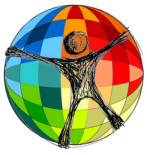
INTERNATIONAL HUMAN RIGHTS INTERNSHIP PROGRAM | WORKING PAPER SERIES

VOL. 11 | NO. 1 | SUMMER 2022

Fueling the Flames: Online Social Platform Responsibility for Harmful Content in Conflict Zones

Charlotte Ridsdale

McGill Centre for Human Rights and Legal Pluralism



Centre sur les droits de la personne et le pluralisme juridique de McGill

McGill FACULTY OF Law

ABOUT CHRLP

Established in September 2005, the Centre for Human Rights and Legal Pluralism (CHRLP) was formed to provide students, professors and the larger community with a locus of intellectual and physical resources for engaging critically with the ways in which law affects some of the most compelling social problems of our modern era, most notably human rights issues. Since then, the Centre has distinguished itself by its innovative legal and interdisciplinary approach, and its diverse and vibrant community of scholars, students and practitioners working at the intersection of human rights and legal pluralism.

CHRLP is a focal point for innovative legal and interdisciplinary research, dialogue and outreach on issues of human rights and legal pluralism. The Centre's mission is to provide students, professors and the wider community with a locus of intellectual and physical resources for engaging critically with how law impacts upon some of the compelling social problems of our modern era.

A key objective of the Centre is to deepen transdisciplinary collaboration on the complex social, ethical, political and philosophical dimensions of human rights. The current Centre initiative builds upon the human rights legacy and enormous scholarly engagement found in the Universal Declaration of Human Rights.

ABOUT THE SERIES

The Centre for Human Rights and Legal Pluralism (CHRLP) Working Paper Series enables the dissemination of papers by students who have participated in the Centre's International Human Rights Internship Program (IHRIP). Through the program, students complete placements with NGOs, government institutions, and tribunals where they gain practical work experience in human rights investigation, monitoring, and reporting. Students then write a research paper, supported by a peer review process, while participating in a seminar that critically engages with human rights discourses. In accordance with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded. Therefore, papers in this series may be published in either language.

The papers in this series are distributed free of charge and are available in PDF format on the CHRLP's website. Papers may be downloaded for personal use only. The opinions expressed in these papers remain solely those of the author(s). They should not be attributed to the CHRLP or McGill University. The papers in this series are intended to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s).

The WPS aims to meaningfully contribute to human rights discourses and encourage debate on important public policy challenges. To connect with the authors or to provide feedback, please contact human.rights@mcgill.ca.

ABSTRACT

The regulation of online harms, including disinformation, propaganda, hate speech, and incitement to violence, faces two core challenges: first is the challenge of striking the proper balance between freedom of expression and harmful speech. Second is the challenge of regulating a borderless, transnational online sphere. In conflict zones, these challenges are especially pressing, as harmful online content can incite or prolong violence. This paper seeks to identify possible solutions for curbing online harms in conflict zones, with a focus on Online Social Platform (OSP) responsibility.

Part I assesses International Human Rights Law's current approach to the moderation of online harms, concluding that international and regional laws are incapable of adapting to the pace of content-sharing on the internet. Part II draws on two case studies that demonstrate the need for OSPs such as Facebook and Twitter to take more action against harmful online content in conflict zones: the case of ethnic cleansing in the Rahkine province of Myanmar; and the recent conflict in the Tigray region of Ethiopia. Part III outlines possible solutions for curbing online harms, focusing on OSP self-regulation.

CONTENTS

"Mass communication, in a word, is neither good nor bad; it is simply a force and, like any other force, it can be used either well or ill. Used in one way, the press, the radio and the cinema are indispensable to the survival of democracy. Used in another way, they are among the most powerful weapons in the dictator's armory."

–Aldous Huxley, Brave New World Revisited

Introduction

There is an inherent tension when it comes to freedom of expression during conflict. On one hand, parties to an armed conflict tend to perpetuate it by spreading misleading information about the other side. On the other hand, a free flow of information can save lives during conflict, so it is precisely in those situations that freedom of expression should be most ardently defended. To balance these competing interests, restrictions on freedom of expression during conflict should be scrutinized.[1]

The advent of the internet has presented new challenges to freedom of expression, as it has created a platform for individuals to share information and ideas on a global scale. This new era of sharing has proved its efficacy at enacting political change through public will, in both pro- and anti-democratic ways. For example, widespread movements for democracy during the Arab Spring were primarily instigated through social media.[2] However,

[1] See Article 19, "Response to the consultation of the UN Special Rapporteur on Freedom of Expression on her report on challenges to freedom of opinion and expression in times of conflicts and disturbances" (19 July 2022) at 1, online (pdf): <ohchr.org/sites/default/files/documents/issues/expression/cfis/conflict/2022-10-07/submission-disinformation-and-freedom-of-expression-during-armed-conflict-UNGA77-cso-article19.pdf> [Article 19, "Response to the consultation"].

[2] See e.g. Kristen McTighe, "A blogger at Arab Spring's genesis", *New York Times* (12 October 2011), online: <nytimes.com/2011/10/13/world/africa/a-blogger-at-arab-springs-genesis.html>. *Contra* Haythem Guesmi, "The social media myth about the Arab Spring", *Al Jazeera* (27 January 2021), online:

authoritarian regimes have been known to maintain their control through social media campaigns[3] and online surveillance with repression of dissidents.[4] While a number of possible state interventions have been proposed in order to maintain a democratic internet, many regulations do not foresee the rate at which online campaigns can cause harms, particularly in conflict zones. Instead, many suggest increasing responsibility of Online Social Platforms (OSPs).

The 2022 report of the UN Special Rapporteur on Freedom of Expression focused on this very topic, and received submissions from civil society groups, tech companies, and stakeholders around the world.[5] The report offers recommendations on how to address issues such as disinformation, misinformation, propaganda, and incitement to violence, specifically in the online sphere.

This work will expand on the evolving discourse over how best to designate responsibility over regulation of content on OSPs. In particular, it will focus on the risks of harmful online content[6] in countries experiencing conflict and explore potential solutions. I will begin by outlining the current state of international human rights law (IHRL) as it pertains to freedom of expression and harmful online content, examining limitations in addressing the challenges of the borderless internet. In Part II I will examine the need for OSPs to take more action against harmful online content, especially in conflict zones, by drawing on two case studies: ethnic cleansing in the Rakhine province of Myanmar, and recent violence in the Tigray region of Ethiopia. Part III seeks to identify possible solutions for curbing online harms, looking at current

---

<aljazeera.com/opinions/2021/1/27/the-social-media-myth-about-the-arab-spring>.

[3] See e.g. Marcel Schliebs et al*, China's Inauthentic UK Twitter Diplomacy: A Coordinated Network Amplifying PRC Diplomats* (Oxford, UK: Programme on Democracy & Technology, 2021).

[4] See e.g. *The Law of the People's Republic of China on Safeguarding National Security in the Hong Kong Special Administrative Region*, 2020.

[5] See Irene Khan, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 77th Sess, UN Doc A/77/288 (2022).

[6] I define harmful online content as hate speech, propaganda, disinformation, and misinformation online.

efforts and how they can be improved and adapted to the online sphere during conflict situations. Throughout the paper I underscore the incapacity of international and regional laws to adapt to the pace of content-sharing on the internet. Instead, I conclude that to prevent online harms in conflict zones, OSPs must take a self-regulatory approach that is grounded in IHRL principles.

I. Online Harms Under International Human Rights Law

Individual freedom of expression must always be balanced against other rights. IHRL addresses this balance generally, but there are concerns about its operation in the realm of online expression. As the UN Special Rapporteur Irene Khan wrote, "social media platforms are highly susceptible to the spread of disinformation, propaganda and incitement."[7] Harmful online content can incite or prolong violence in times of conflict; as a result, this point of balance must sometimes be adjusted to prevent the free flow of disinformation and hate speech. Despite the need for clear guidance in this digital space, IHRL does not translate easily to offer protections for human rights in the online sphere. To explain this, I will begin by outlining the state of international law as it pertains to freedom of expression.

Prior to the rise of the internet, numerous international and regional conventions were developed to balance the right to freedom of expression with the need to address harmful speech, and extensive literature already exists on the subject. Briefly, at the regional level, the African Charter on Human and Peoples' Rights, the American Convention on Human rights and the European Convention on Human Rights all safeguard freedom of expression.[8] The International Covenant on Civil and Political Rights (ICCPR) is a key international convention that recognizes

[7] Khan, *supra* note 5 at 19.

[8] *African Charter on Human and Peoples' Rights*, adopted 27 June 1981, OAU Doc CAB/LEG/67/3 rev 5, 21 ILM 58 (1982), art 9; *American Convention on Human Rights*, signed 22 November 1969, OASTS No 36, 1144 UNTS 123, art 13; Council of Europe, *Convention for the Protection of Human Rights and Fundamental Freedoms*, Rome, 4 XI 1950, art 10 [ECHR].

freedom of expression as a fundamental human right, but also acknowledges that it can be restricted in certain circumstances.[9] Article 19(2) of the ICCPR guarantees individuals the right to seek, receive, and impart information and ideas of all kinds, regardless of borders and through any form of media.[10] However, the three-part test in article 19(3) sets out the conditions that restrictions may be imposed, requiring that they be lawful, necessary, and proportionate.[11] Similarly, the European Convention on Human Rights (ECHR) recognizes that limitations on freedom of expression should be "prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health, or morals, for the protection of the reputation or rights of others."[12]

The importance of this balancing act is essential in democracies; the Human Rights Committee has called article 19 of the ICCPR "the foundation stone for every free and democratic society."[13] However, article 19 is somewhat in conflict with article 20(1) of the ICCPR which obligates states to prohibit propaganda for war, and article 20(2) which prohibits "advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence."[14] States have had to grapple with the types of expression that they should not permit in a liberal democracy, versus the types that should be allowed for the sake of an open exchange of ideas.[15]

One proposed solution is offered in the Rabat Plan of Action. The document provides detailed guidance on how states can distinguish between the kinds of expression that should be

---

[9] *International Covenant on Civil and Political Rights*, 19 December 1966, 999 UNTS 171 [*ICCPR*].

[10] *ICCPR*, *ibid*, art 19(2).

[11] *ICCPR*, *ibid*, art 19(3); permissible restrictions are only those that are "provided by law"; the restriction needs to be for the purpose of a legitimate interest; and the restriction must be necessary and the least restrictive way to achieve the aim.

[12] ECHR, *supra* note 8.

[13] Human Rights Committee, *General comment no 34*, 102nd Sess (2011).

[14] *ICCPR*, *supra* note 9, art 20.

[15] See e.g. First Amendment case law in the USA.

prohibited, and those that should be protected. The plan suggests a high threshold for defining restrictions on freedom of expression and for the application of article 20 of the ICCPR. It also outlines a six-part threshold test for determining hate speech likely to incite violence, therefore justifying restrictions on of freedom of expression; "(1) the social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination and (6) likelihood of harm, including imminence."[16] Although this test offers useful guidance, the imminence and intent criteria may be problematic. If harm is imminent, it may be too late for states or corporations to respond. The intent requirement might also be a too restrictive, as unintended harms are also possible.[17]

Because these conventions were drafted before the internet entirely re-shaped global communications, I posit that IHRL is not sufficient for addressing online speech on its own, especially not in conflict zones. Instead, an approach that places greater responsibility on OSPs for the regulation of online harm is necessary.

Prior to this discussion it is important to note that IHRL primarily targets states and their obligations to protect human rights, while OSPs are private actors whose main objective is to maximize profits. While OSPs have a responsibility to respect human rights, including the right to freedom of expression, they are not held to the same standards as states. Therefore, the regulation of harmful online content should be viewed as a shared responsibility between states and OSPs, with both having a role to play in protecting human rights online.

A. Freedom of Expression Online

[16] *Freedom of Expression vs. incitement to hatred: OHCHR and the Rabat Plan of Action*, OHCHR, 22nd Sess, UN Doc. A/HRC/22/17/Add.4, Appendix, adopted 5 October 2012.

[17] See Jennifer Easterday, Hana Ivanhoe & Lisa Schirch, "Comparing Guidance for Tech Companies in Fragile and Conflict-Affected Situations" (2022) at 15, online (pdf): *TODA Peace Institute* <toda.org/policy-briefs-and-resources/policy-briefs/comparing-guidance-for-tech-companies-in-fragile-and-conflict-affected-situations.html>.

