Platform Governance at the Periphery: Moderation, Shutdowns and Intervention

Giovanni De Gregorio, Nicole Stremlau

Abstract: After illustrating how the spread of dangerous content has led to troubling consequences beyond digital boundaries, this chapter describes how online hate speech has become criminalised in the global south. It analyses Internet shutdowns to understand their socio-legal consequences, and explores the applicability of public international law and the humanitarian doctrine to information interventions.

Keywords: platform governance; global south; Africa; hate speech; internet shutdowns; information intervention; content moderation; disinformation; media; online speech

1. Introduction

The spread of online hate and disinformation is increasingly provoking dramatic and troubling consequences beyond digital boundaries. False information about health treatments during the Covid-19 pandemic,[1] or the use of social media in mobilizing actors for the attack on Capitol Hill,[2] are some prominent examples of how online speech can affect the general public. But offline harms are far broader and often less explicitly tied to online speech. Our focus here is on areas of the world that have not been considered 'priorities' by social media companies.[3] For example, in

1 Julie Posetti and Kalina Bontcheva, *Disinfodemic: deciphering COVID-19 disinformation. Policy brief 1*. (2020), https://en.unesco.org/covid19/disinfodemic.

2 Joan Donovan, Brian Friedberg and Emily Dreyfuss, "The Capitol siege was the biggest media spectacle of the Trump era," *The Guardian*, January 11 (2021) https://www.theguardian.com/commentisfree/2021/jan/11/the-capitol-siege-was-the-biggest-media-spectacle-of-the-trump-era.

3 In April 2021 the Guardian published an except from an email by a top Facebook executive explaining that the company should address concerns of abuse online by focusing on "top countries, top priority areas… and try to somewhat work our way down [to peripheral countries, or those that are seen as less strategic and driving

the Central African Republic, online hate speech has contributed to mass atrocities between Christians and Muslims,[4] and in Sri Lanka, rumours on social media have led to a number of religious attacks, including the 2019 Easter Sunday church and hotel bombings,[5] while the use of Facebook in inciting violence against Myanmar's minority Muslim population has elevated concerns about the role of online platforms in perpetrating genocides.[6]

The *fil rouge* connecting these examples is the role of social media companies[7] in governing speech on a global scale.[8] Online platforms that process content rely on a mix of human moderators and artificial intelligence systems that define which content must be removed according to non-transparent standards and without explanation, providing very few opportunities for remedies.[9] As the global pandemic has altered working arrangements for human coders (along with many office workers) it has also made the implementation of artificial intelligence systems in content moderation more urgent for companies.[10] But this has brought to the fore

---

news]". *See:* https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation.

4 Office of the High Commissioner for Human Rights, *Preventing incitement to hatred and violence in the Central African Republic* (2019) https://www.ohchr.org/EN/NewsEvents/Pages/PeacekeepersDay2019.aspx.

5 Newley Purnell, "Sri Lankan Islamist Called for Violence on Facebook Before Easter Attacks," *Wall Street Journal*, April 30, 2019 https://www.wsj.com/articles/sri-lankan-islamist-called-for-violence-on-facebook-before-easter-attacks-11556650954.

6 Fanny Potkin and Poppy McPherson, "Spreading like Wildfire: Facebook Fights Hate Speech before Myanmar Poll," *Reuters*, November 5, 2020, https://www.reuters.com/article/myanmar-election-facebook-idUSL4N2HQ3QU.

7 When referring to 'social media' we are primarily speaking of user-generated content on large platforms such as Facebook, Twitter, YouTube or TikTok.

8 Hannah Bloch-Wehba, "Global platform governance: private power in the shadow of the state," *SMU L. Rev.* 72 (2019): 27; Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (Yale University Press, 2018), Kate Klonick, "The new governors: The people, rules, and processes governing online speech." *Harv. L. Rev.* 131 (2017): 1598; Luca Belli, David Erdos, Maryant Fernandez Perez, Pedro Augusto P. Francisco, Krzysztof Garstka, Judith Herzog, Krisztina Huszti-Orban et al., *Platform regulations: how platforms are regulated and how they regulate us* (Leeds, 2017).

