Article 19(3), the requirement of the risk of 'egregious harm' does not clarify the exact nature of harm or how imminent such harm should be.

Instagram introduced a policy similar to Facebook's to remove COVID-19 misinformation that focusses on removing "content that has the potential to contribute to real-world harm, including through ... coordination of harm, sale of medical masks and related goods, hate speech, bullying and harassment and misinformation that contributes to the risk of imminent violence or physical harm."¹⁴¹ Like Facebook, Instagram explains that it works with over 60 certified fact-checking organizations that identify false information, altered content or content with no context.¹⁴²

Instagram's policies pose similar challenges with respect to IHRL as Facebook's. The lack of specificity and clear definitions might lead to vagueness and inconsistent application, that might imply lack of compliance under Article 19(3).



Reddit does not define misinformation or disinformation in its Terms.¹⁴³ It only says that it disallows content that impersonates individuals or entities in "a misleading or deceptive manner".¹⁴⁴ This includes using a Reddit account to impersonate someone. It also encompasses "domains that mimic others, as well as deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity." In an interview, a Reddit representative explained Reddit's decentralized method of moderation: "99% of all content moderation decisions on Reddit are undertaken by volunteer community moderators who are just Reddit users. We empower the users to set their own specific rules that apply specifically in their individual subreddits particular to the topic of whatever they want to discuss."¹⁴⁵ Reddit

¹⁴⁰ YouTube, Covid 19 Medical Misinformation Policy,

<https://support.google.com/youtube/answer/9891785?hl=en&ref_topic=9282436>

¹⁴¹ Instagram, Community Guidelines: <<https://www.facebook.com/help/instagram/477434105621119>>

¹⁴² Instagram, Reducing the Spread of Misinformation on Instagram',

<[https://help.instagram.com/1735798276553028/?helpref=hc_fnav&bc\[0\]=Hj%C3%A6lp%20til%20Instagram&bc\[1\]=Politik%20og%20anmeldelse](https://help.instagram.com/1735798276553028/?helpref=hc_fnav&bc[0]=Hj%C3%A6lp%20til%20Instagram&bc[1]=Politik%20og%20anmeldelse)>

¹⁴³ Reddit: Content Policy, <<https://www.redditinc.com/policies/content-policy>>

¹⁴⁴ Reddit, 'Do Not Impersonate an Individual or Entity' <<https://www.reddithelp.com/hc/en-us/articles/360043075032>>

¹⁴⁵ Interview with Reddit representative, 23.04.21.

employees, known as administrators, apply the site's high-level 'Content Policy' to relevant issues such as COVID-19 misinformation on their specific subreddit. Under this policy, "misinformation encouraging physical harm, such as through directing users to inject disinfectant as a means to treat COVID-19, is removed under the site's rule against encouraging violence or physical harm. However, on top of these site-wide rules, which form a baseline, each individual subreddit's community moderators set their own rules for what is allowed and what is prohibited on their specific subreddit. These rules are in addition to the requirements of the Content Policy. Accordingly, a community can choose, for example, whether to only allow COVID-19 information that has been published in peer-reviewed journals and so forth."¹⁴⁶



Like other platforms, **Twitter** does not adequately define what it considers misinformation. Instead, it lays down three criteria to decide when content should be labelled or removed for violating its synthetic and manipulated media policy.¹⁴⁷ *First*, it assesses whether that media, or the context in which it is presented is "significantly and deceptively altered or manipulated." Subtler forms of manipulated media, such as isolative editing, omission of context, or presentation with false context, may be labeled or removed on a case-by-case basis. *Second*, it considers whether the context in which media are shared might result in confusion or misunderstanding or indicates a deliberate intent to deceive people about the nature or origin of the content, for example, by falsely claiming that it depicts reality. *Lastly*, synthetic and manipulated media that are "likely to cause serious harm" are removed altogether. Twitter explains that specific harms might include threats to the physical safety of a person or group, risk of mass violence, threats to privacy/stalking/unwanted and obsessive attention, targeted content that aims to silence someone or voter suppression or intimidation.

