doi:10.1017/aju.2019.46

SYMPOSIUM ON NON-STATE ACTORS AND NEW TECHNOLOGIES IN ATROCITY PREVENTION

SUPPRESSING ATROCITY SPEECH ON SOCIAL MEDIA

*Emma Irving**

In its August 2018 report on violence against Rohingya and other minorities in Myanmar, the Fact Finding Mission of the Office of the High Commissioner for Human Rights noted that "the role of social media [was] significant" in fueling the atrocities.[1] Over the course of more than four hundred pages, the report documented how Facebook was used to spread misinformation, hate speech, and incitement to violence in the lead-up to and during the violence in Myanmar.[2] Concluding that there were reasonable grounds to believe that genocide was perpetrated against the Rohingya, the report indicated that "the Mission has no doubt that the prevalence of hate speech," both offline and online, "contributed to increased tension and a climate in which individuals and groups may become more receptive to incitement."[3] The experience in Myanmar demonstrates the increasing role that social media plays in the commission of atrocities, prompting suggestions that social media companies should operate according to a human rights framework.

As social media becomes ever more accessible and the number of global users continues to grow, the potential for these platforms to be used to fuel existing tensions and cause instability increases. A hateful or inciting message that might previously have had limited exposure can now circulate with great speed and reach a very wide audience. Furthermore, the ability to spread hate speech and incitement is no longer limited to those with a platform on radio or TV, but is now open to anyone with a smart phone.[4]

As the dangerous side of social media becomes more apparent, states and scholars are asking how this phenomenon can and should be addressed. In particular, attention is turning towards the social media companies themselves, and to what actions they must or should take to supress hate speech and incitement posted on their platforms. Thus far, social media companies have been left to regulate themselves. This approach has come under increasing criticism in light of a series of scandals over data breaches,[5] election interference,[6] and hate speech

* *Assistant Professor of Public International Law, Leiden Law School.*

[1] Human Rights Council, Report of the Independent International Fact-Finding Mission on Myanmar, UN Doc. A/HRC/39/64, para. 74 (Aug. 24, 2018) [hereinafter FFM Report, Abbreviated Version].

[2] Human Rights Council, Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar, UN Doc. A/HRC/39/CRP.2 (Sept. 17, 2018) [hereinafter FFM Report, Detailed Version].

[3] *Id.* at para. 1354.

[4] For more on these challenges, *see* Richard Wilson & Matthew Gillet, The Hartford Guidelines on Speech Crimes in International Criminal Law 123–29 (2018).

[5] *Facebook Fined £500,000 for Cambridge Analytica Scandal*, BBC News (Oct. 25, 2018); James Titcomb et al., *Facebook Security Breach Exposed 50 million Accounts to Attackers*, Telegraph (Sept. 28, 2018).

[6] *See, e.g.*, Sarah Frier, *Facebook Discovers an Ongoing Effort to Influence the 2018 Midterm Elections*, Time (July 31, 2018).

and incitement.[7] In response to this criticism, some social media companies—Facebook prominent among them—increasingly acknowledge their role in these events, noting that until now they did not "take a broad enough view of [their] responsibility."[8]

This essay explores whether social media companies are legally obliged to supress hate speech and incitement posted on their platforms under international, regional, or domestic laws. After concluding that there is a significant gap in legal obligations at each level, the essay considers the opportunities presented by self-regulation carried out in a human rights framework.

Responsibilities at the International, Regional, and Domestic Levels

At the international level, only international human rights law contains rules relating to hate speech and incitement that are relevant to the present discussion. Article 20(2) of the International Covenant on Civil and Political Rights stipulates that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." This obligation is directed at states—requiring them to prohibit such conduct under domestic law—rather than at social media companies, which, as domestic corporate actors, are not directly bound by human rights treaties. However, while companies may lack international legal personality, recent years have seen efforts to subject companies to human rights law standards. The UN Guiding Principles on Business and Human Rights[9] detail a number of responsibilities that companies have to respect human rights in their activities. These include the responsibility to avoid contributing to adverse human rights impacts[10] and the responsibility to conduct due diligence to identify the potential human rights impacts of company activities.[11]

Efforts to hold companies to human rights standards have shifted the normative environment in which companies operate. However, the choice of the term "responsibilities" throughout the UN Guiding Principles was deliberate: it denotes that human rights are a standard of expected conduct for companies, not a set of legal obligations.[12] Attempts to draft a treaty that would impose obligations directly on private companies remain at a standstill.[13] As such, in the search for a legal obligation on social media companies to supress hate speech and incitement on their platforms, international human rights law only takes us so far.

