BSG Working Paper Series
Providing access to the latest policy-relevant research

OXFORD INSTITUTE FOR ETHICS, LAW AND ARMED CONFLICT

BLAVATNIK SCHOOL OF GOVERNMENT

UNIVERSITY OF OXFORD

Tackling Online Hate Speech through Content Moderation

The Legal Framework under the ICCPR

Talita Dias

just speech

Jesus College OXFORD

BSG Working Paper Series
Forthcoming in Bahador, Hammer and Livingston (eds), Countering online hate and its offline consequences in conflict-fragile settings (2022)

Tackling Online Hate Speech through Content Moderation:

The Legal Framework under the International Covenant on Civil and Political Rights

Talita Dias*

Abstract: Hate speech has had unprecedented consequences in the digital age. Despite being a global problem, international legal responses to it have been slow-coming and patchy. One of the core international legal instruments on the matter is the International Covenant on Civil and Political Rights (ICCPR). While generally protecting freedom of expression, Article 19(3) of the ICCPR does allow limitations to speech insofar as these are provided by law, necessary and proportionate for a legitimate reason. Likewise, Article 20 requires states parties to prohibit by law war propaganda as well as any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. This paper builds on the existing work of United Nations bodies to provide more granular and well-calibrated guidance on the application of these provisions to online hate speech, particularly in fragile settings. First, it clarifies why international human rights law matters for states, companies and civil society organizations operating in this space. Second, it provides a taxonomy of online hate speech, based on the distinct legal consequences that different forms of hate speech attract under Articles 19 and 20 of the ICCPR. Three

* Shaw Foundation Research Fellow in Law, Jesus College, University of Oxford. Contact: talita.desouzadias@jesus.ox.ac.uk. I am grateful to Laura Livingston for her comments on an earlier version of this chapter and to Sahil Thapa for his research assistance.

categories are proposed, along with respective content moderation measures: a) prohibited, b) limited, and c) free speech acts. The paper then applies this framework to fragile settings.

1. Introduction

Hate speech is not a new phenomenon but it has had unprecedented consequences in the digital age. From developed to developing countries and war-torn regions, online hate speech has led to violence and discrimination against individuals and groups on the basis of race, ethnicity, religion, nationality, social status, gender, sexual orientation, disability, and other characteristics. These, in turn, might escalate armed conflict and hamper economic development, especially in developing countries (see Section 1). Yet international legal rules on the matter are general and patchy, lacking specific guidance as to how online hate speech ought to be moderated on social media platforms.

While Article 19 of the International Covenant on Civil and Political Rights (ICCPR) protects freedom of expression, it provides that speech may be limited to protect the 'rights or reputations of others' as well as 'national security', 'public order, or 'public health or morals'. Likewise, Article 20 of the ICCPR prohibits war propaganda and 'advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence'. However, none of those concepts is subsequently defined in the Covenant. Granted, various United Nations (UN) Special Rapporteurs and the Human Rights Committee have issued some guidance on the implementation of those provisions. But only a few paragraphs deal *specifically* with content moderation of online hate speech on social media and other digital platforms. Moreover, the existing guidance fails to reconcile different regional approaches to limiting freedom of expression, such as constitutional divergences between North America and Europe, as well as Eastern and Western legal traditions.

This chapter is a starting point in filling this legal gap. It builds on the existing work of relevant UN bodies to provide more granular and well-calibrated guidance on how States, online platforms, civil society organizations and other stakeholders should refer to the ICCPR when designing and implementing content moderation policies to tackle online hate speech, with due deference to

regional and other contextual niceties. To this end, the chapter first clarifies the types and the jurisdictional reach of States' human rights obligations under the ICCPR, contrasting them with business responsibilities and the roles of other actors in giving effect to the Covenant and other international instruments. Next, the chapter devises a taxonomy of online hate speech and assesses the rules and measures applicable to their moderation. Specifically, following the legal framework laid down in Articles 20 and 19 of the ICCPR, online hate speech is divided into prohibited, limited, and free speech acts. Based on their likely impact on victims and perpetrators, their context, reach and the author's intention and/or position, different content moderation measures to prevent, mitigate and/or redress online hate speech are proposed for States, tech companies and other stakeholders (see also Section 2). Content moderation is broadly understood as any measure taken to review and manage speech acts online, whether to limit or advance such acts. In this sense, it gives effect to the applicable legal framework. Lastly, the chapter provides specific recommendations for moderating online hate speech in fragile settings.

2. The role of States and non-State actors in giving effect to Articles 19 and 20 of the ICCPR

States are the primary subjects of international law and, as such, are bound by international human rights law, which is found in rules of customary international law as well as international and regional treaties. The ICCPR is an international treaty open to ratification by all States. As a treaty, it is binding on its parties, which include, at present, 173 States in all regions of the world (UN Office of the High Commissioner for Human Rights 2021). Thus, the individual human rights recognized in the Covenant must, first and foremost, be upheld by the States parties thereto. Furthermore, most provisions of the ICCPR, at least on a *general level*, reflect customary international law (Lowe 2013, 535 and 537), which binds all States, irrespective of treaty ratification. This is the case of the provisions most directly implicated by online hate speech, namely, Article 19 of the ICCPR, recognizing the right to freedom of expression (ARTICLE 19 2003, 3; Howe 2017, 12),[1] and Article 20(2) of the ICCPR, prohibiting certain types of particularly harmful speech, i.e., incitement to discrimination, hostility or violence (Callamard 2008, 7, fn 15).[2]

According to Article 2 of the ICCPR, States have two different types of obligations vis-à-vis individual rights-holders. First, they must *respect* human rights, meaning they must refrain from violating such rights through their agents (i.e., negative human rights obligations). Second, States must *protect* human rights

from violations by third parties, i.e., they must prevent, stop, and redress human rights infringements by private entities, individuals and third States (so-called positive human rights obligations) (UN Human Rights Committee 2004, paras 5-8). The obligation to protect human rights does *not* require States to do the impossible to *successfully* prevent or stop such violations but to exercise due diligence or their best efforts in adopting the necessary measures to achieve the relevant aim (UN Human Rights Committee 2004, para 8). In the context of online hate speech, this means that State agents must not only refrain from violating the relevant human rights themselves, particularly non-discrimination and freedom of expression. They must also take reasonable steps in seeking to ensure compliance with those rights by tech companies, individual users, and other public or private entities.

In contrast to customary international law, which applies universally, the obligations to respect and protect human rights under the ICCPR only apply within a State's territory and jurisdiction, in line with Article 2(1) of the Covenant. As noted by the Human Rights Committee, jurisdiction applies extraterritorially insofar as the State has effective control over the individual right-holder (UN Human Rights Committee 2004, para 10), the company whose activity foreseeably causes a human rights violation (UN Human Rights Committee 2019, para 22; Committee; see also Inter-American Court of Human Rights 2017, paras 101-102), or, more broadly, the enjoyment of the rights in question (UN Human Rights Committee 2019, para 63). Control over the enjoyment of rights is functional (Shany 2013), that is, it does not depend on physical proximity to the victim or the events in question but can include remote forms of control through information and communications technologies (see Bundesverfassungsgericht 2020).

Corporations do not (yet) have direct obligations under international law, including international human rights law and the ICCPR in particular. However, as outlined in the UN Guiding Principles on Business and Human Rights (also known as the 'Ruggie Principles'), companies should meet their social responsibilities, that is, their stakeholders' expectations, by voluntarily undertaking to respect human rights (UN Human Rights Council 2008, Principles 2, 11-14). This includes, among other things, the responsibility to exercise due diligence in preventing and mitigating their human rights impact, as well as to provide victims with the necessary remedy for any violation of their rights (UN Human Rights Council 2008, Principles 11, 13, 15).