In addition, Twitter has a separate policy for disinformation that affects civil processes (political elections and major referenda). It prohibits "the distribution of false or misleading information about the procedures or circumstances around participation in a civic process."¹⁴⁸ The policy is aimed at four categories of misleading content: misleading information about how to participate; suppression and intimidation; misleading information about outcomes; and false or misleading affiliation of candidates and political parties. In instances in which misleading information does not seek to manipulate or disrupt civic processes directly but leads to confusion, Twitter does not remove the content but might label it to provide additional context. As compared to other platforms, Twitter's guidelines appear more connected to specific harms that arise from the categories of speech prohibited under Article 19(3) or to the protection of rights under Article 25. It also clarifies what is liable to labelling and what is liable to removal. That said, its compliance with IHRL is weakened by

¹⁴⁶ Interview with Reddit representative, 23.04.21.

¹⁴⁷ Twitter, Synthetic and Manipulated Media Policy, <<https://help.twitter.com/en/rules-and-policies/manipulated-media>>

¹⁴⁸ Twitter, Civic Integration Policy, <<https://help.twitter.com/en/rules-and-policies/election-integrity-policy>>

allowing for blanket removal of deeply manipulated media even if it does not lead to any imminent harm or if it does not fall under a category of speech restricted under Article 19(3).









TikTok's Community Standards on Integrity and Authenticity state that it does not permit misinformation that causes harm to individuals, the community or the larger public regardless of intent. TikTok explains that this includes misinformation that incites hate or prejudice, misinformation related to emergencies that induces panic, medical misinformation that can cause harm to an individual's physical health, content that misleads community members about elections or other civic processes, conspiratorial content that attacks a specific protected group or includes a violent call to action, or denies that a violent or tragic event occurred, digital forgeries (synthetic media or manipulated media) that mislead users by distorting the truth of events and cause harm to the subject of the video, other persons, or society. TikTok's policy appears to be speech-protective in the sense that all categories of speech prohibited are linked to protecting its users from real-world and imminent harm.

On an overall assessment, it seems clear that platform policies on misinformation vary greatly from one another. Some elements of their policies meet the threshold of protection to freedom of expression under Article 19(3) ICCPR, such as the clarity and foreseeability of prohibited content provided by YouTube's explanatory definition of misinformation or how Twitter attempts to meet the necessity requirement by acting only on misinformation that could lead to real-world harm. That said, some other elements of platform policies could stand to become more nuanced and more aligned with IHRL standards, such as a general lack of definitude in most platform policies about what crosses the threshold of misinformation to warrant removal and limiting removal only to misinformation that causes real-world and imminent harm.







Comparison metric – Severity of Enforcement Actions

(This metric ranks platform enforcement policies based on severity)

	 Facebook	 Instagram	 Twitter	 YouTube	 Reddit	 TikTok
Links enforcement action (content removal) to real-world harm			✓	✓		✓
Labels and downranks misinformation	✓	✓	✓		✓	
Removes misinformation	✓ ¹⁴⁹	✓	✓	✓	✓	✓

Comparison Metric – Platform Policies on Misinformation

(This metric shows the level of clarity in platform misinformation policies as well as the expansion of restrictions for different content categories)

	 Facebook	 Instagram	 Twitter	 YouTube	 Reddit	 TikTok
Defines misinformation				✓		✓
Provides examples of what content is removed under the misinformation policy			✓	✓		✓
Exemption for satire and opinion	✓		✓	✓	✓	✓
Policy for medical misinformation and COVID-19	✓	✓	✓	✓	* 150	✓
Policy on election integrity			✓	✓		✓
Policy on synthetic media	✓		✓	✓	✓	✓
Policy on conspiracy theories						✓

¹⁴⁹ Facebook downranks other categories of misinformation but removes misinformation that contains synthetic media and COVID-19 misinformation with a potential for causing real-world harm

¹⁵⁰ Reddit released an explanatory subreddit on medical misinformation after COVID-19, but it did not amend its Community Guidelines.



Identifying a Threshold for Disinformation

In determining the limits of disinformation from an IHRL perspective, social media companies may focus on 1) content, 2) context, 3) intent, and 4) impact in their assessment of cases involving interference with the right to freedom of expression. Although this test is broadly based on the RPA, it has slight variations. The RPA isn't fully relevant for misinformation in the way that it is for hate speech. For example, the RPA lays emphasis on the status of the speaker but this test does not do so. This is because if misinformation leads to imminent physical harm (for example, bleach and COVID 19), it warrants removal under the public health restriction under Article 19(3) even if it is posted by a user with only 20 friends.