Before delving into issues surrounding OSP responsibility for online harms, it is important to outline the state of international law on freedom of expression online, the risks associated with a lawless internet, and the challenges of regulating the online sphere.

Although free media is the backbone of a democratic society, media has long been manipulated to cause harm and incite violence in conflict situations.[18] OSPs and other digital technologies pose new threats globally, as they host billions of active users,[19] allowing false and dangerous speech to proliferate and spread at an unprecedented rate and scale. Despite this, none of the international conventions discussed address human rights online. Noting this gap, in 2018 the UN Human Rights Council passed a resolution on "the promotion, protection and enjoyment of human rights on the Internet," which affirms "that the same rights that people have offline must also be protected online, in particular freedom of expression."[20] This is a positive step; however, the resolution is non-binding and issues of state implementation remain.[21] There remains ambiguity in IHRL over legal classifications of content. In particular, what constitutes hate speech and disinformation.

i) Hate Speech and Disinformation

There is no universally accepted definition of hate speech,[22] because there are varying regional interpretations of free speech,

---

[18] For example, the use of radio during the Rwandan genocide. See e.g. David Yanagizawa-Drott, "Propaganda vs. Education: A Case Study of Hate Radio in Rwanda" in Jonathan Auerbach & Russ Castronovo, eds, *The Oxford Handbook of Propaganda Studies* (Oxford: Oxford University Press, 2013) 378.

[19] 4.59 billion in 2022; see S Dixon, "Number of social media users worldwide from 2018 to 2027" (16 September 2020), online: *Statistica* <statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=In%202021%2C%20over%204.26%20billion,almost%20six%20billion%20in%202027>.

[20] *The promotion, protection and enjoyment of human rights on the Internet*, UNHRC, 47th Sess, UN Doc A/HRC/RES/47/16 (2021).

[21] See Article 19, "Human Rights on the Internet" (15 July 2021), online: <article19.org/resources/un-human-rights-council-adopts-resolution-on-human-rights-on-the-internet/>.

[22] See *General Recommendation No 32 on The Meaning and Scope of Special Measures in the International Convention on the Elimination of Racial Discrimination*, CERD/C/GC/32 (2009) at 9.

and different conceptions of harm around the world. As noted by many scholars, "hate speech lies in a complex nexus with freedom of expression, group rights, as well as concepts of dignity, liberty and equality" and thus any objective definition of hate speech can be contested.[23] The ambiguity of these terms makes it difficult for states to prohibit them. To complicate the matter, international human rights conventions on freedom of expression did not foresee the potential for private citizens to have the capacity to conduct effective hate campaigns independently from the state.[24] As a result, the current international framework does not effectively regulate and prohibit online hate speech, despite that fact that it is linked to real violence[25] and that this phenomenon is exacerbated in conflict zones.[26]

As with hate speech, there is no clear, agreed-upon definition of 'disinformation.' Different national laws and regional standards use different definitions.[27] A variety of power-holders have abused the term to discredit media outlets and to suppress dissenting opinions.[28] Still, IHRL does not prohibit the term *per se*. [29] Although this area of IHRL is murky, the risks of

---

[23] Leandro Silva et al, *Analyzing the Targets of Hate in Online Social Media*, *Proceedings of the Tenth International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Web and Social Media* (2016) at 688.

[24] The historical tragedies that occurred during the Armenian Genocide and the Holocaust influenced the drafting of the Universal Declaration on Human Rights and the ICCPR, and as a result IHR obligations outlawing hate speech were targeted towards states. See Tiran Rahimian, *Whither International Law in Online Content Moderation?* (Bachelor of Civil Law & Juris Doctor, McGill University, 2019) [unpublished] at 10.

[25] See e.g. United Nations Secretary-General, "UN Strategy and Plan of Action on Hate Speech online" (2019), online (pdf): <ohchr.org/en/special-procedures/sr-religion-or-belief/hate-speech-and-incitement-hatred-or-violence>.

[26] See Khan, *supra* note 5 at 4.

[27] See e.g. Council of Europe, *Information Disorder: Toward an interdisciplinary framework for research and policy making*, DGI (27 September 2017), which makes a distinction between "misinformation" (when false information is shared, but no harm is meant), and "disinformation" (when false information is shared knowingly to cause harm). See also definitions of the European Commission, *A multi-dimensional approach to disinformation* (Luxembourg: Publications Office of the European Union, 2018) at 10.

[28] See Article 19, "Response to the consultation", *supra* note 1 at 4.

[29] See *ibid* at 5. States are under an obligation pursuant to art 20 of the ICCPR to only restrict disinformation that doesn't meet the 3-part test under art 19(3) of

disinformation should not be understated, especially during conflict. Distorted understandings of a conflict situation put civilians at heightened risk of physical and emotional harms and can increase violence.[30]

Because they cannot rely on IHRL to define these terms, OSPs have been called upon to develop their own by-laws regarding prohibited content. Hate speech and disinformation online have definitions that vary depending on the platform. For example, YouTube's terms of service define hate speech as "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, or veteran status."[31] Meta's community standards define hate speech as "a direct attack against people – rather than concepts or institutions – on the basis of protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease."[32] These definitions offer a much-needed level of precision not provided for in international law.

Undeniably, issues beyond legal ambiguity create challenges for regulation of online content, the most significant being that of transnationality.

B. Intermediary Liability and the Challenge of Transnationality

In recent years we have seen an alarming trend of disinformation campaigns aimed at shaping public opinion,[33]

---

the ICCPR. Generally, it cannot be prohibited under IHRL unless it amounts to incitement to hostility, violence, or discrimination.

[30] See Khan, *supra* note 5 at 24.

[31] Google Support, "YouTube's Policies: Hate Speech Policy" (2022), online: YouTube <support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436>.

[32] Meta's Transparency Centre, "Facebook Community Standards: Hate Speech" (2022), online: <transparency.fb.com/en-gb/policies/community-standards/hate-speech/>.

[33] See e.g. Davey Alba & Adam Satariano, "At least 70 countries have had disinformation campaigns, study finds", *New York Times* (26 September 2019),

interfering with elections,[34] and increasing the intensity of online hate speech.[35] Opponents of these trends have highlighted that part of the challenge in fighting them is jurisdictional. The human rights obligations of states are generally limited to parties that are within their jurisdiction, but online speech reaches globally with little regard for borders.[36] The transnational nature of the internet creates two key challenges of governance: those of jurisdiction, and those of intermediary liability. I will review these in the following section.

i) Jurisdiction

Central to the challenge of online governance is the incompatibility between the borderless nature of online hate speech and the scope of IHRL. Effectively, IHRL only creates obligations for states,[37] leaving OSPs to be regulated domestically. Major tech platforms are not states, but now wield unprecedented power over individual rights and public discourse; this differentiates OSPs from corporations of decades past. Some tools of international law apply to OSPs, the most robust being the UN's Guiding Principles on Business and Human Rights. It creates soft-law regulations for corporations, imploring them to follow international human rights frameworks in their operations.[38] Broadly speaking the Guiding Principles outline hold that "companies have a responsibility to respect internationally-recognized human rights and to conduct their operations in ways that avoid causing or contributing to 'adverse human rights

---

online: <nytimes.com/2019/09/26/technology/government-disinformation-cyber-troops.html>.

[34] See e.g. Freedom House, Press Release, "Digital Election Interference Widespread in Countries Across the Democratic Spectrum" (7 December 2020), online: *Freedom House* <freedomhouse.org/article/report-digital-election-interference-widespread-countries-across-democratic-spectrum>.

[35] See e.g. Fernand de Varennes, *Recommendations made by the forum on minority issues at its thirteenth session on the theme,"Hate speech, social media and minorities"*, OHCHR, 46th Sess, A/HRC/46/58 (26 January 2021).

[36] Of course, each country has their own national regulation of internet, with countries approaches as varied as China's vs. USA's.

[37] Article 20 of the ICCPR requires a prohibition of certain forms of hate speech by law, thus speaking directly to states. See Rahimian, *supra* note 24 at 23.

[38] See *Guiding Principles for Business and Human Rights: Implementing the United Nations "Protect, Respect, and Remedy" Framework*, OHCHR (2011).

impacts' and to prevent or mitigate such impact."[39] Civil society and stakeholders have voiced some concern over the enforceability, vagueness, and broadness of the Guiding Principles for tech companies specifically,[40] and for good reason; the approach taken by OSPs to tackle harmful content online has been fragmented as a result of the lack of thrust of IHRL, the Guiding Principles included.[41]

ii) Properly Imposing Intermediary Liability

Many states have responded to the challenge of regulating harmful online content by adopting laws that impose strict liability on OSPs who host that content. For example, the German Network Enforcement Act (NetzDG) allows users to flag content that they believe to be illegal and obliges platforms to remove the "violating content" within a short time or face significant financial penalties.[42] Canada and the United Kingdom have tabled similar legislation.[43] On its face these efforts seem positive, however there is a danger of granting authorities excessive discretionary powers to moderate content, to the point of political censorship.

---

[39] See *Khan,* supra note 5.

[40] *Report of the Office of the United Nations High Commissioner for Human rights: The practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies*, OHCHR, 55th Sess, UN Doc A/HRC/50/56 (2022).

[41] See *Khan* supra note 5 at para 85. Each OSP has different policies regarding the kinds of content that they take down, and appeal processes for content that is removed or devalued: "This fragmented approach fails to provide much-needed coherence and predictability to platform practice and has the potential to undermine company compliance with international human rights law."

[42] See *The Network Enforcement Act*, BD, 12 June 2017 [*Network Enforcement Act*].

[43] See Government of Canada, Canadian Heritage, *Discussion Guide (Consultation closed: The Government's proposed approach to address harmful content online)* (2021), online: <canada.ca/en/canadian-heritage/campaigns/harmful-online-content/discussion-guide.html#a1>; *Online Safety Bill* (UK), 209 2022-23. See Appendix.

For example, in Pakistan [44] and the Russian Federation, [45] legislation criminalizing fake news or anti-government sentiments have been passed under the guise of preventing harm but without meeting the exigencies set out in article 19(3) of the ICCPR.[46] Further guidance in designing IHRL-compliant regimes for internet intermediaries and OSPs is needed.

Although a useful framework, IHRL is a limited one in the area of online content. As discussed, it is limited in its enforceability for OSPs and other non-state actors, as there is a large degree of indeterminacy within IHRL norms that would leave platforms with discretionary powers in most hard cases.[47] Without strong top-down accountability, it is in the interest of all platform users for OSPs themselves to put in place strong transparency and accountability mechanisms around content moderation. Moreover, many states have begun to impose legal responsibility on OSPs for their moderation decisions.[48] If anything, this should reinforce the notion that platforms must be proactive about their policies. On top of all these challenges, as long as hate speech and disinformation remain undefined in international law, there will be

---

[44] See *Citizens Protection (Against Online Harm) Rules*, 2020. Imposes penalties of up to 3.2 million USD on social media companies who do not remove "harmful content" within 6-24 hours, but the "harmful content" provision is overly broad. See David Kaye & Michel Forst, *Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the situation of human rights defenders*, OL PAK 3/2020 (19 March 2020).

[45] See *Law on Information, Information Technologies and Information Protection*, Federal Law No 483-FZ0-2020, 2020. Individuals may be fined up to 22,900 USD for disseminating "fake news" and content that shows "blatant disrespect for society, government, official government symbols, constitution or governmental bodies of Russia." See David Kaye, *Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, OL RUS 4/2019 (1 May 2019).

[46] *ICCPR*, *supra* note 9, art 19(3) highlights that freedom of expression is subject to restrictions, but only if they are provided by law and are necessary for the respect of the rights of others, or for the protection of national security, public order, or public health.

[47] See Evelyn Douek, "Limits of International Law in Content Moderation" (2021) 6:1 UC IJIL 37 at 51 [Douek, "Limits"].