One of the main challenges of moderating content online is the fragmented definition of hate speech and other types of illegal and harmful content

across different national laws and platform standards, which often forces platforms to err on the side on censorship (De Streel et al. 2020, at 40-41, 51-52). In this context, the ICCPR offers States and corporations an internationally recognized legal framework and a common language for tackling hate speech in the Internet's boundless environment (Hicks et al., 2021, at 1). Thus, while this contribution focusses on States' obligations under the ICCPR, the *substance* of those duties and the recommendations below – including the proposed taxonomy of online hate speech (Section 3) and respective content moderation measures (Section 4) – should be equally followed by social media platforms and other online platforms, such as search engines and private messaging applications. Facebook, for example, has pledged to follow international human rights instruments when developing its first human rights policy (Sissons 2021).

In the same vein, online platforms should turn to the ICCPR and other international human rights treaties when *challenging* State action or inaction in breach of human rights (see Aswad 2020), including by seeking assistance from other States and international institutions, such as the Human Rights Council. The ICCPR can be a powerful tool in the hands of platforms in at least three scenarios. First, when State actors themselves are the direct perpetrators of online hate speech acts, as has been the case in Myanmar (UN Human Rights Committee 2018, A/HRC/39/64, para 73), Brazil (Lum 2019), and the United States (Goodman and others 2021). Second, when States encroach upon users' freedom of expression to silence opposing or minority views under the pretext of combatting online hate speech (see Hicks et al. 2021), such as in Turkey (Article 19 2021), Israel (7amleh 2021), and Russia (Article 19 2021). Third, States may have insufficient laws on the matter, leaving platforms and users in the dark when it comes to human rights-compliant content moderation policies (see Zuckerberg 2019).

Non-governmental and civil society organizations have also played a prominent role in empowering users to fight State and corporate wrongdoing. In this context, the ICCPR and other international human rights treaties offer a common, universal language, enabling those organizations to engage in advocacy and awareness-raising campaigns across the globe. Notably, the ICCPR and other international treaties have been abundantly (and successfully) used in strategic human rights litigation before different international and domestic courts, either directly or indirectly through tort liability and other legal remedies (see, e.g., Amnesty International 2020; Kohl 2014). In this way, international law and the ICCPR, in particular, are not merely framed as 'State obligations' in a strict legal sense. They are also key legal and policy tools in

the hands of individuals, civil society organizations, governments, corporations, and other relevant stakeholders.

3. A taxonomy of online hate speech and corresponding measures under the ICCPR

'Hate speech' as such is not a legal term of art under international law, including in international human rights instruments such as the ICCPR (UN General Assembly 2019, A/74/486, para 1). In common parlance, it has been broadly defined as any expression of hatred, opprobrium, enmity, detestation, or dehumanization of an individual or group identified by a protected characteristic (ARTICLE 19 2015, 9-10), i.e. race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or another status (ICCPR, Article 26). Given the variety of content and likely effects falling within this definition, the key challenge of regulating hate speech under international human rights law lies in how to strike the appropriate balance between freedom of expression and other protected rights or interests, such as the rights to security, bodily integrity, non-discrimination, health, and reputation. In the online environment, this challenge is compounded by the speed, scale and directness with which content is disseminated by individual Internet users. On the one hand, information and communications technologies may massively increase opportunities for expressing one's views freely and exercising other individual freedoms so dependent, such as the rights to freedom of opinion, to participate in elections and other democratic processes, and to protest. On the other hand, the pervasive nature of the Internet and its various applications may also amplify the negative impact of hate speech, leading to greater hostility, division, and violence in societies (Bieńkowski, Soral and Bilewicz 2021).

Despite those challenges, and the difficulty of striking the right balance between freedom and protection from harm, international human rights law and the ICCPR, in particular, do provide the appropriate baseline according to which hate speech can be classified and tackled, with the necessary adaptations for the online environment and its various service providers.

a. Prohibited Online Hate Speech

The first type of online hate speech that falls within the scope of the ICCPR is captured by Article 20. This provision *requires* States to *prohibit* any type of speech that constitutes propaganda for war, or advocacy for hatred that incites others to discriminate or perpetrate hostile or violent acts against individuals on the basis of nationality, race or religion. As the UN Special Rapporteur for Freedom of Expression has aptly noted, in light of the

growing list of individual or group characteristics protected under international human rights law, prohibited speech ought to be expanded to include incitement to discrimination, hostility or violence on the basis of 'race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity or intersex status' (UN General Assembly 2019, A/74/486, para 9). Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) complements this provision by requiring States to condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one color or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, with due regard to the freedoms of opinion and expression. This includes an obligation to criminalize all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another color or ethnic origin, including the provision of any assistance to racist activities and their financing.

The prohibition of propaganda for war is justified by its historical role in spurring armed conflict and its disastrous consequences for all human rights. In the same vein, the reason behind prohibiting incitement to violence, hostility and discrimination, as well as expressions of racial superiority, lies in their dangerous closeness and potential to contribute to harmful acts that may threaten or violate the victim's rights to life, bodily integrity, non-discrimination, among others. This is in recognition of the well-tested assumption that incitement is not 'just speech' but may and has encouraged crime and other human rights abuses (UN Secretary-General 2019, 1). The reality is that not all individuals or groups who are instigated to commit discrimination, hostility or violence are sufficiently aware of the wrongfulness of the acts in question or resilient enough to resist them. Likewise, especially in societies marked by systemic inequality and discrimination, many vulnerable individuals or groups targeted by incitement to hatred, hostility or violence are not sufficiently empowered to counter or speak out against such acts, as well as to hold those responsible accountable (UN Special Rapporteur on violence against women, its causes and consequences 2018, paras 28-29, 50; Dangerous Speech Project 2021, 19-20; see also Beurger, 2021). In short, enabling individuals to freely incite others to commit discrimination, hostile or violent acts might significantly encroach upon the human rights of vulnerable individuals, particularly by having a chilling effect on

their ability and willingness to speak freely against dominant narratives (UN Human Rights Council 2018, A/HRC/38/47, paras 29, 73; UN Human Rights Council 2018, A/HRC/38/35, para 27).

Yet, because these are still limitations on freedom of expression, such prohibitions must be strictly provided by law, in clear and accessible terms (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, paras 18, 22). Aside from the most serious forms of incitement and the dissemination of ideas of racial superiority, prohibited speech *need not* be criminalized. But if it is, States must be able to justify the necessity and proportionality of the crime and its punishment, as well as to observe additional requirements of legality in the criminal law, including non-retroactivity and specificity (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 29; UN General Assembly 2019, A/74/486, paras 8, 14, 18; UN Human Rights Committee 2011, paras 50-51). Thus, it is not enough for States to generally prohibit online hate speech that constitutes incitement or war propaganda. Laws, whether civil, administrative, or criminal, must *specifically* define what constitutes incitement, advocacy, and propaganda in this context (UN General Assembly 2019, paras 46-50; UN Human Rights Council 2018, A/HRC/38/35, para 26). As the ordinary (and legal) meaning of all three concepts suggests, although the relevant speech acts need not lead to or contribute to a specific result, both intentional conduct and the creation of an imminent risk of war, discrimination, hostility, or violence are necessary for the prohibition incitement and propaganda (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 29(c) and (f) and fn 5). Simply put, these are inchoate acts, that is, incomplete conduct that need not cause or contribute to a result. Therefore, nothing short of intention must be required for their prohibition, especially if the acts in question are criminalized. Similarly, any laws prohibiting hate speech online or offline must tightly define the content of prohibited speech, i.e., war, hatred, hostility, and violence. While war and violence imply physical or kinetic harm, hatred and hostility comprise intense and irrational expressions of opprobrium, enmity, and detestation towards the target group, whether physical or emotional (ARTICLE 19 2009, Principle 12).

Crucially, the fact that certain types of hateful or violent content must be prohibited, whether under criminal, civil or administrative law, means that States must not only sanction them but also actively seek to prevent or halt them. After all, the very purpose of outlawing and sanctioning conduct is to deter and prevent it. However, as noted in the Rabat Plan of Action on the prohibition of incitement to discrimination, hostility or violence, States should distinguish between criminal and

civil forms of prohibited speech (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 12). This division should be based on the seriousness of the speech act, including, in particular, the visibility of the speaker, the risk of harm ensuing, and the vulnerability of victims in context (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 12 and Appendix, paras 20, 29, 34).