While the requirement to assess the context and impact might be slightly easier to conduct, it is especially tricky for platforms to assess the intent (or lack thereof) of the speaker in spreading disinformation. It is difficult for a content-filtering algorithm or even a human content moderator to determine decisively whether a user disseminated the information knowingly or unknowingly. This is because there might be various motivations for someone spreading disinformation. Internet users may have many reasons for sharing pieces of misinformation and disinformation, including a desire to “troll” or for ideological reasons.¹⁵¹ That said, user intent might also be crucial to consider when it comes to issues of humor, satire and parody but also in relation to political dissent/critique of government practices. The question of intent and threshold was discussed in a 2020 case by the Oversight Board. It found Facebook’s misinformation and imminent harm rule to be inappropriately vague and recommended a new Community Standard on health misinformation. The case involved a video that alleged the lack of a health strategy in France and questioned what society had to lose by allowing doctors to prescribe hydroxychloroquine combined with azithromycin for use against Covid-19. The Board noted that the user was opposing governmental policy and did not encourage users to buy or take medicines without a prescription. In this light, the Board held that Facebook did not demonstrate that the post would rise to the level of imminent harm.¹⁵²

Evaluating the content of supposed disinformation also poses difficult challenges across content categories. For example, distinguishing between an incorrect factual claim and an opinion with respect to perceived medical disinformation has proved extremely challenging during COVID-19 because authorities have constantly changed their guidance on scientific issues, such as the efficacy of medical masks.

In removing content based on its “impact” or harm, IHRL clarifies that free expression may only be restricted on the grounds provided in Article 19(3). The European Commission’s ‘Communication on

¹⁵¹ Scott Brennen, Felix Simon, Philip Howard & Rasmus Kleis Nielsen, ‘Types, Sources and Claims of Covid-19 Misinformation’ (2020), *Reuters Institute & University of Oxford*, <<https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation#conclusions>>

¹⁵² Oversight Board: Case 2020-006-FB-FBR <<https://oversightboard.com/news/325131635492891-oversight-board-overturms-facebook-decision-case-2020-006-fb-fbr/>>



Disinformation' released in June 2018 explained 'public harm' to mean "threats to democratic political and policymaking processes as well as public goods such as the protection of EU citizens' health, the environment or security".¹⁵³ However, the relevant threat that disinformation causes to such interests must be serious and immediate in nature.

Thus, from the outset, only limited and qualified instances of intentional or "bad faith" disinformation entailing immediate real-world harm should be subject to the most intrusive restrictive measures such as content removal. Immediate real-world in this scenario means physical or mental injury (for example, self-injecting hydrochloroquine to cure COVID) or substantial impairment of fundamental rights (for example, a morphed nude photo of someone affecting their reputation pursuant to Article 19(3)). Other forms of misinformation likely to result in less serious harm may be subject to less restrictive measures such as labelling or downranking. Based on the principles discussed above, we explain how a platform may use the tests of content, context, intent and impact to regulate disinformation.

¹⁵³ EU Code of Practice on Disinformation, <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54454>

Case Studies

Modi and COVID-19

During the April 2021 COVID-19 wave in India, a Twitter user wrote a Tweet claiming that a 24/7 crematorium had been set up by Prime Minister Narendra Modi.



Analysis of this case under IHRL



CONTENT

Since April 2021, India had been experiencing a devastating second wave of COVID-19. Hospitals and crematoriums in some of India's biggest cities such as Mumbai, New Delhi, Lucknow and Pune ran out of space, and there was a huge shortage of medical oxygen and drugs.¹⁵⁴ The government came under intense criticism for fuelling this surge by allowing a religious festival in India's most populous state and by holding election gatherings with thousands of unmasked attendees in five states heading for elections.¹⁵⁵ By the end of April 2021, India reported over 400,000 new cases and over 3,689 deaths every

¹⁵⁴ Vikas Pandey & Shadab Nazmi, 'India Covid-19: Deadly Second Wave Spreads from Cities to Small Towns' (2021,) *BBC News*, <<https://www.bbc.com/news/world-asia-india-56913047>>