While international human rights law does not bind companies directly, international criminal law was specifically designed to bind non-state actors. Certain regional developments are relevant here: in 2014, the African Union adopted the Malabo Protocol,[14] which will create an international criminal law section within the (yet to

[7] *See, e.g.*, Max Fisher & Amanda Taub, *How Everyday Social Media Users Become Real-World Extremists*, N.Y. Times (Apr. 25, 2018).

[8] Craig Timberg & Tony Romm, *Facebook CEO Mark Zuckerberg to Capitol Hill: "It Was My Mistake, and I'm Sorry"*, Wash. Post (Apr. 9, 2018).

[9] UN Office of the High Commissioner for Human Rights, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, UN Doc. HR/PUB/11/04 (Mar. 21, 2011) [hereinafter UN Guiding Principles].

[10] *Id.* Principle 13.

[11] *Id.* Principle 17.

[12] Human Rights Council, Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, UN Doc. A/HRC/14/27, para. 55 (Apr. 9, 2010) [hereinafter Report of the Special Rapporteur].

[13] For information on this initiative, see the work of the Open-ended intergovernmental working group on transnational corporations and other business enterprises with respect to human rights.

[14] Protocol on Amendments to the Protocol on the Statute of the African Court of Justice and Human Rights, June 27, 2014 [hereinafter Malabo Protocol].

be established) African Court of Justice and Human Rights.[15] This Court will have jurisdiction to prosecute corporations for a number of international crimes, including crimes against humanity and genocide, each of which might encompass hate speech and incitement under certain circumstances.[16] This development departs from the approach taken in the statutes of the International Criminal Court and the International Criminal Tribunals for the Former Yugoslavia and for Rwanda, whereby jurisdiction was limited to natural persons.[17] Instead, it continues the nascent corporate responsibility that began in the post-World War II Nuremberg jurisprudence.[18]

Under the Malabo Protocol, intention on the part of a company to commit a crime can be established "by proof that it was the policy of the corporation to do the act which constituted the offence."[19] Assuming that the Malabo Protocol eventually enters into force,[20] and assuming that jurisdiction can be established in a given case,[21] prosecuting social media companies for unlawful content posted on their platforms is a theoretical possibility. However, given that third parties are the ones posting the unlawful content, rather than social media companies themselves, only in very unusual circumstances would it be possible to show that the company policy was to "do the act which constituted the offence." Simply failing to remove problematic content, even with knowledge that the content is unlawful, is unlikely to qualify. As such, the Malabo Protocol does not currently offer an effective avenue for compelling social media companies to supress hate speech and incitement on their platforms.

There have also been soft law developments at the regional level that are specifically targeted at the growing dangers posed by social media. In 2016, the European Commission presented Twitter, Facebook, YouTube, and Microsoft with a Code of Conduct on Countering Illegal Hate Speech Online.[22] The commitments in this code center around procedures for reporting and removing hate speech from social media platforms. These include a commitment to establish clear and effective processes to review reported content and an undertaking to remove illegal hate speech within twenty-four hours of it being reported. Hate speech is understood as including incitement to violence towards particular groups.

In the 2016 press release that introduced this code, the Commission stressed the importance of member states complying with their EU law obligation to criminalize hate speech within their domestic legal systems. However, with respect to social media companies, the Code of Conduct is careful to use the word "commitments" rather than "obligations." The code is meant to create a normative environment of compliance, in the same way that the UN Guidelines on Business and Human Rights does, but is not intended to create legally binding obligations for

[15] Protocol on the Statute of the African Court of Justice and Human Rights, July 1, 2008.

[16] Jurisdiction over corporations is provided for in Article 46C of the Malabo Protocol; for further information on incitement under international criminal law, *see* Richard Wilson, Incitement on Trial: Prosecuting International Speech Crimes (2017).