In the online environment, this means that States must ensure that online platforms proactively moderate prohibited content, if possible, *before* it is amplified and becomes inevitably viral, thanks to platform recommendation algorithms (Chua 2017) that are geared towards engagement (O'Neil 2016, 180-185, Access Now 2018, 16; Eisenstat 2020; Nicas 2018). When it is *manifestly* clear that the content in question amounts to prohibited speech according to ICCPR-compliant national laws, especially criminal provisions, it should be taken down as promptly as possible (European Commission 2018, Preamble, 3, 5, 24-26). Examples include direct calls to physically assault, harm or kill members of a protected group, such as speech acts amounting to the offence of direct and public incitement to commit genocide against a national, ethnic, racial, or religious group (Convention on the Prevention and Punishment of the Crime of Genocide 1948, Article 3(c); UN General Assembly 2019, para 25).

The imminent risk of harm posed by such clear expressions of advocacy for hatred justifies a precautionary approach to their moderation. Thus, the publication and mass dissemination of expressions or images known to incite discrimination, hostility or violence against individuals or groups on social media should be promptly halted either by a human moderator or an automated system, until a more in-depth analysis of the content justifies its release. *Most* instances of *manifestly* prohibited speech can be initially flagged by artificial intelligence (AI) applications, provided these are sufficiently and regularly trained with the appropriate datasets in different linguistic, cultural and/or regional contexts (see Wijeratne 2020; De Streel et al. 2020, 49, 57; Hao 2021; Hao 2019; Cambridge Consultants 2019, 4-5). These applications include image and speech recognition, or natural language processing (UN General Assembly 2018, A/73/348, paras 1, 13-14; European Commission 2018, 24-25, 36-37, and paras 18, 36; Singh 2019, 12-16). Notably, existing language repositories or hate speech lexicons in different languages (see, e.g., PeaceTech Lab 2021; Wijeratne 2020) could be used to feed relevant AI datasets.

As both human moderators and automated moderation systems need clear and specific standards to moderate content effectively, States should require platforms to keep a public database or

repository of examples of prohibited speech with a brief explanation of their likely or actual impact (UN Human Rights Council 2018, A/HRC/38/35, paras 40, 46, 52, 63). Such explanations could be drawn from the platform's own prior experience, expert research, or civil society input (UN Human Rights Council 2018, A/HRC/38/35, paras 54-55). Nevertheless, precaution cannot justify blanket censorship and indiscriminate content takedowns. Even when it seems clear – to a person or a machine – that a certain instance of online speech constitutes propaganda for war or incitement to discrimination, hostility or violence, such an assessment ultimately depends on linguistic, contextual and common-sense analysis that a computer algorithm – whether powered by AI or not – is currently unable to make (Mitchell 2019, 33-35, 69-70, 108, 136; Boden 2016, 40-44, 56). For instance, it may be that a user seeks to protest, oppose or warn against the impact of prohibited speech by quoting said speech (UN General Assembly 2019, A/74/486, para 10). Thus, content takedown decisions, no matter how prompt, must always be ultimately and meaningfully verified by a human moderator (Singh 2019, 25), especially when prohibited content has been initially detected by an automated system (European Commission 2018, 27, para 20; UN General Assembly 2019, A/74/486, para 50; De Streel et al. 2020, 45, 54). Such human reviews must also provide the author with an opportunity to express their views and challenge the decision, either before or after the initial decision to remove the content has been made (De Streel et al. 2020, 49-50). Granted, in many fragile settings, it is harder to recruit human moderators versed in local languages and rare dialects (see, e.g., Marinescu 2021, on the difficulty of moderating hate speech against the Roma in Romanian; and Wijeratne 2020, on the challenge of moderating content in Sinhala, Sri Lanka's most spoken language). Yet sharing this burden with recognized civil society groups and other qualified users might be a way to scale up content moderation in those contexts. Importantly, in line with Article 2(3) of the ICCPR, judicial avenues and remedies must be available to individuals seeking to challenge platform moderation decisions and demand appropriate compensation (European Commission 2018, 22; UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, paras 10, 27-28; UN General Assembly 2019, A/74/486, paras 33, 55, 57(e)).

At the same time, where such a clear and prompt assessment is not possible, and doubts exist as to whether a certain instance of online hate speech is prohibited or not, one must err on the side of freedom and carefully calibrate any response against the risk of potential harm: content should not be taken down but de-prioritized and/or tagged as

potentially prohibited speech, with the necessary warning to viewers (Associated Press 2020) or perhaps a platform non-endorsement message. Such tags or labels should remain at least until a careful analysis of the material indicates that it is indeed adequate, necessary, and proportionate to remove the content or the label.

b. Limited Online Hate Speech

Whilst protecting the rights to freedom of opinion and expression, Article 19 of the ICCPR *allows* (rather than requires) States to *limit* (without necessarily prohibiting) certain kinds of harmful speech, including online hate speech, provided that certain safeguards are observed. Hateful expression falling under this category includes 'attacks or uses [of] pejorative or discriminatory language' based on an individual or group characteristic (UN General Assembly 2019, A/74/486, para 19). The first of those safeguards is that limitations on speech must be legitimate, that is, grounded in one of the specific reasons laid down in Article 19(3). These are: '(a) [f]or *respect of the rights or reputations of others*;' and '(b) [f]or the *protection of national security or of public order (ordre public), or of public health or morals*' (emphasis added). In contrast to prohibited speech (for which the imminent risk of harm tips the balance in favor of protection), in cases of limited speech, limitations must be the exception and freedom of expression the rule. As noted by the Human Rights Committee, offensive, disrespectful, or pejorative content that does not advocate for war or incite discrimination, hostility, or violence, is, at least in principle, protected (UN Human Rights Committee 2011, para 22). Accordingly, the grounds for limiting these kinds of hate speech, online and offline, must be strictly interpreted (UN Human Rights Committee 2011, para 21). The underlying assumption is that such types of hate speech are *not* so proximate to harm to justify a general prohibition. Nevertheless, in some circumstances, it *may* be legitimate to prohibit, sanction or otherwise restrict certain forms of hateful rhetoric.

Precisely because of the exceptional nature of those limitations, they must also be clearly provided by law, in a way that is accessible to lay individuals (UN Human Rights Committee 2011, 24-27; UN General Assembly 2019, A/74/486, paras 6(a), 20, 31-33; UN Human Rights Council 2018, A/HRC/38/35, paras 7, 46). Legality is an important safeguard against arbitrary limitations to freedom of expression, whether by States or private entities. As with prohibited speech, it requires both the types of limited speech and their respective limiting measures to be specified in law. In the context of online hate speech, this means that it is not enough for social media companies to

regulate their published content by devising community standards or guidelines (UN Human Rights Council 2018, A/HRC/38/35, paras 24, 26, 40, 46). Limitations to otherwise protected speech in the public sphere can only be made by law and are, thus, a State prerogative, albeit a tightly constrained one (UN Human Rights Council 2018, A/HRC/38/35, para 1). Dominant social media platforms such as Facebook, Twitter and YouTube have now become part and parcel of the digital public space. Thus, States must enact laws that specifically lay down which types of online hate speech may be subject to restrictions, what types of limitations are permitted, such as content takedowns, labelling or de-prioritization, in what circumstances they are warranted, and how platforms should apply them.