¹⁵⁵ Ruhi Tehwari, 'Poll Rallies to Kumbh Mela — Modi-Shah's Conscience Must Take a Look at Latest Covid Surge' (2021), *The Print*, <<https://theprint.in/opinion/politricks/poll-rallies-to-kumbh-mela-modi-shahs-conscience-must-take-a-look-at-latest-covid-surge/639526/>>; Hasan Kamal, 'Kumbh Mela and Election Rallies: How Two Super Spreader Events have Contributed to India's Massive Second Wave of Covid-19 Cases,' *Firstpost*, <<https://www.firstpost.com/india/kumbh-mela-and-election-rallies-how-two-super-spreader-events-have-contributed-to-indias-massive-second-wave-of-covid-19-cases-9539551.html>>

day.¹⁵⁶ Social media was abuzz with anger and criticism of the government's failure to contain and tackle the situation.



CONTEXT

A Twitter user shared an image of a crematorium with multiple lit fires. The user captioned the post by stating that the world's first "24*7" crematorium was launched by the Modi government. The picture did not depict PM Modi or any other identifiable persons. It only showed two persons with PPE equipment (face masks and face shield) standing in the corner next to the burning pyres. In this post, the user makes two claims: 1) That a "24*7" crematorium has been launched in India by PM Modi, and 2) the crematorium is the first of its kind in the world. The content of the post does not seem to incite any imminent physical harm based on misinformation and appears to be a commentary on the state of affairs in India during the second wave.



INTENT

The intent of the user, when assessed in the context of the nationwide outrage against the government, appeared to be sarcasm and snark in criticizing the government for its failure to handle the COVID-19 pandemic. Given the social backdrop and political climate, it appears evident that the intent of the user was not to deceive a viewer into actually believing that such a crematorium was launched in India but to contribute to political debate and criticize the government for its inadequate response to the pandemic and the resultant loss of life.



IMPACT

The post does not seem to threaten or cause any imminent harm and does not violate any rights protected under Article 19(3). The Prime Minister, as a public official, would not be in a position to claim protection to his reputation as he is a political figure, and the user is merely contributing to political debate.¹⁵⁷

Thus, the removal of this post for spreading false information would not be justifiable under IHRL.

¹⁵⁶ BBC News, 'India Coronavirus: New Record Deaths as Virus Engulfs India,' <<https://www.bbc.com/news/world-asia-india-56961940>>

¹⁵⁷Lingens v Austria, Application No. 9815/82 (ECHR 8 July 1986) Stijn Smet, 'Freedom of Expression and the Right to Reputation: Human Rights in Conflict' <<https://www.corteidh.or.cr/tablas/r29311.pdf>>

Stanford Professor of Disease Prevention, COVID-19 Interview Removed

In April 2020, YouTube removed a video in which Dr. John Ioannidis, Professor of Disease Prevention at Stanford University in California, USA, questioned the rationale behind the lockdowns imposed by the government. The video was removed by YouTube six weeks after it was originally uploaded for violating its medical misinformation policies.¹⁵⁸

Analysis of this case under IHRL



CONTENT

The video contained a sit-down, long-form interview of Dr. John Ioannidis in which he discussed a recent article he authored that questioned and criticised the approach of locking down entire cities and states in response to COVID-19.¹⁵⁹ In the interview, Dr. Ioannidis highlighted the lack of credible data on COVID-19 to justify the decision-making process by governments around the world. He also said that a “large majority” of medical studies and drug studies had major flaws and were “essentially wrong”. He said, “If we shut everyone in their house, it is a solution. If we manage to even isolate everyone, not even being in touch with any other person, in theory, we are containing the spread of the virus... It has lots of consequences, and for a society like ours, it means that very soon you will start seeing a major impact on the economy”¹⁶⁰. As he was merely questioning the public health rationale of the policies adopted by government, the video constituted content on public health protected under Article 19(3).



CONTEXT

The video was released on 23 March 2020, days after the WHO had declared COVID-19 to be a global pandemic and then-President Trump had declared a national emergency.¹⁶⁰ The interview was recorded at Stanford University in California; and, on 19 March, California became the first state in the US to issue a state-wide stay-at-Home order.



INTENT

When assessed in the context of the evolving governmental response to COVID-19, the speaker intended to contribute to public discourse by critiquing the governmental response on an issue of immediate public importance. Moreover, the video was produced in a documentary form by a verified channel that had previously made videos on other issues of public importance. In addition, the interviewee was a Professor of Disease Prevention.