[17] Rome Statute of the International Criminal Court art. 25(1), July 17, 1998; Statute of the International Criminal Tribunal for the Former Yugoslavia art. 6, May 25, 1993, SC Res. 808/1993; Statute of the International Criminal Tribunal for Rwanda art. 5, Nov. 8, 1994, SC Res. 955/1994.

[18] Kai Ambos, *International Economic Criminal Law*, 29 Crim. L.F. 499, 506–10 (2018).

[19] Malabo Protocol, *supra* note 14, art. 46C.

[20] The Malabo Protocol requires fifteen ratifications to enter into force. As of July 2019, fifteen states have signed it but none has ratified it.

[21] *See* Malabo Protocol, *supra* note 14, arts. 46E, 46E *bis*, and 46F.

[22] European Commission Press Release IP/16/1937, European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech (May 31, 2016). More companies signed up in 2018. *See* European Commission, Countering Illegal Hate Speech Online #NoPlace4Hate (Oct. 18, 2018).

social media companies themselves. Subsequent developments at the EU level, while more specific, detailed, and forceful in their language, essentially continue this voluntary approach.[23]

Germany has transposed the approach of the European Commission Code of Conduct into binding domestic law. The 2017 Network Enforcement Act sets out requirements for the way in which large social media companies in Germany must handle reports of unlawful content. Unlawful content includes, but is not limited to, hate speech and incitement to violence.[24] As with the Code of Conduct, the Network Enforcement Act creates a requirement to establish clear and effective processes to review unlawful content, and an obligation to remove content that is manifestly unlawful within twenty-four hours of receiving a complaint. (For other unlawful content, the deadline is seven days.)[25] Failure to have proper procedures in place can result in financial penalties.[26] Kenya and Honduras have also taken steps to impose legal requirements on social media companies to remove hate speech and incitement to violence on their platforms.[27]

As the above overview shows, initiatives at the domestic level have gone furthest in imposing legal obligations on social media companies, but these remain the exception rather than the norm. The general position at the international, regional, and domestic levels is to rely on voluntary commitments and self-regulation for the moderation of online content. As the following section discusses, this is not necessarily a problem. Given the challenges associated with state-by-state regulation, there are advantages to having companies regulate themselves. How these approaches can be adapted to acute situations, however, is an open question.

The Way Forward?

While the weaponization of social media during conflict is not new, the fact that, in Myanmar, "Facebook *is* the internet"[28] meant that the use of the platform to disseminate hate speech and incitement to violence was particularly prominent. Despite this, the Fact Finding Mission (FFM) did not propose that states impose legal obligations on social media companies, recommending instead that social media companies *voluntarily* adopt human rights law as the framework for moderating content.[29] In other words, the preferred approach is to encourage social media companies to self-regulate in removing problematic content in a way that is consistent with human rights. The UN Special Rapporteur on Freedom of Opinion and Expression[30] and civil society also support this self-regulation approach.[31]

[23] European Commission, Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms, COM(2017) 555 final (Sept. 28, 2017). With respect to terrorist online content specifically, which may constitute hate speech and incitement but which is a much broader category of content, the European Union has initiated steps to impose binding measures on service providers to remove content. Council of the European Union Press Release, Terrorist Content Online: Council Adopts Negotiating Position on New Rules to Prevent Dissemination (Dec. 6, 2018).

[24] Section 1(3) of the Network Enforcement Act (Netzdurchsetzungsgesetz [NetzDG]), defines unlawful content in relation to provisions of the German Criminal Code (Strafgesetzbuch [StGB]), including §§ 91a, 111, and 130, which relate to hate speech and incitement to violence.

[25] Network Enforcement Act, *supra* note 24, § 3.

[26] *Id.* § 4.

[27] High Commissioner for Human Rights, 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, Communication OL KEN 10/2017 (July 26, 2017); Javier Pallero, *Honduras: New Bill Threatens to Curb Online Speech*, Access Now (Feb. 12, 2018).

[28] FFM Report, Abbreviated Version, *supra* note 1, para. 74.