[48] See e.g. *Network Enforcement Act*, *supra* note 42; European Commission, *Digital Services Act*, [2022] OJ 2022/2065 of 19 October 2022 L277/1 [*DSA*]. See Appendix.

ambiguity over the types of expression that should be prohibited online. As I will discuss in Part III, self-regulation of OSPs is necessary until IHRL can develop standards and accountability mechanisms sufficient to handle these challenges. This is especially needed in conflict zones where calls for violence are rampant online. In the next section, I will investigate the role of OSPs in perpetuating harms by looking at the cases of Myanmar and Ethiopia.

II. Responsibility of Online Platforms During Conflicts and Other Violent Situations

All of the most widely-used OSPs in the world are based in the United States.[49] This is significant because third-party liability law in the US is uncommonly permissive, and has been linked to the lack of regulation globally and harms arising from abuse of platforms' tolerance.[50] Section 230 of the Communications Decency Act (CDA) was enacted in 1996 and creates broad immunity from liability for internet intermediaries including OSPs for content they host.[51] As Bowers and Zittrain write, "what might have otherwise been a decades-long process of common law development aimed at defining the specific contours of platform liability with relation to harmful content was instead determined in short order."[52] Up to this point, CDA 230 has essentially made it impossible to legally compel platforms to police harmful content. This may be changing as advocates and lawmakers call for increased OSP accountability, specifically in a Supreme Court

[49] The 3 OSPs with the largest number of users are Facebook, with 2.9 billion, Youtube, with 2.56 billion, and WhatsApp, with 2 billion users in 2022. See S Dixon, "Global social networks ranked by number of users 2022" (26 July 2022), online: *Statistica* <statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.

[50] See John Bowers & Jonathan Zittrain, "Answering Impossible Questions: Content Governance in the Age of Disinformation" (2020) 1:1 Harv Kennedy Sch Misinformation Rev 1 at 2.

[51] *Communications Decency Act*, 47 USC 230, s 230.

[52] Bowers & Zittrain, *supra* note 50.

case currently on the docket.[53] The court will consider whether Google can be liable in part for the 2015 ISIS-led attacks that killed Nohemi Gonzalez, a 23-year-old American student in Paris.[54] This case, alongside a similar case against Twitter,[55] addresses head-on the legal questions surrounding liability for recommender systems that increase recruitment into extremist groups and result in violence.

These cases are specific to the United States, however similar questions arise in other jurisdictions, because these OSPs are American corporations who are granted immunity under CDA 230 (for now). In this section, I seek to understand what the repercussions of this First Amendment value system are when applied to other socio-cultural, economic, and political contexts.

A. *Myanmar*

Since the 2018 UN report on the situation of human rights in Myanmar, the global community has begun to recognise the "determining role" of Facebook in the atrocities against Rohingya Muslims.[56] For example, NGOs such as Amnesty International have denounced Facebook (now Meta) for its role in intensification of the conflict that led to serious human rights violations perpetrated against the Rohingya.[57] The UN's Independent International Fact-Finding Mission on Myanmar called for senior military officials to be investigated and prosecuted for war crimes, crimes against humanity, and

[53] See e.g. Jeffrey Neuburger, "Important CDA Section 230 Case Lands in Supreme Court" (6 October 2022), online: *The National Law Review* <natlawreview.com/article/important-cda-section-230-case-lands-supreme-court-level-protection-afforded-modern>.

[54] See *Gonzalez v Google*, No 18-16700 (9th Cir 2021).

[55] See *Twitter Inc. v Taamneh*, No 21-1496 (9th Cir 2022).

[56] *Report of the Special Rapporteur on the situation of human rights in Myanmar, Advance Unedited Version*, UN Doc A/HRC/37/70 (9 March 2018) at para 65.

[57] See Amnesty International, "Social Atrocity: Meta and the right to remedy for the Rohingya" (2022), online (pdf): *Amnesty International* <amnesty.org/en/documents/asa16/5933/2022/en/> [*Amnesty Report*].

genocide.[58] That same report also concluded that the role of social media was significant in the atrocities that ensued. How did social media play such a large role in the violence?

The history of telecommunications in Myanmar partly explains the severity of the impact. When the internet and mobile phones became globally available around the year 2000, access to these technologies were constrained by state-sanctioned prohibitive pricing in Myanmar.[59] This restrictive policy began to shift as the government recognized that increased telecommunications infrastructure could be a conduit to international investment and economic prosperity. As an illustration of this, when then-President Thein Sein moved to open the telecommunications ecosystem in Myanmar, the U.S. government reengaged with Myanmar after over 50 years of disconnection.[60]

Following a series of constitutional amendments in 2008 allowing for the broadening of rights relating to self-expression and relaxing of media censorship, foreign companies began to invest in the construction of telecommunications infrastructure across Myanmar.[61] This resulted in an unprecedented boom in mobile usage; in 2018, there were 61 million mobile phone users, up from 6.8 million in 2013.[62] Facebook also benefited from these developments in Myanmar, as an expanding number of mobile phone providers were offering cheap social media data packages with Facebook pre-installed.[63] As a result, the number of

---

[58] *Report of the independent international fact-finding mission on Myanmar*, OHCHR, 39th Sess, UN Doc A/HRC/39/64 (2018) [*Report Myanmar*].

[59] See Jeffrey Sablosky, "Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar" (2021) 43:6 Media Cult Soc 1017 at 1024; Sophie Song, "Internet in Myanmar remains slow, unstable and affordable to less than 1% of the population", *International Business Times* (6 December 2013), online: <ibtimes.com/internet-myanmar-remains-slow-unstable-affordable-less-1-population-1402463>.

[60] See Sablosky, *ibid* at 1025.

[61] See *ibid*.

[62] See The Worldbank Data, "Mobile cellular subscriptions – Myanmar" (2022), online: <data.worldbank.org/indicator/IT.CEL.SETS?end=2020&locations=MM&start=1960&view=chart>.

[63] Part of the reason for this was that Facebook was the only social media that supported Burmese text. See Thomas Dowling, "Shooting the (Facebook)

Facebook users reached 18 million by 2018.[64] Importantly, the rapid uptake of social media use occurred in a context in which digital media literacy was extremely low, and misinformation was rampant.[65] In addition, Myanmar was at that time in the midst of a delicate transition to democracy after decades of military rule. Some hoped that Facebook could act as a vehicle for open debate and democratization. Instead, it became a tool for genocide and the collapse of liberal government in the country.

Anti-Rohingya rhetoric, policies and violence had been prevalent in Myanmar throughout its tumultuous history,[66] but the internet offered an even more efficient vehicle for hate speech and misinformation compared to other forms of media. In the years leading up to the atrocities in the Rakhine State against the Rohingya in 2017, Facebook became an echo-chamber of anti-Rohingya sentiments. Accounts linked to the Tatmadaw (Myanmar's military) and Buddhist nationalist groups made countless posts calling for explicit violence against the Rohingya, spreading disinformation about an impeding Muslim takeover of the country, and otherwise dehumanizing them. Much of the most extreme content has been documented by bodies such as the UN and Amnesty International, and was posted by actors including senior government and military officials, prominent civilian nationalist groups, and 'news' pages.[67]

The UN Fact-Finding Mission documented "over 150 public social media accounts, pages and groups [that] regularly spread messages amounting to hate speech against Muslims in general or Rohingya in particular,"[68] and recommended that "the extent to which Facebook posts and messages led to real-world

---

Messenger" (21 January 2019), online: *Tea Circle - A forum for new perspective on Myanmar* <teacircleoxford.com/essay/shooting-the-facebook-messenger-part-i/>.

[64] See BBC Trending, "The country where Facebook posts whipped up hate", *BBC* (12 September 2019), online: <bbc.com/news/blogs-trending-45449938>.

[65] See *Amnesty Report*, *supra* note 57 at 17.

[66] See *Report Myanmar*, *supra* note 58 at 6, 28.

[67] See *Report Myanmar*, *ibid; Amnesty Report*, *supra* note 57.

[68] *Report Myanmar*, *ibid* at para 131.

discrimination and violence must be independently and thoroughly examined."[69]

i) The Investigations

Amnesty International conducted investigations after the 2018 violence, their core source being interviews of hundreds of survivors, now refugees. All the refugees interviewed cited deterioration of communal relations between the Rohingya and other ethnic groups after 2012.[70] This deterioration coincided with the growth and popularity, both online and offline, of radical Buddhist nationalist groups in Myanmar. One of the most infamous groups is the Association for the Protection of Race and Religion (otherwise known as the MaBaTha). Their Facebook page had the most followers in the country from 2015 onwards, and their platform was used to spread a vehement anti-Muslim campaign throughout 2016.[71]

The connection between anti-Rohingya content online and the realization of actual violence is not fully understood. However, the UN Fact-finding mission and Amnesty International's reporting document several specific incidents in which viral Facebook posts were linked to incidents of offline violence.[72] Moreover, academic literature on genocide and mass violence more generally has highlighted the relationship between vilification, dehumanization and acts of violence.[73] Jones highlights that acts of violence can become normalized in environments where the persecuted group is socially excluded and depersonalized within the framework of a national or religious ideology.[74] This trend was also evident in the context of the Rwandan genocide. One study examined the effects of the propaganda disseminated by Radio Television Libre Miles Collines, and found that killings were 65-77% higher in

---

[69] *Ibid* at para 74.

[70] See *Amnesty Report*, *supra* note 57 at 27: "We used to live together peacefully alongside the other ethnic groups in Myanmar. Their intentions were good to the Rohingya, but the government was against us. The public used to follow their religious leaders, so when the religious leaders and government started spreading hate speech on Facebook, the minds of the people changed."

[71] See *ibid*.

[72] See *ibid* at 32.

[73] See *ibid*.

[74] See *ibid*.

Rwandan villages that received signal, compared with those that did not receive signal.[75] As Amnesty's report puts it, "narratives of dehumanization, of impending threat or takeover from their 'other' and false information regarding the wrongs they have supposedly perpetrated"[76] create an environment that makes mass violence possible.

In the case of Myanmar, Facebook was the primary tool for creating this environment. A study by Victoire Rio found that both the Tadmadaw and MaBaTha engaged in systemic operations that employed hundreds of staff between them in order to amplify their anti-Rohingya sentiments through the power of Facebook.[77] After the UN Fact-Finding Mission was published, Facebook responded by publishing an independent human rights impact assessment (HRIA) for Myanmar, which came to the same conclusions.[78] The HRIA acknowledged that Facebook played a role in the violence against the Rohingya, stating, "Facebook has become a useful platform for those seeking to incite violence and cause offline harm. Though the actual relationship between content posted on Facebook and offline harm is not fully understood, Facebook has become a means for those seeking to spread hate and cause harm, and posts have been linked to offline violence."[79]

Thus, it has been established that hate speech on Facebook played the role of increasing the likelihood of mass violence against the Rohingya between 2017 and 2018. What is less clear is how the company failed so badly at removing the type of content that has been linked to the violence.

ii) Content Moderation

There have been internal policies against the use of "hate speech" on Facebook for many years, with systems in place to remove or demote harmful content from the platform. The company's Community Standards recognize that hate speech on

---

[75] See *ibid* at 33.

[76] *Ibid*.

[77] See *ibid* at 30.

[78] See BSR, "Human Rights Impact Assessment: Facebook in Myanmar" (2018), online (pdf): <habout.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf>.

[79] *Ibid* at 24.

the platform can lead to offline violence,[80] yet Facebook failed to act during the outpouring of violent rhetoric leading up to the 2018 genocide in the Rakhine province of Myanmar, and continued to fail even after being called to account. For example, Facebook has never disclosed the number of Burmese-language content moderators it employed during the atrocities. An independent investigation later found that the company only had five Burmese language speakers to monitor content in a country of over 18 million Facebook users.[81]

Amnesty International's report highlights survivor testimony of Facebook's inadequate content moderation leading up to the Rakhine crisis. Several Rohingya refugees told Amnesty that they tried to 'report' anti-Rohingya content on Facebook before and during the violence, to no avail. Facebook either never responded or told them that the content did not violate the platform's Community Standards.[82]

Algorithms and code also played a role in the disastrous lack of moderation on Facebook; a role that was more political than it seemed. Due to Myanmar's late entry into the internet world, Burmese was not included in Unicode script.[83] As a result, the Burmese people invented their own encoding standard – Zawgyi. Without getting too technical, Zawgi is restricted to devices in Myanmar, and is incompatible with Facebook's language-detecting algorithms.[84] Famously, the phrase "kill all the Kalars that you see in Myanmar; none of them should be alive" became "I shouldn't have a rainbow in Myanmar" through Facebook's content-moderation software. [85] Moreover, the platform's choice to encode only Burmese as the sole dialect for Facebook exemplifies the inability to center the socio-cultural context of Myanmar in their decision making. There are several other languages commonly used in the country that have a distinct

---

[80] See *Amnesty Report*, *supra* note 57 at 34.