¹⁵⁸ Michael A. Alcorn, ‘How Wrong was Ioannides?’ (2020) <<https://michaelaalcorn.medium.com/how-wrong-was-ioannidis-5940e49c9af6>>

¹⁵⁹ Full transcript of the interview: <<https://www.thepressandthepublic.com/post/perspectives-on-the-pandemic-i>>

¹⁶⁰ AJMC, ‘A Timeline of COVID-19 Developments in 2020’ (2021), <<https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020>>

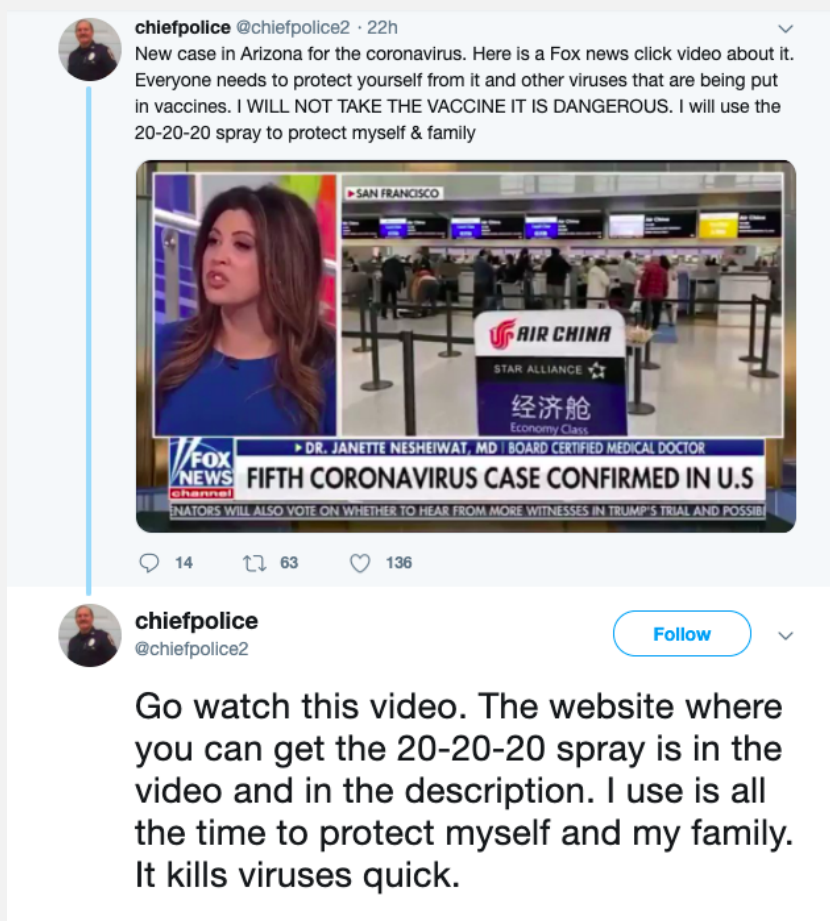


The video did not threaten or cause any imminent harm and did not violate any rights protected under Article 19(3). Dr. Ioannidis did not encourage viewers to breach restrictions imposed by local laws, and he did not offer any medical advice that might put a viewer at risk for increased infection or harm. Before the video was taken down, it was viewed more than half a million times.¹⁶¹

Thus, the removal of this post for spreading false information would not be justifiable under IHRL.

Prescribing plant fertilizer to kill COVID-19

On January 27, 2020, a Twitter user made a post encouraging his followers to administer a plant fertilizer called 20-20-20 spray to themselves and their family members to kill the Coronavirus. The user was a popular QAnon supporter with over 17.2K followers. Although the Tweet was removed by Twitter, web archival records show that the Tweet gathered over 63 retweets and 136 likes shortly after it was posted.



¹⁶¹ Peter Jamison, 'A Top Scientist Questioned Virus Lockdowns on Fox News. The Backlash was Fierce' (2020), *The Washington Post*, <<https://www.washingtonpost.com/dc-md-va/2020/12/16/john-ioannidis-coronavirus-lockdowns-fox-news/>>

Analysis of this case under IHRL



CONTENT

A user Tweeted a screenshot of a Fox News video about a fifth Coronavirus case in the United States. He claimed that everyone needed to protect themselves from the Coronavirus and “other viruses being put in vaccines”. The user said that he would not take “the vaccine” as it was dangerous and that he would instead take the 20-20-20 spray to protect himself and his family. He claimed that the 20-20-20 spray kills viruses quickly.¹⁶² He also encouraged users to check out a linked website from where they might buy the product.