In keeping with the exceptional nature of limitations to otherwise protected speech, Article 19(3) of the ICCPR also requires those limitations to be necessary. Necessity, in this context, has been understood as a two-part test requiring States to assess not only whether the limitation is the least restrictive means to fulfil the legitimate aim(s) but also whether the restriction to free speech is proportionate to the right or interest it aims to protect (UN Human Rights Committee 2011, paras 22, 33; UN General Assembly 2019, A/74/486, paras 6(c), 51; UN Human Rights Council 2018, A/HRC/38/35, paras 7, 28, 44-45). In other words, the type of limitation chosen must be carefully calibrated to the importance of the legitimate aim protected. For instance, in situations of armed conflict or violent confrontations (see Nassiwa 2021; Nyheim and Veisalova 2021), it may be necessary and proportionate to temporarily prohibit and take down vague, short-of-incitement, expressions of hatred against an individual or group to prevent the escalation of violence. Conversely, in more resilient societies where different groups are empowered, it may suffice to simply tag or label an instance of limited speech with a warning to viewers and/or a platform non-endorsement message. However, these are not easy decisions. For instance, in many social settings, it may be necessary to prohibit the denial of certain historical facts, such as the Holocaust in Europe, whereas, in others, the risk of harm ensuing from such types of hate speech is negligible to justify any limitation (see Bazyler 2017, 184; but see *contra* UN General Assembly 2019, A/74/486, para 22 and ARTICLE 19 2015, 32-33). Likewise, images or symbols that are innocuous in one place may constitute expressions of hatred or group superiority, such as the monkey emoji (Burge 2021). And all these assessments may change over time within the same social settings, if events or situations that are prone to generate violence, harm or division emerge. This is the case with elections in otherwise stable environments, as well as situations of

political transition and armed conflict in more fragile settings.

Although necessity and proportionality are essential legal parameters, they are, in and of themselves, too broad to guide daily online content moderation decisions of big and small platforms alike. Thus, more granular criteria have been proposed for the application of necessity and proportionality to limitations on online hate speech. These include a) the socio-historical context, such as whether the same or similar content was used to spur violence or discrimination in the past; b) the position of the speaker, i.e., the more powerful or popular the speaker, the greater the likelihood that their speech will influence the audience's attitudes; c) the target audience, including its resilience or susceptibility to act upon or be persuaded by the speech (see Warren 2021; Nelson and Gilberds 2021); d) the degree of hatred expressed, including the directness or specificity of the speech act; and e) the means of dissemination, i.e. the more massive the medium used the more necessary and proportionate it may be to restrict the content (Benesch 2013, 2-6; Dangerous Speech Project 2021, 19-23). On digital platforms, the role of AI-powered algorithms in amplifying hateful or violent content, including when hashtags or bots are used to increase the number of clicks, must also be taken into account (see Daily Sabah 2021).

To this list of factors, I would add the vulnerability of the individuals or groups targeted by the relevant speech act(s): the more vulnerable the victims, the more likely it will be that hate speech will affect their mental wellbeing and the enjoyment of other rights, including by leading to self-censorship (UN Human Rights Council 2018, A/HRC/38/47, para 29; Abeyesekera and Cain 1992, 238, 242-243). Vulnerability, in this context, may be defined as the systemic and/or historical denial of rights or interests on the basis of protected characteristics (see Peroni and Timmer 2013, 1058-1060). Groups that are particularly vulnerable to online hate speech include women, children, racial or ethnic minorities, such as indigenous populations, persons with disabilities, and members of the LGBTQ+ community (UN Human Rights Council 2018, A/HRC/38/35, para 27; UN General Assembly 2019, A/74/486, para 25). The special vulnerability of some of these groups is reflected in the adoption of specific human rights treaties seeking to protect them, such as the Convention on the Elimination of All Forms of Discrimination Against Women, the Convention on the Rights of the Child, the International Convention on the Elimination of All Forms of Racial Discrimination, and the Convention on the Rights of Persons with Disabilities. Although not yet binding, the Declaration on the Rights of Indigenous Peoples (UN General Assembly 2007) and the Human

Rights Council Resolution on Human Rights Protection against violence and discrimination based on sexual orientation and gender identity (UN Human Rights Council 2016) indicate that there is growing consensus among States that such groups deserve special protection under international law.

As the foregoing analysis indicates, clearly spelling out in law the scope and rationale of online hate speech restrictions will not remove the complexity of the moral judgements required to make such content moderation decisions. This means that they should never be subject to automated content takedowns, even if temporary and no matter how advanced the moderation technology purports to be (UN General Assembly 2019, A/74/486, para 34). It bears recalling that, when it comes to limited speech, one must err on the side of freedom and that algorithms, including machine-learning ones, can only make quantitative decisions, not qualitative ones (Mitchell 2019, 70-72, 122). Thus, human moderation, coupled with the necessary anti-bias and discrimination training, is indispensable in those cases (see, e.g., Facebook 2021). But in a digital environment dominated by a handful of platforms, where the scale of content moderation is inevitably massive, it is difficult to trust that such companies will be making careful, independent decisions at scale (UN Human Rights Council 2018, A/HRC/38/35, paras 34, 58, 63). Even the most well-resourced social media companies, like Facebook and Twitter, have resisted calls to employ additional local moderators (Scheck, Purnell and Horwitz 2021; Barret 2020, 4, 19-20, 24-25), and struggled to find individuals who are sufficiently trained in the local languages and cultural niceties of the countries where they operate (Hicks et al. 2021; O'Neil 2021; Jee 2020). And this task is probably bound to fail: there will never be enough human moderators to carefully and independently look through the millions if not billions of suspected instances of online hate speech posted every day (Wijeratne 2020; Cambridge Consultants 2019, 4; Koebler and Cox 2018).

Given the inherent moral and practical challenges of moderating limited online hate speech at scale, platforms need to combine human judgement with scalable technologies. One way to achieve that, whilst ensuring greater representation and impartiality of decision-making bodies, is to harness existing technologies, such as consensus algorithms (Wikipedia Contributors 2021) to decentralize, democratize and humanize the moderation process. This idea is inspired by the jury system, which is used precisely when difficult moral decisions call for peer, common-sense judgement. Specifically, a certain number of individual platform users could be randomly selected to sit on country or

region-specific 'juries' and vote on the merits of decisions to keep, remove or otherwise limit the relevant content, following necessary instructions on applicable human rights standards. All decisions should be stored on a public and immutable chain, increasing their transparency and public scrutiny (Doubleday 2018). To avoid a pure majority-vote and the silencing of minority voices, particularly vulnerable or visible users, such as female and non-white journalists, activists, and politicians, should be able to nominate recognized, third-party content moderators to decide whether the content they author or receive may indeed be limited (de Souza Dias and Thapa 2021). Such moderators may include civil society organizations with expertise on the topic, academic institutions, and independent content hotlines, whose ongoing work should be further harnessed beyond the mere flagging of problematic content. As noted elsewhere, many of the largest platforms have not sufficiently involved end-users and civil society organizations in their content moderation processes (De Streel et al. 2020, 45-46, 51). In sum, content moderation processes should leverage the massive user pool of social media platforms, along with the work of numerous independent organizations to generate scalable yet fully human content-moderation decisions. In this way, they may wield greater independence, legitimacy, and representation than company oversight boards (see Paul 2021).