[81] See Ceclia Kang & Sheera Frenkel, *An Ugly Truth: Inside Facebook's Battle for Domination* (NY: Harper Collins, 2021) at 191.

[82] See *Amnesty Report*, *supra* note 57 at 35.

[83] Unicode is the most commonly used information technology standard for consistent encoding and representation of text in computer programming.

[84] See Sablosky, *supra* note 59 at 1032.

[85] *Ibid;* Kalar became a hateful ethnic slur for a Rohingya.

writing style, including Tai, with 3.2 million users.[86] The Tatmadaw has been imposing strict laws to prevent the use of languages other than Burmese in the country for generations, so Facebook's choice goes beyond convenience and demonstrates an unintentional allegiance to the Myanmar military and their politics. As Sablosky puts it, "despite their proclaimed detachment and support for global expression, Facebook's action in Myanmar attest to the outsized impact its decisions have on international matters."[87] In Myanmar, we witnessed the dystopian reality of a push for technology over considerations of risk, with grave results.

iii) Moving Forward

There is currently a high-profile class-action suit against Facebook (now Meta) for their inaction in Myanmar,[88] and a separate case brought before the International Court of Justice against Myanmar that has led to discovery of many internal documents within Facebook. [89] These cases, alongside investigations into internal documents conducted by Amnesty International and leaked records such as the Facebook Papers[90] have together generated enough evidence to demonstrate that Facebook "substantially contributed" to human rights harms suffered by the Rohingya.[91] Facebook has taken action following these findings, hiring more Burmese-speaking content moderators and establishing the Meta Oversight Board, which will be discussed in detail later. It remains to be seen if these changes are enough to protect vulnerable populations from ethnic violence incited online.

---

[86] See *ibid* at 1033.

[87] *Ibid* at 1036.

[88] See *Jane Doe v Meta Platforms Inc*. No 3 2022cv07557 (ND Cal). Rohingya Refugees are claiming over $150 billion for compensatory damages from Meta.

[89] *Application of the convention on the prevention and punishment of the crime of genocide (The Republic of the Gambia v Myanmar)* (2019); *The Republic of the Gambia v Facebook*, No 20-mc-36-JEB-ZMF (ND Cal 2021).

[90] See Facebook Papers, "Facebook and Responsibility" (last visited 30 August 2023), online (pdf): <documentcloud.org/documents/21594152-tier2_rank_other_0320>. Facebook executives were briefed on the legal implications of algorithmic recommendations. On p.1: "Actively ranking content in News Feed and promoting content on recommendations surfaces makes us responsible for any harm caused by exposure to that content."

[91] See *Amnesty Report*, *supra* note 57 at 62.

In Myanmar, we saw limited telecommunications that were followed by a Facebook monopoly over communications without protective moderation mechanisms in place. In an unstable environment fraught with ethnic tensions, this can lead to hate speech online circulating rapidly and causing real-world violence. In the next section, I will investigate the recent Tigray crisis, highlighting the role of online disinformation in aggravating and prolonging conflict.

B. Ethiopia

Tensions between the Tigray region and Ethiopia's federal government have been high throughout the country's history, and have mounted recently when Prime Minister Abiy Ahmed dissolved the former dominant political party to form his Prosperity Party in 2019, which wishes to move away from Ethiopia's system of ethnic federalism. Tigrayans and other minority ethnic groups oppose this move as they worry that a dissolution of ethnic federalism would result in their oppression. The armed conflict broke in November 2020 when the Tigray People's Liberation Front (TPLF) attacked the Ethiopian National Defence Forces in Tigray. The federal government responded by shutting down the internet and telecommunication services in the region and sending troops loyal to the federal government.

Since November 2020, Tigrayans have endured extreme ethnic-based violence,[92] they have been pushed out of Western Tigray and banned from speaking their language of Tigrinya. Attacks on civilians by all parties have been found to be in violation of international humanitarian law which may amount to war crimes[93] and it is estimated that the death toll at the end of summer 2022 was in the hundreds of thousands, with the violence

[92] Amounting to war crimes, see Human Rights Watch, "We Will Erase You From This Land: Crimes Against Humanity and Ethnic Cleansing in Ethiopia's Western Tigray Zone" (2022), online (pdf): *Human Rights Watch* <hrw.org/sites/default/files/media_2022/04/ethiopia0422_web_1.pdf>.

[93] See *Report of the OHCHR-EHRC joint investigations into alleged violations of international human rights, humanitarian and refugee law committed by all parties to the conflict in the Tigray region of the Federal Democratic Republic of Ethiopia*, OHCHR (3 November 2021) at 83, online: (pdf) <ohchr.org/sites/default/files/2021-11/OHCHR-EHRC-Tigray-Report.pdf>.

exacerbating the already dangerous risk of famine.[94] In addition, from November 2020 to August 2021, communications were not fully restored in the region, making these estimates challenging to verify.[95] Without a good understanding of the conflict as it evolves, the international community remains helpless to intervene. What information does circulate should be verified, and OSPs have a responsibility to label disinformation and remove hate speech that worsens conflicts.

As I will describe below, this conflict has highlighted the destructive impacts that disinformation can have on conflict situations.

i) The Ethiopian Media Landscape: Aiding the Spread of Misinformation

The media landscape in Ethiopia, as well as a failure of OSPs to invest in language resources, has influenced the spread of disinformation and intensified the severity of the conflict. Media in Ethiopia has historically been under heavy State control, with a continuous prominence of State-centralised media outlets in the country.[96] This heavy media control explains how the Ethiopian government was so effective at establishing a media blackout at the start of the conflict, shutting down internet and communications infrastructure in the Tigray region, and expelling foreign

---

[94] It is very difficult to know the exact number of deaths due to absence of communications in the region, and blackmailing of NGOs working there. Staff and volunteers at Ghent University have researched and recorded details of civilian casualties, with a low estimate around 300,000 and a high estimate of 600,000. The death toll contains casualties due to direct killings, deaths due to lack of healthcare, and deaths due to famine (the latter having the greatest impact). See Martin Plaut, "New estimate of the Tigray death toll" (19 October 2022), online: *Martin Plaut* <martinplaut.com/2022/10/19/new-estimate-of-the-tigray-death-toll/>.

[95] See Claire Wilmot, Ellen Tveteraas & Alexi Drew, "Dueling Information Campaigns: The war over the narrative in Tigray" (20 August 2021), online: *Media Manipulation Casebook* <mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray#footnoteref4_98p4l98>.

[96] See Muna Shifa & Fabio Andres Diaz Pabon, "The Interaction of Mass Media and Social Media in Fueling Ethnic Violence in Ethiopia" (15 March 2022), online: *Accord* <accord.org.za/conflict-trends/the-interaction-of-mass-media-and-social-media-in-fuelling-ethnic-violence-inethiopia/#:~:text=In%20Ethiopia%2C%20social%20media%20is,into%20mass%20atrocities%20and%20genocide.>.

journalists from the area.[97] In turn, this media blackout resulted in social media becoming a central source of information and of disinformation outside of the Tigray region, where telecommunications were also shut down by government orders. Although internet use in Ethiopia is low,[98] online coverage of the conflict is popular amongst members of the diaspora who want updated information on their home country. As a result, a battle over the narrative around the conflict is brought to an international stage.

There have been numerous online campaigns purportedly dedicated to documenting the Tigray conflict on social media, particularly on Twitter, [99] but many of the posts contained disinformation meant to vilify the other side. Thousands of Ethiopians from the diaspora of both sides of the conflict also joined Twitter after the fighting began, sharing contradictory information that caused confusion and anger.[100]

For example, an echo chamber of disinformation was created at the start of the Tigray conflict, when Tigrayan supporters claimed that Eritrean troops systematically killed hundreds of civilians in the city of Axum on the 28th of November 2020.[101] Government supporters called these claims "fake news," and their support networks online described these as attempts to undermine the credibility of the government. Tigrayan supporters, on the other hand, claim that this was a mode of discrediting

---

[97] See Salem Solomon, "Journalists struggle through information blackout in Ethiopia", *VOA* (2 December 2020), online: <voanews.com/a/press-freedom_journalists-struggle-through-information-blackout-ethiopia/6199045.html>.

[98] This is lower than the 60% global average. See Lars Kamer, "Internet usage in Africa" (17 November 2022), online: *Statistia* <statista.com/topics/9813/internet-usage-in-africa/#:~:text=Africa's%20online%20shoppers%20amounted%20to,with%20the%20rising%20internet%20penetration.>.

[99] See *ibid*.

[100] See Wilmot, *supra* note 95 at 4.

[101] Prime Minister Abiy Ahmed announced on November 4, 2020 that the Ethiopian National Defense Forces (ENDF) had been ordered to fight the TPLF and militia loyal to them. The ENDF has relied on the support of special forces from the Amhara region, and on the Eritrean Defence Force. Ethiopian and Eritrean authorities have made conflicting statements about the involvement of Eritrean troops in Tigray.

survivors and reporters covering Tigrayan suffering.[102] Amnesty International conducted an independent report on the occurrences in Axum, corroborating Tigrayan supporter stories of the massacre there, which offered clarity to the situation. The report describes the violence as an atrocity which "ranks among the worst documented so far in this conflict. Besides the soaring death toll, Axum's residents were plunged into days of collective trauma amid violence, mourning and mass burials."[103] At least 240 citizens were indiscriminately killed by Eritrean forces in two days, according to the report. Supporters of the government actively tried to discredit the report, launching a fake "fact-checking" campaign on Facebook.[104] The government later publicly acknowledged the Axum massacre.[105]

A lack of understanding of the location, extent, and reasons for violence leaves innocent citizens vulnerable to harm and might even influence some to take up arms without real cause.[106] The Tigray conflict has demonstrated that when disinformation is circulated in environments with poor access to information, it increases fear, instability and prolongs violence.

It is clear that war crimes and humanitarian law are the most pressing concerns now that peace talks have begun between the TPLF and the federal government. For the scope of this paper, however, I want to understand what role social media disinformation and hate campaigns had on the escalation of the conflict.

ii) Twitter and Facebook's Response

As exemplified by the tragedies in Myanmar, capacity for content moderation in non-Anglophone countries has been a challenge for tech companies, and African countries are no

---

[102] See Wilmot, *supra* note 95.

[103] Amnesty International, "The Massacre in Axum" (26 February 2021), online: *Amnesty International* <amnesty.org/en/documents/afr25/3730/2021/en/>.

[104] See Wilmot, *supra* note 95.

[105] See Cara Anna, "Ethiopia now calls Axum massacre allegations 'credible'", *AP News* (3 March 2021), online: <apnews.com/article/abiy-ahmed-ethiopia-massacres-belgium-kenya4e5eda7bb2753973951269039d5ab802>.

[106] See Jack Burnham, "From the Internet to Ashes: Disinformation and the Tigray War", *Nato Association* (8 September 2022), online: <natoassociation.ca/from-the-internet-to-ashes-disinformation-and-the-tigray-war/>.

exception. In April 2021, Twitter announced that it would open its first Africa office in Ghana.[107] However, most of the job openings for that office were in engineering, advertising, and communications instead of content moderation. Without investing in local content moderators, it will be difficult for OSPs to fight ethnocentric disinformation. This is particularly important because Africa is home to 2,133 languages and almost 1.4 billion people. Although Twitter has millions of users on the continent, its current language support does not include any major spoken African language other than Arabic.[108] This oversight demonstrates an indifference from the company regarding online safety in Africa, a place where state-sanctioned disinformation has had impacts ranging from unfair elections[109] to intensification of cultural tensions resulting in violence. Underscoring the importance of OSP action in the face of online harms, a 2015 study examined of the relationship between communication technologies and political violence in twenty-four African countries, and found that the expansion of social media was associated with an increase in the incidence of collective violence.[110]

In Ethiopia specifically, there are between 45 and 86 languages spoken, with 29% of the population speaking Amharic, 33% speaking Oromo, and 6% speaking Tigrinya.[111] It is unclear the extent to which content in any of these languages is moderated effectively by Twitter. In the case of the recent Tigray conflict,

---

[107] See Kevin Keykopour & Uche Adegbite, "Establishing Twitter's presence in Africa" (12 April 2021), online: *Twitter Blog* <blog.twitter.com/en_us/topics/company/2021/establishing-twitter-s-presence-in-africa>.