Note: The 20-20-20 spray is a greenhouse plant fertilizer for soil application containing ammonium, nitrogen, phosphorous and potassium. Given its intended use as a plant fertilizer, it appears to be unsuitable for human consumption.



CONTEXT

The post was made on January 27, 2020 at a time when COVID-19 had not been declared a pandemic by the WHO and it was slowly starting to spread outside of China.¹⁶³ At this point, there existed scarce information amongst the public and health authorities about the nature of the virus, its origin, and its treatment. As the screenshot also depicts, the Coronavirus was slowly spreading in the US and creating global fear and panic given the strict isolation measures and deaths being reported in Chinese cities, specifically Wuhan. The user was a popular supporter of the QAnon group of conspiracy theorists that enjoy the loyal support of thousands of supporters worldwide. In this context, the post appeared to present a decisive ‘treatment’ to kill the virus with the 20-20-20 plant fertilizer. As a vaccine for COVID-19 had not been developed at the time, it appears unclear which vaccine the user is referring to.



INTENT

The intent of the user appears to encourage users to order the 20-20-20 plant fertilizer from the website linked in the video and administer it to themselves and their family members to kill the virus. The user intended to warn users that vaccines were dangerous and had viruses “put into” them.

Although the user’s account was eventually suspended by Twitter, secondary sources suggest that he had over 17.2K followers at the time of being

¹⁶² Archived Tweet by @chiefpolice2 obtained from waybackmachine.org, <https://web.archive.org/web/20200129054922/https://twitter.com/chiefpolice2/status/1222066290040307712>.

¹⁶³ AJMC: ‘A Timeline of COVID-19 Developments in 2020’ (2021), <https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020>.



removed.¹⁶⁴ The impact of the video could have been that users ordered and self-administered a plant fertilizer unauthorized for medical treatment. Most users on the post appeared to accept the claim and sought to administer the prescribed spray. A user enquired that he tried to order the product but the link appeared to be broken. Other users on the post commented saying that he would try it as even if it does not “cure” his heart issues, it would be worth it if it could improve his breathing issues.

Article 19(3) ICCPR lays down that speech might be reasonably restricted on the grounds of protection of public health. As the post was encouraging Twitter users to self-administer a plant fertilizer to themselves and their family and thus cause real-world imminent harm to their health, the post could reasonably be restricted under Article 19(3).

Disinformation and violence after the 2020 US Elections

Around voting day in the 2020 US Presidential Elections, a Facebook user claims in a #StopTheSteal group that “conservative poll watchers were not being permitted” and another user responds by asking if other members were “going to give up without a fight” and asks them to “gather up arms and meet at the election headquarters” quickly.

Analysis of this case under IHRL



Around November 4, 2020, multiple groups titled ‘Stop The Steal’ emerged on Facebook, many Tweets used the #StopTheSteal hashtag and many subreddits appeared that contained false claims about the voting process and certification of the results of the election.¹⁶⁵ A report by the Tech Transparency Project documents one of these posts where a Facebook user claims that “conservative poll watchers were not being permitted” to view the counting of votes in swing states. Another user then asks, “are you guys going to give up without a fight, I thought we were Americans”. The second user says “who’s with me we can’t let this injustice happen” and asks people to “gather up arms and meet at the election headquarters quickly”.¹⁶⁶ Claims about election watchers not being permitted to view the counting of votes were verifiably false and had been fact-checked by authorities throughout the election cycle.¹⁶⁷

¹⁶⁴ Chiefpolice2’s account suspended, <https://twitter.com/chiefpolice2>.

¹⁶⁵ Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here, BuzzFeed News, <https://www.buzzfeednews.com/article/ryanmac/full-facebook-stop-the-steal-internal-report>.