Nonetheless, even such peer decisions should always be open to judicial review, which requires close cooperation between social media platforms and domestic courts. In more detail, States should find technical solutions to integrate platform-based, out-of-court dispute resolution or conflict prevention mechanisms into their judicial system, perhaps by having readily accessible online court systems, where all evidence and submissions are presented online, and court decisions are directly implemented 'on-chain', i.e., on the platform (Susskind 2020; Lawtech UK 2021; Tworek et al. 2020, 9, 12-16). And it goes without saying that public oversight by civil society organizations remains an essential part of any content moderation exercise (UN Human Rights Council 2018, A/HRC/38/35, paras 51, 54, 56, 58, 70; UN Human Rights Council 2018, A/HRC/38/47, paras 86, 109). To enable such public scrutiny, not only should content moderation decisions by platforms and public courts be made public but companies and courts should also periodically publish reports summarizing the types of online hate speech taken down, tagged, redacted or otherwise limited in clear and accessible terms (UN Human Rights Council 2018, A/HRC/38/35, paras 38, 63).

c. Free Online Hate Speech

The third and final category of online hate speech regulated by the ICCPR is free speech. This category comprises hateful content that neither constitutes prohibited speech under Article 20 (propaganda for war and incitement to discrimination, hostility, or violence) nor merits limitation under Article 19(3) of the ICCPR (to protect the rights or reputations of others, national security, public order, or morals), considering all relevant factors (UN General Assembly 2019, A/74/486, para 24; UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 20). For these types of hate speech, online or offline, the benefits of exposure to public debate and critical thinking outweigh the risks of harm to other protected rights or interests (see Nyheim and Veisalova 2021; Vogt 2021). In most cases, these types of offensive or discriminatory expressions are directed not at natural persons, whether individuals or groups, but at *institutions*, such as a particular religion, its tenets or figures, a State or government. Because there is no direct expression of hatred towards rights-holders, these types of hate speech must be permitted as a general rule. For this reason, the Human Rights Committee has noted that blasphemy or treason laws are incompatible with the ICCPR to the extent that they prohibit criticism of a religious or public institution, whether offensive or not (UN Human Rights Committee 2011, para 48; UN Human Rights Council 2018, A/HRC/38/35, para 13). After all, human rights law is about protecting humans, not institutions per se. However, contextual and linguistic analysis, coupled with common sense and human judgment, remains important to ensure that an expression of hatred seemingly directed at an institution is not in fact targeting an individual or group and calls for some form of limitation.

In the online environment, this means that under the ICCPR 'institutional' hate speech enjoys a presumption of freedom and should not in principle be taken down or otherwise censored, whether by a human moderator or an automated system. Yet, if careful human analysis of the content reveals that it amounts to limited or prohibited speech, then the content should be appropriately de-prioritized, tagged, or removed. At the same time, one must recognize the potential of this type of hate speech to generate division and intolerance, especially in the online environment, where it can be easily amplified and constantly fed to like-minded individuals by recommendation algorithms (Wu 2016; Cinelli et al. 2021; Tufekci 2017). Therefore, States, platforms and civil society organizations should find ways to raise awareness of such risks, build resilience in societies, especially among vulnerable groups, and foster intercultural dialogue and tolerance (UN High Commissioner for

Human Rights 2013, A/HRC/22/17/Add.4, paras 35-39; UN General Assembly 2019, A/74/486, paras 24, 28, 54-55, 58(e); UN Human Rights Council 2018, A/HRC/38/47, paras 50, 87, 97, 110, 119). These measures have the potential to address the root causes of all types of online hate speech – prohibited, limited and free –, thereby contributing to their prevention (UN High Commissioner for Human Rights 2013, A/HRC/22/17/Add.4, para 37; see also Weiss 2021; Buerger 2021; Gichuhi 2021; Morrison 2021).

Although it is beyond the scope of this contribution to provide a comprehensive and detailed list of such preventive measures, particularly useful ones include: i) digital literacy campaigns about how social media and their algorithms work (UN General Assembly 2018, A/73/348, paras 57, 66); ii) educational campaigns (Weiss 2021) and public awareness courses on basic international human rights standards at schools, governmental institutions and tech companies; iii) changes in algorithmic design to promote positive types of engagement (Shaer 2014; Ucciferri and Marechal 2021), counter-speech (including reactive messages posted by chatbots, e.g. Cambridge Consultants 2019, 9) and a diversity of views on political, social and moral issues (UN General Assembly 2019, A/74/486, paras 18, 28, 51, 54, 58(f); Buerger 2021); iv) giving users more power over the content they want to receive, such as by opting out from platform-curated feeds (UN Human Rights Council 2018, A/HRC/38/35, paras 60-61); v) verifying and auditing social media algorithms, bearing in mind the need to protect proprietary rights and trade secrets (UN Human Rights Council 2018, A/HRC/38/35, para 56; Access Now 2018, 33-36); v) promoting the use of smaller, non-profit, open-source and/or decentralized social media companies, such as Diaspora, Minds and Mastodon (Meritt 2019; Roose 2018); and vi) piloting paid, ad-free version of their platforms, where users have even greater control over their feeds. Of course, though potentially beneficial, each of these measures may have different drawbacks, such as high design and implementation costs and lack of sufficient technical expertise. Thus, careful, case-by-case considerations ought to inform their selection, design, and implementation.

d. The Dos and Don'ts of Moderating Hate Speech in Fragile Settings

A key question that remains is how to apply Articles 19 and 20 of the ICCPR, along with the taxonomy and recommendations discussed earlier, to online hate speech in fragile settings, such as young democracies and conflict-affected States? The short answer is there is no one-size-fits-all approach or a perfect solution to moderating online hate speech in either developing or developed

countries. In fact, one struggles to find a single legislative model that has not been challenged on the basis of compliance with international human rights law (Hicks 2021, 1). Even Western democracies (on both sides of the Atlantic) have struggled with content moderation (see, e.g., Index on Censorship 2021, on the United Kingdom; ARTICLE 19 2021, on Italy; Schulz 2020, on France; Noyan 2021, on Germany; and Burwell 2021, on the United States). Thus, automatically importing their models to fragile settings could be a recipe for disaster.

Legal commentators have also disagreed on the right form and amount of platform regulation, reflecting fundamentally different traditions to balancing freedom of expression, non-discrimination, and other rights: the North American emphasis on freedom and self-regulation versus the protective, State-centric approach to safeguarding human dignity spearheaded in Europe and Asia (see, e.g., Siripurapu and Merrow 2021; Laub 2019; Klonick 2017). Notably, this disagreement manifests itself in the debates surrounding platform intermediary liability and duties of care: while some argue that platforms should be held liable for illegal content that they fail to remove, others propose the mere regulation of platform moderation processes rather than specific outcomes (De Streel et al. 2020, 53).

However, those difficulties are not insurmountable. Even though legal frameworks on online hate speech and content moderation measures should be tailored to the needs and resources of each State, including the platforms and civil society organizations that operate therein, there is a minimum common denominator of 'dos and don'ts' which should be followed in fragile settings to ensure consistency with the ICCPR.

The starting point of *any* regulatory model of content moderation, whether based on some form of intermediary liability or a duty of care, is an accessible, foreseeable, and sufficiently clear definition of different forms of online hate speech and the respective measures that platforms, users and other stakeholders may or must implement. Overly broad definitions and wide platform discretion, especially when coupled with intermediary liability and high penalties for failing to remove content, may legitimize private censorship (Index on Censorship 2021, 4 and 11), which can significantly hinder the development of young democracies. On the other hand, the lack of relevant legal definitions and requirements for content moderation may result in the proliferation of hate or otherwise harmful speech, and government abuse of content takedown requests, especially in conflict-affected or divided States (UN General Assembly 2019, A/74/486, para 31). Overall, regulatory models should ensure that the

burden and responsibility for posting and moderating content are shared and balanced across platforms, users, public entities, and other relevant stakeholders (De Streel et al. 2020, 54).

As hinted at earlier, a robust legal framework for online hate speech must be accompanied by judicial remedies against errors made by platforms, independent moderators, or States in the classification and treatment of online hate speech. As mentioned earlier, content moderation, whether by humans or machines, is not foolproof. Errors such as wrongful content takedowns and failure to remove or limit content are inevitable. Thus, it is indispensable that the judiciary has the final word, as it normally does in human rights or constitutional issues (Hicks et al., 2021, 10). One incipient case in point is Germany, which has amended its NetzDG law to provide for an appeals system and arbitration tribunals to hear content moderation disputes (Library of Congress 2021), following initial criticism for its lack of provisions on judicial oversight (Human Rights Watch 2018). The scale and speed of online hate speech and other forms of illegal content will be a challenge. However, further engagement of independent bodies in content moderation processes, such as the use of trusted flaggers (Digital Europe 2021; De Streel et al. 47, 79) and recognized moderators, may avoid the need for judicial dispute settlement, whereas 'online content' courts could be a cost-effective solution to adjudicate those disputes in fragile settings (see De Streel et al. 2020, 51, 55; Tworek et al. 2020, 9, 12-16). Independent oversight boards, such as Facebook's, are a positive step in this direction (Milanovic 2021). But they ought to be integrated with or complemented by proper judicial bodies (UN General Assembly 2019, A/74/486, paras 33, 57(e)).