[29] FFM Report, Detailed Version, *supra* note 2, para. 1718.

[30] Report of the Special Rapporteur, *supra* note 12.

[31] Article 19, Self-Regulation and "Hate Speech" on Social Media Platforms (2018) [hereinafter Article 19 report]; Global Civil Society Initiative, Manila Principles on Intermediary Liability (May 30, 2015).

This reluctance to recommend state-imposed legal regulation is likely attributable to the fact that moving away from a self-regulation model has some significant disadvantages. First, increased legal pressure on social media companies to suppress certain types of content can lead to the overremoval of content. As social media companies are guided by economic considerations, there is a significant risk that they will prioritize avoiding liability over the protection of free speech, and so remove more content than is warranted.[32] After Germany's Network Enforcement Act came into force, allegations circulated that social media companies were removing legitimate speech.[33] Second, domestic legal regulation would not necessarily mean that human rights would be better protected. On the contrary, where legal regulation does exist in domestic contexts, it often contains vaguely formulated rules and compels social media companies to suppress political dissent online under the guise of suppressing hate speech and incitement.[34]

If self-regulation is preferred to state-imposed regulation, how should the human rights framework shape this self-regulation? The UN Special Rapporteur on Freedom of Opinion and Expression stressed some core human rights concepts that should guide social media companies in moderating content, including legality, necessity and proportionality, nondiscrimination, transparency, and access to a remedy. One particularly innovative idea is the proposal to create social media councils.[35] These councils, composed of both social media company representatives and other relevant stakeholders, would operate as independent self-regulatory bodies for the industry. They could elaborate ethical guidelines and content moderation policies, and could act as a focal point for individual complaints, thereby providing access to a remedy for users and accountability for the companies. To gain public trust, these councils would have to work in an open and consultative manner.

A self-regulation approach to content moderation that is informed by a human rights framework, such as a system of social media councils, has much to recommend it. In promoting the independence of social media companies, it reduces the opportunities for states to use these companies to silence opposition; in focusing on transparency and access to a remedy, it may reduce the likelihood that social media companies act overly broadly in removing content, because it would require them to provide justifications for their decisions.

However, there is still work to be done to understand how these broad human rights concepts should guide social media companies in acute situations, such as that in Myanmar. How social media companies approach hate speech and incitement on their platforms in the context of stable, secure societies may differ from how they should approach such content in unstable, insecure countries where violence often lurks just below the surface. The FFM Report recommends that "all death threats and threats of harm in Myanmar should be treated as serious and immediately removed when detected."[36] Such an all-or-nothing approach may be appropriate in the Myanmar context, but less appropriate in more stable contexts. In the latter, the companies or the councils should take time to understand the background of the speech, such as whether it was made in jest, in order to ascertain whether removal is justified. While ultimately the outcome may be the same, in that the speech is removed, the process for arriving at the decision may differ. A further recommendation of the FFM is that all social media platforms active in Myanmar establish "early warning systems for emergency escalation."[37] Again, this may be desirable for unstable countries, but these systems require a degree of data collection that may not accord with human rights principles if done in other contexts.

[32] Report of the Special Rapporteur, *supra* note 12, para. 17. On the problem of overbroad removals, *see* Avi Shapiro, *YouTube and Facebook Are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals*, The Intercept (Nov. 2, 2017).

[33] Bernhard Rohleder, *Germany Set out to Delete Hate Speech Online. Instead, It Made Things Worse*, Wash. Post (Feb. 20, 2018).

[34] Report of the Special Rapporteur, *supra* note 12, paras. 15–24.

[35] Described in detail in Article 19 report, *supra* note 31.

[36] FFM Report, Detailed Version, *supra* note 2, para. 1724.

[37] *Id.*

In the absence of an appropriate international law framework, and given the disadvantages of state-imposed legal regulation, self-regulation informed by human rights norms would seem to be the best option on the table. Indeed, the human rights framework is likely flexible and robust enough to accommodate different approaches in different communities. What is needed now is further guidance on how to strike the balance between supressing hate speech and incitement, and the need to protect human rights, particularly in acute situations. Perhaps here, the role of social media councils may be particularly useful.