[108] See Torinmo Salau, "How Twitter Failed Africa: Big Tech ignored policies that enable disinformation and propaganda across the continent", *Foreign Policy* (19 January 2022), online:

<foreignpolicy.com/2022/01/19/twitter-africa-ghana-dorsey-disinformation/>.

[109] See Kinife Micheal Yilma, "On Disinformation, Elections and Ethiopian Law" (2021) 65:3 J Afr Law 351.

[110] See T Camber Warren, "Explosive Connections? Mass Media, Social Media, and the Geography of Collective Violence in African States" (2015) 52:3 J Peace Research 297.

[111] See Translators Without Borders, "Language data for Ethiopia" (last visited 30 August 2023), online: *Translators Without Borders* <translatorswithoutborders.org/language-data-for-ethiopia>.

Twitter's main action was to halt Trending Topics in Ethiopia.[112] However, a study out of NYU found that this did not significantly decrease the volume of English-language tweets related to the conflict (they were unable to conduct the same analysis for tweets in Amharic or Tigrinya).[113] Twitter's efforts to curb disinformation were thus insufficient.

Facebook is another OSP that has come under scrutiny due to their mishandling of harmful online content during the conflict. The more prevalent use and misuse of social media in non-English languages has resulted in more limited moderation of hate speech and disinformation. For example, in 2020, a disinformation campaign was used to vilify Hachalu Hundessa, a prominent Ethiopian musician and activist; he was later assassinated. After his death, there was rampant hate speech and incitement to violence on Facebook, resulting in mob violence that led to hundreds of deaths. This ultimately drove the government to conduct a full internet shutdown.[114] At that time, Facebook had not yet made its hate-speech detection software compatible with Amharic or Oromo, and some blame this oversight in part for the violence in the wake of Hundessa's killing.[115]

After receiving global criticism for their role in Hudnessa's murder, Facebook has taken some action; they have expanded its capacity to review content in Amharic, Oromo, Somali and Tigrinya, developed technology to automatically identify hate speech and ethnic slurs, removed coordinated unauthentic

---

[112] See Twitter Safety, "Given the imminent threat of physical harm, we've temporarily disabled Trends in Ethiopia. Alongside continued efforts to disrupt platform manipulation, we hope this measure will reduce the risks of coordination that could incite violence or cause harm." (5 November 2021 at 22:40), online: *Twitter* <twitter.com/TwitterSafety/status/1456813765387816965?s=20>.

[113] See Megan A Brown, "Trendless Fluctuation? How Twitter's Ethiopia Interventions May (Not) Have Worked", *Tech Policy Press* (11 January 2022), online: <techpolicy.press/trendless-fluctuation-how-twitters-ethiopia-interventions-may-not-have-worked/>.

[114] See Prabha Kannan, "Digital Extractivism in Africa Mirrors Colonial Practices" (15 August 22), online: *Stanford University, Human-Centred Artificial Intelligence* <hai.stanford.edu/news/neema-iyer-digital-extractivism-africa-mirrors-colonial-practices>.

[115] See *ibid*.

behavior,[116] released political transparency tools,[117] and run a media literacy billboard campaign across Addis Ababa.[118]

What are the impacts of these actions? The Meta Oversight Board has ruled on two separate occasions that Facebook should remove content that could incite further violence in Ethiopia.[119] In one of the decisions, the Board reiterated that:

> In situations of armed conflict in particular, the risk of hateful, dehumanizing expressions accumulating and spreading on a platform, leading to offline action impacting the right to security of person and potentially life, is especially pronounced. Cumulative impact can amount to causation through a gradual build-up of effect, as happened in the Rwandan genocide.[120]

As I will discuss below, the Oversight Board's policy recommendations are not binding on Facebook, and the OSP still continues to fail in removing harmful content from its platform, especially in non-English languages. To highlight this, Global Witness conducted a study in June 2022 to test Facebook's hate speech detection system specifically for Ethiopian content. The group identified the twelve worst examples of Amharic-language hate speech that had been posted on Facebook and submitted

---

[116] See Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior From Ethiopia" (16 June 2021), online: *Meta* <about.fb.com/news/2021/06/removing-coordinated-inauthentic-behavior-from-ethiopia/>.

[117] See Meta Newsroom, "Ethiopia: Preparing for Elections Day" (5 May 2021), online: *Meta* <facebook.com/gpa/blog/ethiopia-preparing-for-election-day?_rdc=2&_rdr>.

[118] See Meta Newsroom, "An Update on Our Longstanding Work to Protect People in Ethiopia" (9 November 2021), online: *Meta* <about.fb.com/news/2021/11/update-on-ethiopia/>.

[119] See Oversight Board, "Oversight Board upholds Meta's original decision: Case 2021-014-FB-UA" (December 2021), online: *Oversight Board* <oversightboard.com/news/927673894608838-oversight-board-upholds-meta-s-original-decision-case-2021-014-fb-ua/> [Oversight Board, "Case 2021"]; Oversight Board, "Oversight Board upholds Meta's decision in "Tigray Communication Affairs Bureau" case 2022-006-FB-MR" (October 2022), online: *Oversight Board* <oversightboard.com/news/592325135885870-oversight-board-upholds-meta-s-decision-in-tigray-communication-affairs-bureau-case-2022-006-fb-mr/> [Oversight Board, "Case 2022"].

[120] Oversight Board, "Case 2021", *ibid*.

them for approval as advertisements. All twelve were approved.[121] When Global Witness informed Facebook of this serious failure in content moderation, a spokesperson acknowledged that the advertisements should not have been approved, as they violate Facebook's policies on hate speech and incitement to violence. Two weeks later however, Global Witness submitted an additional two ads for approval, and they were also approved. The study concludes that Facebook's actions have not made a significant impact on reducing harmful speech in Ethiopia. It calls on the company to properly resource content moderation in all of the countries in which they operate. Indeed, additional investigations have found that Facebook is not doing enough to remove posts that call for violence and spread dangerous disinformation.[122]

Overall, the lack of independent media in Ethiopia, and state-sanctioned regional telecommunications shutdowns created poor access to information in conflict-affected areas. In turn this allows more unverified and potentially false information to be circulated, both internally and externally to the country. This increases confusion and disruption for the parties suffering in the conflict. OSPs, especially Twitter and Facebook in this case, have a responsibility to act against disinformation on their platforms.

The OB has made several recommendations to Facebook, including for them to clarify that in situations of violent conflict and war, unverified rumors pose higher risk to the rights of life and security of persons. As such, this should be reflected at all levels of the content moderation process.[123]

---

[121] See Global Witness, "'Now is the time to kill': Facebook Continues to Approve Hate Speech Inciting Violence and Genocide During Civil War in Ethiopia" (June 2022), online (pdf): *Global Witness* <globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/>.

[122] See Jasper Jackson et al, "Facebook accused by survivors of letting activists incite ethnic massacres with hate and misinformation in Ethiopia", *The Bureau of Investigative Journalism* (20 February 2022), online: <thebureauinvestigates.com/stories/2022-02-20/facebook-accused-of-letting-activists-incite-ethnic-massacres-with-hate-and-misinformation-by-survivors-in-ethiopia>.

[123] See Oversight Board, "Case 2021" *supra* note 119; Oversight Board, "Case 2022", *supra* note 119.

C. *Case Studies: Conclusions*

As these case studies demonstrate, OSPs can be used in conflict zones to spread disinformation, fueling uncertainty and upholding instability, as well as to spread hate speech to dehumanize the opposition. It is challenging to moderate harmful online content, in part because states sometimes do not abide by IHRL. For example, we saw that in Myanmar Facebook was used by members of the military, the government, and extremist groups to fuel ethnic tensions that ultimately lead to extreme acts of violence against the Rohingya. In Ethiopia, internet shutdowns mandated by government were used to control the narrative over the conflict to people outside of the Tigray region, which influenced political discourse and perpetuated instability. Both of these situations are likely to be found contrary to article 20 of the ICCPR.[124]

Although distinct situations, the common denominator in these cases is that harmful online content can prolong and worsen historical tensions, and incite people to do real-world harm. OSPs have a responsibility to prevent these harms by improving their content moderation capacities and ensuring that their Community Guidelines are consistent with IHRL. To accomplish their due diligence for upholding human rights, local understandings of socio-cultural, political, and economic contexts and challenges, as well as increased language proficiencies are required from OSPs.

The challenge of implementing these changes lies partly in the willingness of OSPs to take action, and this willingness may only result by increasing their liability. In the next section, I will discuss current frameworks that are in place to hold OSPs accountable, and how they might be improved.

III. Towards a Policy Agenda Confronting OSP Responsibility and Platform Governance

[124] *ICCPR*, *supra* note 9, arts 20(1), 20(2). Article 20(1) of the ICCPR obligates states to prohibit propaganda for war, and article 20(2) prohibits "advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence."

As the case studies demonstrate, OSP policies and actions (or inaction) impact human rights in meaningful ways; this is particularly true in states with widespread ethnic violence such as Myanmar and Ethiopia. OSPs must take a more active role in regulating online content, and they will only do so if they are regulated more carefully, and with an IHRL-compliant balancing of free expression and prohibition of harmful content. In the following section, I discuss self-regulation as a potential solution towards reducing online harms in conflict zones, using Meta's Oversight Board as a model. I will also highlight how regional regulations may push OSPs to implement self-regulatory governance structures that are in line with IHRL norms.

A. *International Human Rights Law as a Framework for OSPs*

Although the weaknesses of IHRL in this area are important (see Part I), it can nevertheless provide a helpful framework generally for holding governments, OSPs, and individuals accountable in a transnational context. IHRL resulted from decades of debate and collaboration between states and experts around the world and offers the only set of tested international rules and principles for legal rights. Moreover, in conflict-ridden states, the rule of law is ambiguous. IHRL is the only mode to hold such states accountable for human rights violations as its global norms are as close to universal state consent as anything. With these points in mind, I present two suggestions to reduce online harms in conflict zones: self-regulation and regional legislation imposing third-party liability on OSPs.

i) Self-regulatory Mechanisms

Experts in this area have been highlighting the need for OSP regulatory mechanisms that are agile and responsive enough to be effective in dynamic situations that require rapid and broad action, [125] conflict zones being one of them. High-level international and even domestic legislative efforts are not up to this task on their own due to their slow processes. Instead, scholars

[125] See e.g. David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 70, UN Doc A/HRC/38/35 (6 April 2018); Rory Van Loo, "Federal Rules of Platform Procedure" (2021) 88: 4 U Chi L Rev 829.

suggest that the focus should be on legitimizing the process by which platforms make decisions about speech.[126] One mechanism for this is self-regulation, whereby OSPs draft platform rules that are compliant with IHRL and maintain independent oversight bodies to audit the platforms, holding them accountable for non-compliance. Some have called this model a form of "digital constitutionalism."[127] In the following section, I will outline the main advantages and challenges of this approach, particularly for states that experience conflict.

ii) The Oversight Board

Meta's Oversight Board (OB) is a good example of how this might be done (interestingly, the company launched the OB partly in response to the violence in Myanmar).[128] Briefly, the OB is an independent institution established to review Facebook and Instagram's content moderation decisions. Legally independent from Meta, the OB's mandate is twofold; to issue binding decisions on content moderation questions, and to offer non-binding recommendations regarding platform policies.[129] The OB is governed by its Charter and Bylaws, outlining the process through which the Board selects content moderation decisions to take on. The Charter enshrines the independence of the OB from Meta and highlights the importance of precedent and IHRL norms in its decision-making.[130] In so doing, Meta's Oversight Board seems to be taking into account academics' calls for platforms to implement basic principles of administrative law – transparency, participation, reason-giving, and review.[131] However, the Bylaws

---

[126] See Evelyn Douek, "Verified Accountability: Self-regulation of content moderation as an answer to the special problems of speech regulation" (2019) Hoover Institution Series 1903 at 7 [Douek, "Accountability"].