Relatedly, users on both sides of the table, i.e., content authors and recipients, as well as platforms themselves and competent regulatory bodies, must be able to *challenge* content moderation outcomes, including before domestic courts. Indeed, many platforms have already put in place such 'counter-notice' or appeals processes, deemed essential to curb abusive or unsubstantiated flagging practices and to safeguard authors' right to freedom of expression (De Streel et al. 2020, 48-49). Likewise, the anonymity of those reporting and flagging content must be safeguarded, except in cases of defamation or copyright infringement (De Streel et al. 2020, 51). Moreover, not just *substantive* standards but also the content moderation *process*, including notices and counter-notices, must be made accessible to users and relevant stakeholders (De Streel et al. 2020, 51). According to some, this is the case of the flagging interfaces designed by YouTube and Twitter, but not Facebook (Tworek and Leesen 2019, 5; Singh 2019, 25).

Lastly, ensuring transparency in the content moderation process is key in the implementation of any chosen legal framework. Good examples of transparency frameworks can be found in countries around the world, such as India's Guidelines for Intermediaries and Digital Media Ethics Code (Hicks et al. 2021, 2-3), and the European Union's forthcoming Digital Services Act (European Commission 2020, 2, 5, 11 and Section3). Transparency must come from all stakeholders involved, including platforms, public bodies, trusted flaggers, recognized moderators and the judiciary. To this end, reporting mechanisms must be put in place and enforced, such as information on flagged, limited, and removed content and the legal basis of the limiting measure adopted (see Santa Clara Principles 2018, Principle 1; De Streel et al. 2020, 50).

4. Conclusion

The ICCPR is one of the most important international human rights instruments to date, given its quasi-universal reach. Under the ICCPR and its customary counterparts, States have duties not only to respect human rights in the context of online hate speech but also to protect the human rights of individual victims and speakers. These obligations apply extraterritorially insofar as States are home to social media companies operating abroad or otherwise exercise some form of control over the enjoyment of relevant human rights, such as by controlling an online service or device like a computer server where online content is stored. Companies, including social media platforms, do not *yet* have binding obligations under international human rights law, but have a social responsibility to follow those universal standards, given the global reach of online platforms. The ICCPR is also a powerful tool in the hands of civil society organizations and individual users seeking to give effect to internationally recognized human rights around the world.

The ICCPR does not contain specific rules for online hate speech. But its general rules for prohibited speech under Article 20, limited speech under Article 19(3), and protected speech under Article 19(2) apply to the phenomenon and provide the baseline for its regulation by States and moderation by companies and other relevant stakeholders. Building on the interpretative guidance provided by the UN Human Rights Committee and the UN Special Rapporteurs on different topics, this Chapter recommends that States, companies and civil society organizations adapt Articles 19 and 20 of the ICCPR to the phenomenon of online hate speech. Specifically, States should enact legislation requiring public institutions and online platforms to adopt the necessary and proportionate technical, remedial, and educational measures to counter different

types of online hate speech – prohibited, limited and protected. To be sure, these do not exhaust the types of action needed to tackle online hate speech. Yet the interpretative framework and measures proposed in this Chapter can provide States, platforms, users, and civil society organizations with additional clarity over what needs to be done to counter online hate speech in line with Articles 19 and 20 of the ICCPR. Importantly, each of those stakeholders has a role to play in the design and implementation of a robust legal framework, effective remedies, and transparency mechanisms well-suited to moderate online hate speech in different fragile settings around the world.

[1] Although a few States parties to the ICCPR have made declarations or reservations with respect to Article 19, such as by reserving the right to license television or radio broadcasts or to derogate from this right in certain situations, no State has opposed the essence of the right to freedom of expression (see United Nations Treaty Collection 2021).

[2] Note that Article 20(1) ICCPR, requiring states to prohibit propaganda for war, has been the subject of reservations and declarations by several States, including Denmark, Finland, Liechtenstein, Iceland, Luxembourg, The Netherlands, and the Republic of Ireland. Although a few States have reserved the right to adopt legislation or further legislation, civil or criminal, prohibiting advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, pursuant to Article 20(2) ICCPR, or declared that this provision must be read in line with the right to freedom of expression in Article 19, no State party to the ICCPR has questioned the unlawfulness of this type of hate speech (see United Nations Treaty Collection 2021).

Bibliography

"Germany: Network Enforcement Act Amended to Better Fight Online Hate Speech." *Library of Congress*, 2021. https://www.loc.gov/item/global-legal-monitor/2021-07-06/germany-network-enforcement-act-amended-to-better-fight-online-hate-speech/.

"Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG)". *German Law Archive*, 2018. https://germanlawarchive.iuscomp.org/?p=1245.

"Pandemic of Bots: Half of Twitter trends are fake, says new study". 2021. *Daily Sabah*, June 3, 2021. https://www.dailysabah.com/life/science/pandemic-of-bots-half-of-twitter-trends-are-fake-says-new-study.

"The Santa Clara Principles on Transparency and Accountability in Content Moderation." As of September 27, 2021. https://santaclaraprinciples.org/.

7amleh. 2021. "The Attacks on Palestinian Digital Rights: Progress report." May 6-19, 2021. https://7amleh.org/storage/The%20Attacks%20on%20Palestinian%20Digital%20Rights.pdf.

Abeyesekera, Sunila and Cain, Kenneth L. 1992. "Incitement to Inter-Ethnic Hatred in Sri Lanka," in Coliver, Sandra. *Striking a Balance: Hate Speech, Freedom of Expression and Non-discrimination* (London and Essex: ARTICLE 19 and University of Essex, 1992): 238-244.

Access Now. 2018. "Human Rights in the Age of Artificial Intelligence." 8 November 2018. https://www.accessnow.org/human-rights-in-the-age-of-AI

Advisory Opinion OC-23/17 of November 15, 2017, Requested by the Republic of Colombia: The Environment and Human Rights (Inter-American Court of Human Rights November 15, 2017)

Amnesty International. 2020. "Going to court to protect the rights of refugees and migrants: An overlooked tool for positive change." August 18, 2020. https://www.amnesty.org/en/latest/research/2020/08/going-to-court-to-protect-the-rights-of-refugees-and-migrants/.

ARTICLE 19. 2003. "Statement on the Right to Communicate." Feb. 2003. As of September 27, 2021, https://www.article19.org/data/files/pdfs/publications/right-to-communicate.pdf

ARTICLE 19. 2009. "The Camden Principles on Freedom of Expression and Equality." April 2009. As of September 27, 2021, https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf.

ARTICLE 19. 2015. "Hate Speech Explained: A Toolkit." As of September 27, 2021, https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf.

ARTICLE 19. 2021. "Italy: New anti-discrimination bill must meet international free speech standards." July 15, 2021. https://www.article19.org/resources/italy-free-speech-standards-should-inform-debate-on-the-zan-bill/.

ARTICLE 19. 2021. "Turkey: Facebook and other companies 'in danger of becoming an instrument of state censorship.'" January 18, 2021. https://www.article19.org/resources/turkey-facebook-and-other-companies-in-danger-of-becoming-an-instrument-of-state-censorship/.

ARTICLE 19. 2021." Russia: Laws enabling massive online censorship must be repealed." February 15, 2021. https://www.article19.org/resources/russia-laws-enabling-massive-online-censorship-must-be-repealed/.

Associated Press. 2020. "Twitter will label and may remove media designed to mislead." *The Guardian*, February 5, 2020. https://www.theguardian.com/technology/2020/feb/04/twitter-label-remove-manipulated-media.