¹⁶⁶ Capitol Attack was months in the making on Facebook, Tech Transparency Project, <https://www.techtransparencyproject.org/articles/capitol-attack-was-months-making-facebook>

¹⁶⁷ Trump’s wrong claim that election observers were barred in Pennsylvania, Michigan, Politifact,

<https://www.politifact.com/factchecks/2020/nov/12/donald-trump/trumps-wrong-claim-election-observers-were-barred-/>; Over 100



The post was made in the context of the US Presidential Elections 2020. In the days leading up to voting day on November 4, platforms including Facebook, Twitter, Reddit, and others had reported that various groups had emerged that were spreading election-related misinformation and openly inciting violence. The Facebook group in which this post was made appeared to be a popular ‘#StopTheSteal’ group with over 338K votes. There were also reports that election officials around the country, especially in swing states, were being threatened and intimidated with physical harm.



Although the exact intent of the user cannot be decisively determined, we could try to estimate the second user’s intent by looking at the totality of the circumstances and the tone of his comments in the post. The second user’s tone does not seem to suggest that he is urging his followers to peacefully ensure that election watchers could observe the counting of votes. His opening comment suggestively asks if they were ready to give up without a “fight” and stirs up people by asking “who’s with me we can’t let this injustice happen”. Finally, his incitement to imminent violence appears clearer as he asks to bring their guns and gather quickly. The call to bring arms, as well as the sense of immediacy in his tone make it clear that his intention was: 1) either intimidate poll workers at counting station; or 2) actually cause viol violence at ing booths, possibly even gun violence. In both of these cases, the intention does not appear to promote democratic discourse but appears to promote violence and intimidation.



Given the vast membership of the group with over 338K followers, the disinformation around denial to access of counting booths, the inciteful tone of the post, the tense atmosphere around the elections with weeks of violent rhetoric, and the easy access of assault weapons in the US, there was a high likelihood that the impact of this post could have been group users gathering at a vote counting booth and indulging in violence.

Article 19(3) ICCPR lays down that speech might be reasonably restricted on the grounds of protecting public order. Further, the right to participate in elections and express one’s political will is safeguarded by Article 25. As a result of the objectively false claims about election fraud, the incitement to “fight” and the imminent call to bring arms, the removal of the post would have been compatibly acceptable under Article 19(3).



Chapter Conclusions

This chapter highlighted the various issues that may exist in dealing with disinformation. It looked at foundational questions around devising an appropriate definition of disinformation, envisaged what a human rights-centred approach to dealing with disinformation might look like, devised a test or threshold for disinformation and, finally, applied this test to real-life case studies.

There is a strong argument to be made that social media platforms should adopt a cautious approach in dealing with disinformation. If social media platforms were to adopt overbroad policies, it might lead to the systematic removal of legitimate information and opinion and unreasonably restrain the right to expression under Article 19 ICCPR.

As the EU HLEG on Disinformation also noted in March 2021, simplistic solutions to disinformation should be disregarded. Any form of censorship, both public and private, must be viewed with caution and avoided as far as possible.¹⁶⁸ Tech-oriented solutions to deal with misinformation need to focus on reducing incentives to produce and circulate disinformation and changing the product design of algorithms to create more friction, so that disinformation does not go viral. Moreover, platforms should also work to build cross-platform cooperation, so that disinformation can be dealt with at an ecosystem level. As highlighted previously in the report, misinformation cannot be tackled by IHRL-compliant Terms alone. The focus needs to be on a hybrid approach that adopts IHRL-compliant Terms and other measures suggested in the Code of Practice on Disinformation of the EU, such as promoting media literacy, transparency in political advertising, labelling and downranking. Lastly, as the Inter-American Commission on Human Rights and the SRFOE have also noted, any approach by platforms to deal with disinformation must be accompanied by maximum transparency with respect to company policies, and platforms must engage in ongoing due diligence to determine their policies' impact on freedom of expression.¹⁶⁹

¹⁶⁸ European Commission, 'Report of the independent High level Group on fake news and online disinformation' (2018), <<https://www.ecsite.eu/sites/default/files/amulti-dimensionalapproachtodisinformation-reportoftheindependenthighlevelgrouponfakenewsandonlinedisinformation.pdf>>

¹⁶⁹ Catharine Christie, Edison Lanza & Michael Camilleri, 'Covid-19 and Freedom of Expression in the Americas' (2020), The Dialogue, <<https://www.thedialogue.org/wp-content/uploads/2020/08/Covid-19-and-Freedom-of-Expression-in-the-Americas-EN-Final.pdf>>