[127] *Ibid* at 1.

[128] Public controversies including Facebook's implication in the Myanmar crisis precipitated creation of Oversight Board. See Kate Klonick, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129:2418 Yale LJ 2418 at 2447–48.

[129] See David Wong & Luciano Floridi, "Meta's Oversight Board: A Review and Critical Assessment" (2023) 33:261 Minds & Machines 261.

[130] See Meta, "Oversight Board Charter" (last visited 31 August 2023), art 2(2), online (pdf): <about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf>.

[131] See Van Loo, *supra* note 125 at 843.

provide legal constraints on the Board, preventing it from reviewing cases where "the underling content is unlawful in a jurisdiction with a connection to the content."[132] This effectively limits its ability to review decisions in jurisdictions where laws exist that amount to citizen censorship, which is an obstacle that must be addressed by Meta.

With these elements in mind, the most important weaknesses of the OB are its limited jurisdiction and capacity. Indeed, the OB can only rule on posts, not accounts, and with only forty Board members (mostly from the high-income countries),[133] there are a limited number of decisions that can be rendered. It is also notable that two-thirds of appeals came from the high-income countries in 2021, with significant geographic regions such as Sub-Saharan Africa and Central and South Asia representing only 2% of appeals.[134] It is unclear if these numbers are reflective of a lack of awareness of or access to the OB processes, or insufficient attention from the Board to users in low-income countries. Furthermore, the OB's lack of policy influence within Meta is a significant weakness that will slow change; this system should be adjusted to grant the OB more policy influence over Meta.

Despite these weaknesses, there are important takeaways to be gleaned from Meta's Oversight Board. The most significant strengths of the OB are its ability to enhance the transparency of content moderation decisions and processes, the ability to effect OSP reform indirectly through policy recommendations, and its assertiveness in overruling Meta on moderation decisions.[135] Due to its independent judiciary-like structure, the OB is an approach to content moderation that transcends borders, offers flexibility, accountability, and transparency. Whenever a controversial

---

[132] Meta, "Oversight Bylaws" (last visited 31 August 2023), s 1.2.2, online (pdf): <about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf> which reads in full, "where the underlying content is criminally unlawful in a jurisdiction with a connection to the content (such as the jurisdiction of the posting party and/or reporting party) and where a board decision to allow the content on the platform would lead to adverse governmental action against Facebook."

[133] Most of the OB's members are from the USA or Europe, which is problematic given that Southeast Asia contained four of the top 10 countries with the largest Facebook audiences in 2019. See Wong and Floridi, *supra* note 129 at 270.

[134] See Wong & Floridi, *supra* note 129 at 271.

[135] See *ibid* at 266.

content moderation decision is taken by an OSP, there tends to be public outcry. Systems like the OB allow for open contestation and explanation of the norms that OSPs are developing. As a result, OSP users may become aware of the rules and then aid in generating compliance.[136] Additionally, an independent oversight mechanism provides a procedure for OSPs to outsource controversial decisions and avoid commercial, political, or majoritarian pressure, [137] and maintain legitimacy. Importantly, the OB's decisions are based on precedent and are aligned with IHRL.[138]

In the case of regions experiencing conflict, OSP's self-regulatory mechanisms may work towards quickly imposing IHRL-based norms on what constitutes harmful online content to uphold the universal right to freedom of expression. Certainly, a high degree of local consultation will be required to prevent the imposition of 'global' norms that do not consider local contexts. In addition, content must be moderated under adapted standards in conflict situations. During conflict, there should be mechanisms to include higher sensitivity to hate speech and disinformation, and blocking of accounts that spread this type of harmful content, even if they are state-affiliated. Until IHRL and regional legislative frameworks can regulate on these issues with clarity and flexibility, self-regulation by OSPs is necessary to prevent harmful online content from causing real-world harms. As I discuss further below, OSPs are unlikely to implement IHRL-compliant regulation on their own. Instead, states must also push OSPs by legislating on platform responsibility. Currently, the EU is leading by example in this area.

iii) Regional Legislation

There have been major legislative changes around the world in recent days with the common goal of imposing stricter

---

[136] See Douek, "Accountability", *supra* note 126 at 17.

[137] See *ibid* at 18.

[138] The OB has cited IHRL in overturning Meta's content moderation decisions and policies; see Oversight Board, "Case 2021", *supra* note 119, where they called Facebook's decision to remove a post as "an unnecessary and disproportionate restriction on free expression under international human rights standards."

OSP responsibility.[139] The most sweeping is the European Union's Digital Services Act (DSA).[140]

The DSA rules entered into force on November 16th, 2022, intending to address three of the main problems related to the governance of digital services in the EU; increasing exposure to illegal and harmful activities online, lack of cooperation between national authorities, and risks of legal fragmentation and legal barriers for digital services.[141] The DSA addresses these issues through binding EU-wide obligations that apply to all digital services that connect consumers to goods, services, or content, and is thus the broadest and most sweeping legislative effort in the area of online safety to date. Throughout the legislation there is a focus on the need to protect users' fundamental rights online, and it thus serves as a tool to promote the application of the EU Charter online.[142]

To summarize this enormous piece of legislation as it relates to OSPs, the DSA imposes penalties on OSPs that do not take down defined types of harmful content[143] within a prescribed period.[144] It also obligates different rules for different actors, depending on their role, size, and impact. For example, Very Large Platforms[145] must conduct additional annual risk assessments[146] and an independent audit[147] to understand illegal content dissemination through their services, any negative impacts

[139] See Appendix for a sampling of legislative efforts in democracies on digital regulation.

[140] *DSA*, *supra* note 48.

[141] See *ibid*, Preamble.

[142] See Giancarlo Frosio, "Platform Responsibility in the Digital Services Act: Constitutionalising, Regulating and Governing Private Ordering", forthcoming in Andrej Savin & Jan Trzaskowsi, eds, *Research Handbook on EU Internet Law* (Edward Elgar, 2021) at 12.

[143] Defined as content that is illegal according to EU law. See *DSA*, *supra* note 48, art 3(h).

[144] See *DSA*, *supra* note 48, art 6.

[145] Defined as those that provide services to over 10% of the EU population (45 million at the moment), thus presumed to have the largest impact and highest risk.

[146] See *DSA*, *supra* note 48, art 34.

[147] See *ibid*, art 37.

on fundamental rights,[148] and to ensure compliance with its obligations under the law. Very Large Platforms also now have obligations to provide transparent information about their recommender systems and content removal processes,[149] and to share data with researchers and authorities.[150] Procedural guarantees for content removal notice, counter notice, and complaint procedures must additionally be ensured by these platforms.[151]

The reception towards this legislation has been mixed.[152] Proponents believe that it is positive overall, as it imposes a process of constitutionalism for online regulation that is based heavily in the rights enshrined by the EU Charter,[153] and includes multiple oversight mechanisms. Critics highlight the many liability exemptions for OSPs and the lack of specific obligations to address algorithmic opacity.[154] I will detail these in turn.

iv) Liability Exemptions

A long list of liability exemptions that were enshrined in the DSAs predecessor is still present,[155] which exempt platforms from liability for illegal content they host if they did not have actual knowledge of it. The broad liability exemptions allow leeway for OSPs to not moderate sufficiently against harmful content. Instead, it should be the responsibility of OSPs to develop systems to monitor defined types of harmful content, particularly in conflict

---

[148] Specifically, freedom of expression and information, the right to private life, the right to non-discrimination, and the rights of the child.

[149] See *DSA*, *supra* note 48, arts 27, 16–18.

[150] See *ibid*, art 40.

[151] See *ibid*, art 16.

[152] See e.g. David Morar, "The Digital Services Act's lesson for the U.S. policymakers: Co-regulatory mechanisms" (23 August 2022), online: *Brookings* <brookings.edu/blog/techtank/2022/08/23/the-digital-services-acts-lesson-for-u-s-policymakers-co-regulatory-mechanisms/>; *contra* Julia Keseru, "The EU's Digital Services Act Doesn't Go Far Enough" (16 May 2022), online: *CIGI* <cigionline.org/articles/data-rights-protections-are-overdue-for-an-upgrade/>.

[153] See Frosio, *supra* note 142.

[154] See *ibid*. There are also concerns that the *DSA* does not sufficiently protect user privacy, but this goes beyond the scope of this paper.

[155] See European Commission, *E-Commerce Directive,* [2000], OJ, 2000/31/EC, arts 12–15 are now *DSA*, *supra* note 48, arts 4–6.

situations. As demonstrated in the cases of Myanmar and Ethiopia, a lack of contextual understanding and failure to invest in moderators with relevant language proficiency results in harmful content spreading online and causing real-world harm. An effective option would be to attribute liability to OSPs when they are involved in the creation, optimization, or promotion of the impugned content, as this would capture dangerous algorithmic processes that promote viral but harmful content.[156] IHRL norms should also be implemented and to determine what kinds of content are considered harmful, the Rabat Plan of Action should be used.

v) Algorithmic Opacity

Although not discussed in detail in this paper, transparency of algorithmic tools is incredibly important for online safety. Many algorithmic systems prioritize virality over safety, [157] thus amplifying content that can incite violence over content that contains verified information, [158] increasing the rate at which harmful content circulates online. There needs to be more regulation of algorithms, increased transparency over their functions, and prohibitions on the algorithmic promotion of harmful content.

vi) Self-regulation is Key

As discussed, the global recognition of the unprecedented harms that can occur under the unchecked power of OSPs has been increasing in recent years. Government initiatives such as the DSA are not without flaws but are ultimately positive because they at least push OSPs to get serious about self-regulation. I posit that self-regulation remains the most effective mechanism to protect users against harmful content in conflict zones, as they can adapt to changing online environments without bureaucratic burdens. The biggest barrier to self-regulation is its higher costs, but as shown through various corporate histories, credible threats of government regulation result in changed corporate

---

[156] See Miguel Peguera, "The platform neutrality conundrum and the Digital Services Act" (2022) 53 IIC 681 at 684.

[157] See e.g., Matthew Shaer, "What Emotion Goes Viral the Fastest" (April 2014), online: *Smithsonian Magazine* <smithsonianmag.com/science-nature/what-emotion-goes-viral-fastest-180950182/>.

[158] See *Amnesty Report*, *supra* note 57 at 9.

behavior.[159] Moreover, though it will be logistically challenging for OSPs to apply different rules for every jurisdiction,[160] tragedies such as the Rohingya genocide in Myanmar underscore the vital importance of timely intervention.

Additionally, self-regulation is best for conflict-heavy states that do not necessarily abide by international or regional human rights norms. More generally, there is a danger of over-censorship if governments were granted the primary power to manage rules for online speech. Instead, they should create broad guidelines for OSP content moderation, all within the restraints of their constitutions, and hopefully, IHRL. [161] OSP self-regulation mechanisms should abide by these guidelines, while accepting that they cannot be applied universally, to sufficiently protect vulnerable groups against human rights violations online.

As Douek writes, IHRL may be the "least-worst" option for OSP regulation. [162] However, the current framework lacks enforceability for OSPs, and indeterminacy within IHRL norms make breaches difficult to identify. Additionally, much of the scholarship on potential regulatory frameworks is considered only for democratic states.[163] These frameworks lack applicability for states with different governance structures. Indeed, there is much more work to be done to understand how OSPs might reduce their harm in conflict zones.

Conclusion

This paper has aimed to examine the current framework of IHRL in relation to freedom of expression online. It argued that

---

[159] For example, tobacco advertising shut down after government regulation became a real possibility. Also, when President Biden and former President Trump called for the elimination of the liability exemptions in CDA s 230, company CEOs were suddenly more eager to participate in regulatory changes.

[160] See Douek, "Accountability", *supra* note 126 at 11.

[161] See *ibid* at 6.

[162] See Douek, "Limits", *supra* note 47 at 72.