Aswad, Evelyn Mary. 2020. "To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?" *Washington and Lee Law Review* 77: 609-659.

Barret, Paul M. 2020. "Who Moderates the Social Media Giants? A Call to End Outsourcing." *New York University, Stern, Center for Business and Human Rights*, June 2020. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020.

Bazyler, Michael. 2017. *Holocaust, Genocide, and the Law: A Quest for Justice in a Post-Holocaust World*. Oxford, England: Oxford University Press.

Benesch, Susan. 2013. "Dangerous Speech: A Proposal to Prevent Group Violence." *Dangerous Speech Project*, February 23, 2013. https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf.

Boden, Margaret A. 2016. *AI: Its Nature and Future.* Oxford, England: Oxford University Press.

Bundesverfassungsgericht (BVerfG). 2020. Bundesnachrichtendienst Case, 1 BvR 2835/17. May 19, 2020. As of September 27, 2021, https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2020/bvg20-037.html.

Burge, Jeremy. 2021. "How the Monkey Emoji is Racist." *Emojipedia*, July 12, 2021. https://blog.emojipedia.org/how-the-monkey-emoji-is-racist/.

Burwell, Frances. 2021. "Free speech and online content: What can the US learn from Europe?" *Atlantic Council*, February 1, 2021. https://www.atlanticcouncil.org/blogs/new-atlanticist/free-speech-and-online-content-what-can-the-us-learn-from-europe/.

Callamard, Agnes. 2008. "Conference Room Paper # 2." Paper presented at the *Expert meeting on the links between articles 19 and 20 of the ICCPR: Freedom of expression and advocacy of religious hatred that constitutes incitement to discrimination, hostility or violence*, United Nations Office of the United Nations High Commissioner for Human Rights, Geneva, October 2-3, 2008. As of September 27, 2021, https://www.article19.org/data/files/pdfs/conferences/iccpr-links-between-articles-19-and-20.pdf.

Cambridge Consultants. 2019. "Use of AI in Online Content Moderation: A 2019 Report Produced on Behalf of OFCOM." As of September 27, 2021, https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

Chua, Roger. 2017. "A simple way to explain the Recommendation Engine in AI." *Medium*, June 26, 2017. https://medium.com/voice-tech-podcast/a-simple-way-to-explain-the-recommendation-engine-in-ai-d1a609f59d97

Cinelli, Matteo et al. 2021. "The echo chamber effect on social media." *Proceedings of the National Academy of Sciences* 118 (9): 1-8. https://doi.org/10.1073/pnas.2023301118

Dangerous Speech Project. 2021. "Dangerous Speech: A Practical Guide." As of September 27, 2021, https://dangerousspeech.org/wp-content/uploads/2020/08/Dangerous-Speech-A-Practical-Guide.pdf.

de Souza Dias, Talita and Thapa, Sahil. 2021. "Tackling Football-Related Online Hate Speech: The Role of International Human Rights Law: Parts I & II." *EJIL: Talk!*, July 30, 2021. https://www.ejiltalk.org/tackling-football-related-online-hate-speech-the-role-of-international-human-rights-law-part-i/; https://www.ejiltalk.org/tackling-football-

related-online-hate-speech-the-role-of-international-human-rights-law-part-ii/.

De Streel, Alexandre et al. 2020. "Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform." *European Parliament*, June 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/De Streel et al._STU(2020)652718_EN.pdf.

Digital Europe. 2021. "Digital Services Act Position Paper." March 31, 2021. https://www.digitaleurope.org/resources/digital-services-act-position-paper/.

Doubleday, Kevin. 2018. "Blockchain Immutability – Why Does it Matter?" *Fluree*, November 21, 2018. https://medium.com/fluree/immutability-and-the-enterprise-an-immense-value-proposition-98cd3bf900b1.

Eisenstat, Yaël. 2020. "Dear Facebook, this is how you're breaking democracy." Filmed August 2020 at TED2020. As of September 27, 2021, https://www.ted.com/talks/yael_eisenstat_dear_facebook_this_is_how_you_re_breaking_democracy.

European Commission. 2018. *Recommendation of 1.3.2018 on measures to effectively tackle illegal content online.* C(2018) 1177 final, March 1, 2018. As of September 27, 2021, https://digital-strategy.ec.europa.eu/en/library/commission-recommendation-measures-effectively-tackle-illegal-content-online.

European Commission. 2020. *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*. COM(2020) 825 final, December 15, 2020. As of September 27, 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en.

Facebook. 2021. "Managing Unconscious Bias". As of September 27, 2021, https://managingbias.fb.com/.

Goodman, Ryan, Dugas, Mari and Tonckens, Nicholas. 2021. "Incitement Timeline: Year of Trump's Actions Leading to the Attack on the Capitol." *Just Security*, January 11, 2021. https://www.justsecurity.org/74138/incitement-timeline-year-of-trumps-actions-leading-to-the-attack-on-the-capitol/.

Hao. 2019. "This is how AI bias really happens—and why it's so hard to fix." *MIT Technology Review*, February 4, 2019. https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

Hao. 2021. "AI still sucks at moderating hate speech." *MIT Technology Review*, June 4, 2021. https://webcache.googleusercontent.com/search?q=cache:7kXnm2GlA_0J:https://www.technologyreview.com/2021/06/04/1025742/ai-hate-speech-moderation/+&cd=1&hl=en&ct=clnk&gl=uk.

Hicks, Peggy et al. 2021. "Press briefing: Online content moderation and internet shutdowns." *UN Human Rights Office*, 14 July 2021. https://www.ohchr.org/Documents/Press/Press%20briefing_140721.pdf.

Howe, Emily. 2018. "Protecting the human right to freedom of expression in international law." *International Journal of Speech-Language Pathology*, 20 (1): 12–15. https://doi.org/10.1080/17549507.2018.1392612.

Human Rights Watch. 2018. "Germany: Flawed Social Media Law - NetzDG is Wrong Response to Online Abuse." February 14, 2018. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

Index on Censorship. 2021. "Right to Type: How the 'Duty of Care' model lacks evidence and will damage free speech." June 17, 2021. https://www.indexoncensorship.org/wp-content/uploads/2021/06/Index-on-Censorship-The-Problems-With-The-Duty-of-Care.pdf.

Jee, Charlotte. 2008. "Facebook needs 30,000 of its own content moderators, says a new report." *MIT Technology Review*, June 8, 2020. https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/.

Kohl, Uta. 2014. "Corporate human rights accountability: the objections of western governments to the Alien Tort Statute." *International and Comparative Law Quarterly*, 63(3): 665-697.

Klonick, Kate. 2017. "The New Governors: The People, Rules, and Processes Governing Online Speech." *Harvard Law Review*, 131(6):1598-1670.

Koebler, Jason and Cox, Joseph. 2018. "The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People." *Vice*, August 23, 2018. https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works.

Laub, Zachary. 2019. "Hate Speech on Social Media: Global Comparisons." *Council of Foreign Relations*, June 7, 2019. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

Lawtech UK. 2021. "UK Jurisdiction Taskforce: Digital Dispute Resolution Rules." April 22, 2021. As of September 27, 2021, https://technation.io/lawtech-uk-resources/#rules.

Lowe, Anne. 2013. "Customary International Law and International Human Rights Law: A Proposal for the Expansion of the Alien Tort Statute." *Indiana International & Comparative Law Review* 23 (3): 523-553. https://doi.org/10.18060/17886.

Lum, Kathryn. 2019. "The effects of Bolsonaro's hate speech on Brazil." Monitor Racism, January 2019. http://monitoracism.eu/the-rise-of-bolsorano/.

Marinescu, Delia. 2021. "Facebook's Content Moderation Language Barrier." *New America*, September 8, 2021. https://www.newamerica.org/the-thread/facebooks-content-moderation-language-barrier/.