Concluding comments

The Internet is the most revolutionary breakthrough in communications technology since the printing press. In theory, the Internet should have made free speech invincible, banishing censorship to the ash heap of history. Yet, unmediated access to free and equal speech has caused resurgent autocracies to fight back and democracies to think twice about whether the Internet should be seen as more of a blessing than a curse. Due to the process of “platformization”, dominant platforms are caught between their initial techno-optimistic impulses towards free speech maximalism and the techno-dystopian backlash fueled by their facilitation of the dark sides of free speech, including disinformation and hate speech. Consequently, the practical exercise of free speech on privately-owned platforms has become subject to ever more restrictive terms and state regulation. This has left users vulnerable to opaque content moderation with little transparency and no overarching principles for how to reconcile the fundamental value of freedom of expression with online harms capable of justifying restrictions of free expression and information. As demonstrated in this report, neither the Terms nor the practical content moderation of major platforms conform to IHRL norms on freedom of expression. Consequently, much content is purged that would otherwise be protected under IHRL (while also leaving some content – including incitement to violence - in place that should be removed). This has real consequences for the practical exercise of freedom of expression, particularly in countries where social media are the only alternative to traditional media, which are dominated by official propaganda and censorship.

To be compliant with IHRL, platform content moderation practices must be legitimate, necessary and proportional and must occur within the framework of one of the grounds set forth in Article 19(3) ICCPR. With respect to hate speech, platforms should frame Terms based on an Article 20(2) ICCPR threshold and must strictly take into consideration the RPA’s six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm before taking enforcement action.

For disinformation, platform Terms must be tailored to protect the grounds in Article 19(3) ICCPR and Article 25 ICCPR, and platforms must refrain from adopting vague, blanket policies for removal. Only disinformation entailing real and immediate harm should be subject to the most intrusive restrictive measures such as content removal; whereas other forms of misinformation may be subject to less restrictive measures such as labelling or downranking. In determining the limits of disinformation at the enforcement stage, platforms must focus on the post’s content, its context, the speaker’s intent, its impact and its likelihood of causing imminent harm.

To ensure a more viable future for free speech on social media, Justitia believes that turning to IHRL as a “framework of first reference” will provide more clarity and legitimacy to content moderation on major social media platforms – particularly, when it comes to the contested categories of hate speech and disinformation, as set forth above. Accordingly, Justitia recommends that major platforms – starting with those that have signed up to the EU’s Code of Conduct on Illegal Hate Speech and Code of Practice on Disinformation – formally commit to adopting an IHRL approach to content

moderation by signing a voluntary Free Speech Framework Agreement (FSFA). The FSFA would not require the harmonization of Terms or content moderation practices among platforms themselves. The content of the FSFA would focus on issues such as how to develop Terms that reflect IHRL norms, how to integrate IHRL into human and automated content moderation, and how to ensure transparency and accountability vis-à-vis the commitment to an IHRL approach. While legally non-binding, Justitia proposes that the FSFA be administered by the Office of the UN High Commissioner for Human Rights (OHCHR) under the specific auspices of the SRFEO.

The FSFA could also address how to combine an IHRL approach with greater user control over content in order to enable decentralized content moderation. This will allow users to avoid content that may be permissible under IHRL but deeply offensive to them. The content of the FSFA should be agreed upon in a series of meetings with the participating platforms, the OHCHR and relevant subject-matter experts such as the SRFOE, regional institutions (including but not limited the OSCE, the OAS and the ACHPR) as well as participants from civil society. For purposes of overcoming paternalistic approaches to content moderation and steering the ship in the direction of user empowerment, the FSFA could be made available for public consultation. The FSFA should require platforms to provide the OHCHR access to relevant data in order to assess overall compliance, issue recommendations that address best practices and to identify shortcomings regarding platform compliance with the FSFA.

We acknowledge that an IHRL approach and the proposed FSFA will not address or solve the entire multitude of challenges that come with private online content moderation and that IHRL as a “framework of first reference” will create new challenges and dilemmas. However, we hope to have demonstrated that grounding content moderation on IHRL as a “framework of first reference” offers a practical way forward to approximate the jurisprudential, jurisdictional and perceptual variations on the issue of freedom of expression on global private platforms within a set of widely-accepted global norms.



**THE
FUTURE
OF
FREE
SPEECH**

**REBUILDING THE
BULWARK OF LIBERTY**