[163] Evelyn Douek's pieces discuss enhancing state-platform cooperation in order to for platform regulators to have democratic legitimacy, but in many states (Myanmar, Ethiopia) this would not be possible.

the IHRL framework alone is insufficient to protect users from online harm, particularly in conflict zones. The case studies of Myanmar and Ethiopia demonstrate that the lack of attention from OSPs to specific language needs, and the inadequate moderation of hate speech and disinformation, may result in real-world harms. Through an investigation into current self-regulatory and regional regulatory frameworks, I have highlighted a need for government intervention to push for a better self-regulatory framework, and the need for these interventions to be informed by IHRL norms.

With all this in mind I have several recommendations:

1. Self-regulation is required to respond effectively to changing dynamics in conflict situations, and OSPs must be pushed to implement them.
2. OSPs need specific policies for operating in conflict settings, based on IHRL,[164] and be required to conduct HR due diligence reporting (as in the DSA).
3. Content moderation policies must be aligned with IHRL standards on freedom of expression.
4. OSPs should abide by transparency reporting obligations on algorithms and content moderation decisions.
5. IHRL should develop clear definitions for hate speech and disinformation online.
6. Self-regulatory frameworks should include independent oversight bodies that have policy-making power and follow precedent.

International frameworks for balancing freedom of expression and other rights can also be usefully applied in non-conflict zones. For instance, the Oversight Board at Facebook invoked the Rabat Plan of Action in its recent analysis of the suspension of Donald Trump's account after the January 6 riots.[165] However, it is essential to recognize that harmful online content has even more serious effects on conflict-affected areas and should receive significant attention and investment.

---

[164] See Khan, *supra* note 5 at para 123.

[165] See Facebook Oversight Board, "Case decision 2021-001-FB-FBR: Case Summary" (2021), online (pdf): <oversightboard.com/sr/decision/2021/001/pdf-english>.

Bibliography

LEGISLATION

African Charter on Human and Peoples' Rights, adopted 27 June 1981, OAU Doc CAB/LEG/67/3 rev 5, 21 ILM 58 (1982).

American Convention on Human Rights, signed 22 November 1969, OASTS No 36, 1144 UNTS 123.

Citizens Protection (Against Online Harm) Rules, 2020.

Communications Decency Act, 47 USC 230.

Council of Europe, *Convention for the Protection of Human Rights and Fundamental Freedoms*, Rome, 4 XI 1950.

European Commission, *Digital Services* Act, [2022] OJ 2022/2065 of 19 October 2022 L277/1.

European Commission, *E-Commerce Directive,* [2000], OJ, 2000/31/EC.

International Covenant on Civil and Political Rights, 19 December 1966, 999 UNTS 171.

Law on Information, Information Technologies and Information Protection, Federal Law No 483-FZ0-2020, 2020.

The Law of the People's Republic of China on Safeguarding National Security in the Hong Kong Special Administrative Region, 2020.

The Network Enforcement Act, BD, 12 June 2017.

INTERNATIONAL AND GOVERNMENT DOCUMENTS

Council of Europe, *Information Disorder: Toward an interdisciplinary framework for research and policy making,* DGI (27 September 2017).

de Varennes, Fernand, *Recommendations made by the forum on minority issues at its thirteenth session on the theme,"Hate speech, social media and minorities"*, OHCHR, 46th Sess, A/HRC/46/58 (26 January 2021).

European Commission, *A multi-dimensional approach to disinformation* (Luxembourg: Publications Office of the European Union, 2018).

Freedom of Expression vs. incitement to hatred: OHCHR and the Rabat Plan of Action, OHCHR, 22nd Sess, UN Doc. A/HRC/22/17/Add.4, Appendix, adopted 5 October 2012.

General Recommendation No. 32 on The Meaning and Scope of Special Measures in the International Convention on the Elimination of Racial Discrimination, CERD/C/GC/32 (2009).

Government of Canada, Canadian Heritage, *Discussion Guide (Consultation closed: The Government's proposed approach to address harmful content online)* (2021), online: <canada.ca/en/canadian-heritage/campaigns/harmful-online-content/discussion-guide.html#a1>.

Guiding Principles for Business and Human Rights: Implementing the United Nations "Protect, Respect, and Remedy" Framework, OHCHR (2011).

Human Rights Committee, *General comment no 34*, 102nd Sess (2011).

Kaye, David, *Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, OL RUS 4/2019 (1 May 2019).

Kaye, David, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 70, UN Doc A/HRC/38/35 (6 April 2018).

Kaye, David & Michel Forst, *Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the situation of human rights defenders*, OL PAK 3/2020 (19 March 2020).

Khan, Irene, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 77th Sess, UN Doc A/77/288 (2022).

Online Safety Bill (UK), 209 2022-23.

Report of the independent international fact-finding mission on Myanmar, OHCHR, 39th Sess, UN Doc A/HRC/39/64 (2018).

Report of the Office of the United Nations High Commissioner for Human rights: The practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, OHCHR, 55th Sess, UN Doc A/HRC/50/56 (2022).

Report of the OHCHR-EHRC joint investigations into alleged violations of international human rights, humanitarian and refugee law committed by all parties to the conflict in the Tigray region of the Federal Democratic Republic of Ethiopia, OHCHR (3 November 2021), online: (pdf) <ohchr.org/sites/default/files/2021-11/OHCHR-EHRC-Tigray-Report.pdf>.

Report of the Special Rapporteur on the situation of human rights in Myanmar, Advance Unedited Version, UN Doc A/HRC/37/70 (9 March 2018).

The promotion, protection and enjoyment of human rights on the Internet, UNHRC, 47th Sess, UN Doc A/HRC/RES/47/16 (2021).

United Nations Secretary-General, "UN Strategy and Plan of Action on Hate Speech online" (2019), online (pdf): <ohchr.org/en/special-procedures/sr-religion-or-belief/hate-speech-and-incitement-hatred-or-violence>.

JURISPRUDENCE

Application of the convention on the prevention and punishment of the crime of genocide (The Republic of the Gambia v Myanmar) (2019).

Gonzalez v Google, No 18-16700 (9th Cir 2021).

Jane Doe v Meta Platforms Inc. No 3 2022cv07557 (ND Cal).

The Republic of the Gambia v Facebook, No 20-mc-36-JEB-ZMF (ND Cal 2021).

Twitter Inc. v Taamneh, No 21-1496 (9th Cir 2022).

SECONDARY MATERIAL

Aikins, Enoch Randy, "West Africa/ECOWAS" (2 December 2022), *ISS African Futures*, online: <futures.issafrica.org/geographic/regions/west-africa-ecowas/#cite-this-research>.

Alba, Davey & Adam Satariano, "At least 70 countries have had disinformation campaigns, study finds", *New York Times* (26 September 2019), online: <nytimes.com/2019/09/26/technology/government-disinformation-cyber-troops.html>.

Amnesty International, "Social Atrocity: Meta and the right to remedy for the Rohingya" (2022), online (pdf): *Amnesty International* <amnesty.org/en/documents/asa16/5933/2022/en/>.

——, "The Massacre in Axum" (26 February 2021), online: *Amnesty International* <amnesty.org/en/documents/afr25/3730/2021/en/>.

Anna, Cara, "Ethiopia now calls Axum massacre allegations 'credible'", *AP News* (3 March 2021), online: <apnews.com/article/abiy-ahmed-ethiopia-massacres-belgium-kenya4e5eda7bb2753973951269039d5ab802>.

Article 19, "Human Rights on the Internet" (15 July 2021), online: <article19.org/resources/un-human-rights-council-adopts-resolution-on-human-rights-on-the-internet/>.

——, "Response to the consultation of the UN Special Rapporteur on Freedom of Expression on her report on challenges to freedom of opinion and expression in times of conflicts and disturbances" (19 July 2022), online (pdf): <ohchr.org/sites/default/files/documents/issues/expression/cfis/conflict/2022-10-07/submission-disinformation-and-freedom-of-expression-during-armed-conflict-UNGA77-cso-article19.pdf>.

BBC Trending, "The country where Facebook posts whipped up hate", *BBC* (12 September 2019), online: <bbc.com/news/blogs-trending-45449938>.

Brown, Megan A, "Trendless Fluctuation? How Twitter's Ethiopia Interventions May (Not) Have Worked", *Tech Policy Press* (11 January 2022), online: <techpolicy.press/trendless-fluctuation-how-twitters-ethiopia-interventions-may-not-have-worked/>.

BSR, "Human Rights Impact Assessment: Facebook in Myanmar" (2018), online (pdf): <about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf>.

Bowers, John & Jonathan Zittrain, "Answering Impossible Questions: Content Governance in the Age of Disinformation" (2020) 1:1 Harv Kennedy Sch Misinformation Rev 1.

Burnham, Jack, "From the Internet to Ashes: Disinformation and the Tigray War", *Nato Association* (8 September 2022), online: <natoassociation.ca/from-the-internet-to-ashes-disinformation-and-the-tigray-war/>.

Dixon, S, "Global social networks ranked by number of users 2022" (26 July 2022), online: *Statistica* <statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>.

——, "Number of social media users worldwide from 2018 to 2027" (16 September 2020), online: *Statistica* <statista.com/statistics/278414/number-of-worldwide-social-network-users/#:~:text=In%202021%2C%20over%204.26%20billion,almost%20six%20billion%20in%202027>.

Douek, Evelyn, "Limits of International Law in Content Moderation" (2021) 6:1 UC IJIL 37.

——, "Verified Accountability: Self-regulation of content moderation as an answer to the special problems of speech regulation" (2019) Hoover Institution Series Paper 1903.

Dowling, Thomas, "Shooting the (Facebook) Messenger" (21 January 2019), online: *Tea Circle - A forum for new perspective on Myanmar* <teacircleoxford.com/essay/shooting-the-facebook-messenger-part-i/>.

Easterday, Jennifer, Hana Ivanhoe & Lisa Schirch, "Comparing Guidance for Tech Companies in Fragile and Conflict-Affected Situations" (2022), online (pdf): *TODA Peace Institute* <toda.org/policy-briefs-and-resources/policy-briefs/comparing-guidance-for-tech-companies-in-fragile-and-conflict-affected-situations.html>.

Facebook Oversight Board, "Case decision 2021-001-FB-FBR: Case Summary" (2021), online (pdf): <oversightboard.com/sr/decision/2021/001/pdf-english>.

Facebook Papers, "Facebook and Responsibility" (last visited 30 August 2023), online (pdf): <documentcloud.org/documents/21594152-tier2_rank_other_0320>.

Freedom House, Press Release, "Digital Election Interference Widespread in Countries Across the Democratic Spectrum" (7 December 2020), online: *Freedom House* <freedomhouse.org/article/report-digital-election-interference-widespread-countries-across-democratic-spectrum>.

Frosio, Giancarlo, "Platform Responsibility in the Digital Services Act: Constitutionalising, Regulating and Governing Private Ordering", forthcoming in Andrej Savin & Jan Trzaskowsi, eds, *Research Handbook on EU Internet Law* (Edward Elgar, 2021).

Gleicher, Nathaniel, "Removing Coordinated Inauthentic Behavior From Ethiopia" (16 June 2021), online: *Meta* <about.fb.com/news/2021/06/removing-coordinated-inauthentic-behavior-from-ethiopia/>.

Global Witness, "'Now is the time to kill': Facebook Continues to Approve Hate Speech Inciting Violence and Genocide During Civil War in Ethiopia" (June 2022), online (pdf): *Global Witness* <globalwitness.org/en/campaigns/digital-threats/ethiopia-hate-speech/>.

Google Support, "YouTube's Policies: Hate Speech Policy" (2022), online: YouTube <support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436>.

Guesmi, Haythem, "The social media myth about the Arab Spring", *Al Jazeera* (27 January 2021), online: <aljazeera.com/opinions/2021/1/27/the-social-media-myth-about-the-arab-spring>.

Human Rights Watch, "We Will Erase You From This Land: Crimes Against Humanity and Ethnic Cleansing in Ethiopia's Western Tigray Zone" (2022), online (pdf): *Human Rights Watch*

<hrw.org/sites/default/files/media_2022/04/ethiopia0422_web_1.pdf>.

Jackson, Jasper et al, "Facebook accused by survivors of letting activists incite ethnic massacres with hate and misinformation in Ethiopia", (20 February 2022), online: *The Bureau of Investigative Journalism* <thebureauinvestigates.com/stories/2022-02-20/facebook-accused-of-letting-activists-incite-ethnic-massacres-with-hate-and-misinformation-by-survivors-in-ethiopia>.