Meritt, Tom. 2019. "Top 5 decentralized social networks." *TechRepublic*, March 21, 2019. https://www.nytimes.com/2018/03/28/technology/social-media-privacy.html.

Milanovic, Marko. 2021. "The Facebook Oversight Board Made the Right Call on the Trump Suspension." *EJIL: Talk!*, May 6, 2021. https://www.ejiltalk.org/the-facebook-oversight-board-made-the-right-call-on-the-trump-suspension/.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. Harlow, England: Penguin Books.

Nicas, Jack. "How YouTube Drives People to the Internet's Darkest Corners." *The Wall Street Journal*, February 7, 2018. https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.

Noyan, Oliver. 2021. "Germany's online hate speech law slammed by opposition, Commission." *Euractiv*, May 10, 2021. https://www.euractiv.com/section/digital/news/germanys-online-hate-speech-law-slammed-by-opposition-commission/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. Harlow, England: Penguin Books.

O'Neil, Cathy. 2021. "Facebook and Twitter Can't Police What Gets Posted." *Bloomberg*, February 19, 2021. https://www.bloomberg.com/opinion/articles/2021-02-19/facebook-and-twitter-content-moderation-is-failing.

Paul, Kari. 2021. "Facebook ruling on Trump renews criticism of oversight board." *The Guardian*, May 5, 2021. https://www.theguardian.com/technology/2021/may/05/facebook-oversight-board-donald-trump.

PeaceTech Lab. 2021. "PeaceTech Lab Lexicons." As of September 27, 2021, https://www.peacetechlab.org/toolbox-lexicons.

Peroni, Lourdes and Timmer, Alexandra. 2013. "Vulnerable groups: The promise of an emerging concept in European Human Rights Convention law." *International Journal of Constitutional Law* 11(4): 1056–1085, https://doi.org/10.1093/icon/mot042.

Roose, Kevin. 2018. "Can Social Media be Saved?" *The New York Times*, March 28, 2018. https://www.nytimes.com/2018/03/28/technology/social-media-privacy.html;

Scheck, Justin, Purnell, Newley, and Horwitz, Jeff. 2021. "Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show." *The Wall Street Journal*, September 16, 2021. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953?mod=article_inline.

Schulz, Jacob. 2020. "What's Going on With France's Online Hate Speech Law?" *Lawfare*, June 23, 2020. https://www.lawfareblog.com/whats-going-frances-online-hate-speech-law.

Shaer, Matthew. 2014. "What Emotion Goes Viral the Fastest?" *Smithsonian Magazine*, April 2014. https://www.smithsonianmag.com/science-nature/what-emotion-goes-viral-fastest-180950182/.

Shany, Yuval. 2013. "Taking Universality Seriously: A Functional Approach to Extraterritoriality in International Human Rights Law." *The Law & Ethics of Human Rights* 7 (1): 47-71. https://doi.org/10.1515/lehr-2013-00047 47.

Singh, Spandana. 2019. "Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User- Generated Content." New America Open Technology Institute, July 15, 2019. newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial- intelligence-moderate-user-generated-content/.

Siripurapu, Anshu and Merrow, William. 2021. "Social Media and Online Speech: How Should Countries Regulate Tech Giants?" *Council of Foreign Relations*, February 9, 2021. https://www.cfr.org/in-brief/social-media-and-online-speech-how-should-countries-regulate-tech-giants.

Sissons, Miranda. 2021. "Our Commitment to Human Rights." *Facebook*, March 16, 2021. https://about.fb.com/news/2021/03/our-commitment-to-human-rights/.

Susskind, Richard. 2020. "The Future of Courts." *Remote Courts* 6 (5). https://thepractice.law.harvard.edu/article/the-future-of-courts/.

The Human Rights Council. 2016. "Resolution on Human Rights Protection against violence and discrimination based on sexual orientation and gender identity", A/HRC/RES/3 (30 June 2016).

The United Nations General Assembly. 1948. "Convention on the Prevention and Punishment of the Crime of Genocide." Treaty Series 78 (December): 277.

Tufekci, Zeynep. 2017. "We're building a dystopia just to make people click on ads." Filmed September 17, 2017, at TEDGlobalNYC, New York. As of September 27, 2021, https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads.

Tworek, Heidi, et al. 2020. "Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective." *Transatlantic Working Group on Content Moderation Online and Freedom of Expression*, January 14, 2020. https://www.ivir.nl/publicaties/download/Dispute_Resolution_Content_Moderation_Final.pdf.

Tworek, Heidi and Leerssen, Paddy. 2019. "An Analysis of Germany's NetzDG Law." *Transatlantic Working Group on Content Moderation Online and Freedom of Expression*, April 15, 2019. https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

United Nations General Assembly. 1966. "International Covenant on Civil and Political Rights." Treaty Series 999 (December): 171.

United Nations General Assembly. 2007. Declaration on the Rights of Indigenous People, A/RES/61/295 (13 September 2007).

United Nations General Assembly. 2008. *Protect, respect and remedy: a framework for business and human rights: Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie.* A/HRC/8/5 (7 April 2008).

United Nations General Assembly. 2018. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. A/73/348 (29 August 2018).

United Nations General Assembly. 2019. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. A/74/486 (9 October 2019).

United Nations Human Rights Committee. 2004. General comment No. 31 [80]: The nature of the general legal obligation imposed on States Parties to the Covenant. CCPR/C/21/Rev.1/Add.13 (26 May 2004).

United Nations Human Rights Committee. 2011. General Comment No. 34, Article 19: Freedoms of Opinion and Expression.

United Nations Human Rights Committee. 2018. Report of the independent international fact-finding mission on Myanmar: Advance Edited Version. A/HRC/39/64 (12 September 2018).

United Nations Human Rights Committee. 2019. General comment no. 36, Article 6 (Right to Life). CCPR/C/GC/36 (3 September 2019).

United Nations Human Rights Council. 2013. *Annual report of the United Nations High Commissioner for Human Rights, Addendum, Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred.* A/HRC/22/17/Add.4 (11 January 2013).

United Nations Human Rights Council. 2018. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.* A/HRC/38/35 (6 April 2018).

United Nations Human Rights Council. 2018. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. A/HRC/38/35 (6 April 2018).

United Nations Human Rights Council. 2018. *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*. A/HRC/38/47 (18 June 18).

United Nations Office of the High Commissioner for Human Rights. 2021. "Status of Ratification Interactive Dashboard: International Covenant on Civil and Political Rights." As of September 27, 2021, https://indicators.ohchr.org/.

United Nations Secretary-General. 2019. "United Nations Strategy and Plan of Action on Hate Speech." As of September 27, 2021, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20a

nd%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

United Nations Treaty Collection. 2021. "International Covenant on Civil and Political Rights: Status as at 02-07-2021 03:15:43 EDT." As of September 27, 2021, https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-4&chapter=4&clang=_en.

United Nations. 1966. "International Convention on the Elimination of All Forms of Racial Discrimination." Treaty Series 660 (March): 195

United Nations. 1988. "Convention on the Elimination of All Forms of Discrimination against Women." Treaty Series 1249 (December): 13.

United Nations. 1989. "Convention on the Rights of the Child." Treaty Series 1577 (November): 3

United Nations. 2006. "Convention on the Rights of Persons with Disabilities." Treaty Series 2515 (December): 3.

Yudhanjaya Wijeratne. 2020. "Facebook, language and the difficulty of moderating hate speech." *Media@LSE*, July 23, 2020. https://blogs.lse.ac.uk/medialse/2020/07/23/facebook-language-and-the-difficulty-of-moderating-hate-speech/.

Wikipedia Contributors. 2021. "Consensus (computer science)". As of September 27, 2021, https://en.wikipedia.org/wiki/Consensus_(computer_science).

Wu, Tim. 2016. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Vintage Books.

Zuckerberg, Mark. 2019. "Opinion: The Internet needs new rules. Let's start in these four areas." *The Washington Post*, March 30, 2019. https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.