Kamer, Lars, "Internet usage in Africa" (17 November 2022), online: *Statistia* <statista.com/topics/9813/internet-usage-in africa/#:~:text=Africa's%20online%20shoppers%20amounted%20to,with%20the%20rising%20internet%20penetration.>.

Kang, Ceclia & Sheera Frenkel, *An Ugly Truth: Inside Facebook's Battle for Domination* (NY: Harper Collins, 2021).

Kannan, Prabha, "Digital Extractivism in Africa Mirrors Colonial Practices" (15 August 22), online: *Stanford University, Human-Centred Artificial Intelligence* <hai.stanford.edu/news/neema-iyer-digital-extractivism-africa-mirrors-colonial-practices>.

Keseru, Julia, "The EU's Digital Services Act Doesn't Go Far Enough" (16 May 2022), online: *CIGI* <cigionline.org/articles/data-rights-protections-are-overdue-for-an-upgrade/>.

Keykopour, Kevin & Uche Adegbite, "Establishing Twitter's presence in Africa" (12 April 2021), online: *Twitter Blog* <blog.twitter.com/en_us/topics/company/2021/establishing-twitter-s-presence-in-africa>.

Kinife Micheal Yilma, "On Disinformation, Elections and Ethiopian Law" (2021) 65:3 J Afr Law 351.

Klonick, Kate, "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression" (2020) 129:2418 Yale LJ 2418.

McTighe, Kristen, "A blogger at Arab Spring's genesis", *New York Times* (12 October 2011), online: <nytimes.com/2011/10/13/world/africa/a-blogger-at-arab-springs-genesis.html>.

Meta, "Oversight Bylaws" (last visited 31 August 2023), s 1.2.2, online (pdf): <about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf>.

——, "Oversight Board Charter" (last visited 31 August 2023), online (pdf): <about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf>.

Meta Newsroom, "An Update on Our Longstanding Work to Protect People in Ethiopia" (9 November 2021), online: *Meta* <about.fb.com/news/2021/11/update-on-ethiopia/>.

——, "Ethiopia: Preparing for Elections Day" (5 May 2021), online: *Meta* <facebook.com/gpa/blog/ethiopia-preparing-for-election-day?_rdc=2&_rdr>.

Meta's Transparency Centre, "Facebook Community Standards: Hate Speech" (2022), online: <transparency.fb.com/en-gb/policies/community-standards/hate-speech/>.

Morar, David, "The Digital Services Act's lesson for the U.S. policymakers: Co-regulatory mechanisms" (23 August 2022) online: *Brookings* <brookings.edu/blog/techtank/2022/08/23/the-digital-services-acts-lesson-for-u-s-policymakers-co-regulatory-mechanisms/>.

Neuburger, Jeffrey, "Important CDA Section 230 Case Lands in Supreme Court" (6 October 2022), online: *The National Law Review* <natlawreview.com/article/important-cda-section-230-case-lands-supreme-court-level-protection-afforded-modern>.

Oversight Board, "Oversight Board upholds Meta's decision in "Tigray Communication Affairs Bureau" case 2022-006-FB-MR" (October 2022), online: *Oversight Board* <oversightboard.com/news/592325135885870-oversight-board-upholds-meta-s-decision-in-tigray-communication-affairs-bureau-case-2022-006-fb-mr/>.

——, "Oversight Board upholds Meta's original decision: Case 2021-014-FB-UA" (December 2021), online: *Oversight Board* <oversightboard.com/news/927673894608838-oversight-board-upholds-meta-s-original-decision-case-2021-014-fb-ua/>.

Peguera, Miguel, "The platform neutrality conundrum and the Digital Services Act" (2022) 53 IIC 681.

Plaut, Martin, "New estimate of the Tigray death toll" (19 October 2022), online: *Martin Plaut* <martinplaut.com/2022/10/19/new-estimate-of-the-tigray-death-toll/>.

Rahimian, Tiran, *Whither International Law in Online Content Moderation?* (Bachelor of Civil Law & Juris Doctor, McGill University, 2019) [unpublished].

Sablosky, Jeffrey, "Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar" (2021) 43:6 Media Cult Soc 1017.

Salau, Torinmo, "How Twitter Failed Africa: Big Tech ignored policies that enable disinformation and propaganda across the continent", *Foreign Policy* (19 January 2022), online: <foreignpolicy.com/2022/01/19/twitter-africa-ghana-dorsey-disinformation/>.

Schliebs, Marcel et al, *China's Inauthentic UK Twitter Diplomacy: A Coordinated Network Amplifying PRC Diplomats* (Oxford, UK: Programme on Democracy & Technology, 2021).

Shaer, Matthew, "What Emotion Goes Viral the Fastest" (April 2014), online: *Smithsonian Magazine* <smithsonianmag.com/science-nature/what-emotion-goes-viral-fastest-180950182/>.

Shifa, Muna & Fabio Andres Diaz Pabon, "The Interaction of Mass Media and Social Media in Fueling Ethnic Violence in Ethiopia" (15 March 2022), online: *Accord* <accord.org.za/conflict-trends/the-interaction-of-mass-media-and-social-media-in-fuelling-ethnic-violence-in ethiopia/#:~:text=In%20Ethiopia%2C%20social%20media%20is,into%20mass%20atrocities%20and%20genocide.>.

Silva, Leandro et al, *Analyzing the Targets of Hate in Online Social Media, Proceedings of the Tenth International AAAI (Association for the Advancement of Artificial Intelligence) Conference on Web and Social Media* (2016).

Solomon, Salem, "Journalists struggle through information blackout in Ethiopia", *VOA* (2 December 2020), online: <voanews.com/a/press-freedom_journalists-struggle-through-information-blackout-ethiopia/6199045.html>.

Song, Sophie, "Internet in Myanmar remains slow, unstable and affordable to less than 1% of the population", *International*

Business Times (6 December 2013), online: <ibtimes.com/internet-myanmar-remains-slow-unstable-affordable-less-1-population-1402463>.

The Worldbank Data, "Mobile cellular subscriptions – Myanmar" (2022), online: <data.worldbank.org/indicator/IT.CEL.SETS?end=2020&locations=MM&start=1960&view=chart>.

Translators Without Borders, "Language data for Ethiopia" (last visited 30 August 2023), online: *Translators Without Borders* <translatorswithoutborders.org/language-data-for-ethiopia>.

Twitter Safety, "Given the imminent threat of physical harm, we've temporarily disabled Trends in Ethiopia. Alongside continued efforts to disrupt platform manipulation, we hope this measure will reduce the risks of coordination that could incite violence or cause harm." (5 November 2021 at 22:40), online: *Twitter* <twitter.com/TwitterSafety/status/1456813765387816965?s=20>.

Van Loo, Rory, "Federal Rules of Platform Procedure" (2021) 88: 4 U Chi L Rev 829.

Warren, T Camber, "Explosive Connections? Mass Media, Social Media, and the Geography of Collective Violence in African States" (2015) 52:3 J Peace Research 297.

Wilmot, Claire, Ellen Tveteraas & Alexi Drew, "Dueling Information Campaigns: The war over the narrative in Tigray" (20 August 2021), online: *Media Manipulation Casebook* <mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray#footnoteref4_98p4l98>.

Wong, David & Luciano Floridi, "Meta's Oversight Board: A Review and Critical Assessment" (2023) 33:261 Minds & Machines 261.

Yanagizawa-Drott, David, "Propaganda vs. Education: A Case Study of Hate Radio in Rwanda" in Jonathan Auerbach & Russ Castronovo, eds, *The Oxford Handbook of Propaganda Studies* (Oxford: Oxford University Press, 2013) 378.

Appendix

Comparative Table 1 – Digital Regulation Laws in Democracies (Europe)

	EU (DSA, 2022)	**UK** (Online Safety Bill -not passed)	**Germany** (Network Enforcement Act, 2017)	**France**
Nature of scheme	Structural regulations concerning illegal content moderation, algorithm transparency, reporting and audit mechanisms. Digital Services Coordinators for each Member State to supervise the application of the DSA and to enforce it where necessary.	Duty of care model; Regulations targeting online platforms in different ways, depending on their size (# of users). Risk assessments, published definitions of categories of prohibited content, notice and takedowns.	Structural regulations concerning removal of illegal content, fines for contraventions; user complaint system imposed.	No overarching scheme; Digital Republic Bill and Law Concerning Respect of the Principles of the Republic.
Obligations created	Obligations vary depending on size of online platform: Duty to take action against illegal content, due diligence obligations for transparency reporting, cooperation with national authorities, etc.	Duty on online platforms to moderate content, commission risk assessments, publish reports.	Obligation to remove "manifestly unlawful" content, falling under one of the listed offences in the German Penal Code.	N/A
Moderation obligations for illegal content or defined by other statute(s)?	Illegal content under EU law.	Illegal content and named categories of harmful content.	Illegal content, defined in the German Criminal Code.	Illegal content only, defined in the French Criminal Code and 1881 Press Act.
Liability exemptions	Prohibits general monitoring by platforms to protect users' fundamental freedoms article 24.	Duty of care rather than intermediary liability model.	Does not apply to platforms offering journalistic or editorial content; fewer reporting & complaints system requirements for platforms with < 2million users.	N/A

Comparative Table 2 – Digital Regulation Laws in Democracies (Americas and Africa)

	Brazil (Law of freedom, responsibility and transparency on the internet, Draft bill 2020)	**Canada** (Act to Address Online Harms)	**USA** (Communications Decency Act s.230)	**South Africa** (Films and Publications Amendment Act, 2022)	**Kenya** (Information and Communications (Amendment) Act, 2019)
Nature of scheme	Prohibits OSPs to remove content only by reference to their Community Guidelines, increases criminal penalties for libel and defamation.	Proposal to establish 3 bodies to oversee a regulatory regime whereby online service providers have obligations to take reasonable steps to make harmful content inaccessible. Monetary penalties for non-compliance.	Most prominent legislation is the addition of s.230 of title 47 of the US code (1996). Provides immunity from liability for online service providers.	Regulates online distribution of films, games and publications, including all user-generated content posted to OSPs.	Regulates OSPs and users through licencing requirements and platform rules.
Obligations created	OSPs must publish reports on content moderation decisions, remove 'fake news' and specified prohibited content.	Remove defined types of harmful content, reporting of illegal content.	Only to provide information on the kinds of parental control protections are commercially available to users.	Prohibits OSP users from sharing certain types of content (pornography, violent video, etc).	Requires licences for OSPs, physical officers in the country, and obligates them to share user data upon request.
Moderation obligations	Yes; specified in the Bill.	Yes; harmful content specified in the bill.	Some illegal content only (copyright or sex trafficking/child pornography).	N/A	Yes; penalties and fines for illegal content.
Liability exemptions	N/A	Only responsible to take "reasonable action."	s.230(c)(2) provides immunity from liability for any take down or decision not to take down content.	N/A	N/A

Comparative Table 3 – Digital Regulation Laws in Democracies (Asia & Oceania)

	Australia (Online Safety Act, 2021)	**Singapore** (Broadcasting Act, 1994)	**India** (Intermediary Guidelines and Digital Media Ethics Code, 2021)	**Indonesia** (Ministerial Regulation 5 of 2020 and Ministerial Regulation 10 of 2021)
Nature of scheme	Grants E-safety commissioner with new powers to enact regulatory legislation to promote online safety; new regulations and civil penalties.	Codes of Practice used to regulate broadcasting, only allows "acceptable content" as defined under Protection from Online Falsehoods and Manipulation Act, 2019.	Introduces take-down rules and increases liability risks for intermediaries who do not comply.	Requires all digital services to register with the government or face fines.
Obligations created	OSPs must take down specific types of defined content.	OSPs must take down content Minister declares an offence.	OSPs must publish rules on prohibited content.	OSPs must take down content that the government deems unlawful.
Moderation obligations	Harmful content defined in the Online Safety Act.	Yes; defined types of content in POFMA.	Yes; remove illegal content within 36 hours of takedown order.	Yes; remove content government deems unlawful or face fines.
Liability exemptions	N/A	N/A	Yes; if OSPs take all "reasonable and practicable measures" to remove or disable access to prohibited content.	Yes; liability exemption policy for e-commerce platforms, only responsible for content if they are unable to prove content was generated by users.