9 Sarah T Roberts, *Behind the screen* (Yale University Press, 2019).

10 Sana Ahmad, "COVID-19 and the Future of Content Moderation," Coronavirus and its Societal Impact-Highlights from WZB Research, 2020. https://www.wzb.eu/en/research/corona-und-die-folgen/covid-19-and-the-future-of-content-moderation.

just how problematic AI can be when it comes to effectiveness; during the pandemic there have been significant cases when the involvement of human moderators was restricted and an over-reliance on the automated system led to the spread of disinformation and blocking of accounts that were actually countering disinformation.[11]

Against this opaque framework of governance and fragmented responses by social media companies, a variety of actors, from non-governmental organizations to various public authorities around the world have tried to tackle the harm produced by the spread of hate and disinformation online.[12] Governments have reacted in different ways, particularly in poorer and less geopolitically influential countries. They have accused online platforms of disseminating hate and disinformation online, criminalised the spread of hate and disinformation,[13] have used platforms for surveillance, worked to push alternative narratives (sometimes flooding platforms with disinformation), and have attempted to censor content.[14] The spread of hate on social media has also been one of the primary reasons why governments have increasingly justified the use of Internet shutdowns, which can involve a range of tools from slowing down the internet (making it practically unusable) to completely switching it off.[15] Whereas only a few years ago such forms of censorship would have been seen as a grave violation of freedom of expression, increasingly they being seen to be one of the few mechanisms available for addressing online speech and offline harms in a moment of crisis.

The role of media in contributing to disseminate hate and violence is not new.[16] In some cases of violence and mass atrocities, international actors, including the United Nations (UN), have relied on "information

11 Statt Nick, "How Facebook Is Using AI to Combat COVID Misinformation and Detect 'Hateful Memes,'" The Verge, May 12, 2020, https://www.theverge.com/2020/5/12/21254960/facebook-ai-moderation-covid-19-coronavirus-hateful-memes-hate-speech.

12 Roxana Radu, Fighting the 'Infodemic': Legal Responses to COVID-19 Disinformation. *Social Media+ Society*, *6*(3), 2020.

13 Dickens Olewe, "Kenya, Uganda and Tanzania in "anti-fake news campaign," *BBC News*, 16 May 2018.

14 Adrian Shahbaz, "The Rise of Digital Authoritarianism: Freedom on the Net 2018," *Freedom House*, October (2018) https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism.

15 De Gregorio, Stremlau, " Internet Shutdowns and the Limits of Law".

16 Robert Edwin Herzstein. *The war that Hitler won: The most infamous propaganda campaign in history* (Putnam Publishing Group, 1978). Nicole Stremlau. *Media, Conflict and the State in Africa* (Cambridge University Press, 2018).

interventions," an expression developed in the 1990s in response to the conflict in Rwanda and the Balkans.[17] While information interventions have been applied to traditional media outlets, we ask whether such a response could be relevant for social media, particularly when online platforms have a leading role in disseminating content directly associated with mass atrocities.

Within this framework, this chapter explores the challenges raised by online hate and disinformation in areas of the world that are less of a business priority for large social media companies. By focusing on content moderation as an expression of platform governance, we underline how the spread of online hate and disinformation have led to troubling consequences beyond digital boundaries. In the first part, we focus on the criminalisation of online hate and disinformation as a response to the consequences this content produces in the online and offline world. The second part explores how the spread of online hate speech and disinformation has provided governments with further justifications, that are increasingly becoming internationally acceptable (or at least understood), to censor the Internet for protecting national security or other public interests. The third part focuses on the role of international actors in addressing the spread of online hate and disinformation by looking at the applicability of the doctrine of information intervention to social media.

Our focus in this chapter is on the variety of legal and censorship responses to online hate. We recognize that there are considerable efforts on the part of governments to address online speech with different techniques ranging from attempting to shift narratives through flooding social media with specific content (as seen with the role of Cambridge Analytica), or using surveillance and both online and offline coercion or harassment to silence certain voices. In this chapter, however, our emphasis is on the intersection of concerns around content moderation and the use of law or force to address these concerns.

17 Monroe E Price and Mark Thompson, eds. *Forging peace: intervention, human rights, and the management of media space* (Indiana University Press, 2002); Jamie F. Metzl, "Information intervention: When switching channels isn't enough," *Foreign Affairs* (1997): 15-20.

2. An initial response: Criminalising online hate and disinformation

Social media companies have a critical role in determining the standards of protection of online speech on a global scale. Although these companies do not always have offices in the country where hate and mass atrocities are perpetrated, they exercise broad discretion in determining the rules according to what information circulates online and, therefore, how content is shared between communities.[18] And this does not change even in situations of conflicts or violence where these actors have determine how to moderate hate and disinformation according to their ethical, business and legal framework.

This process, with its strengths and limitations, was evident during the Arab Spring.[19] As observed by Zeitoff,[20] communications in conflicts have typically been defined in two ways: "elite-level communication" focused on tactical and logistical aims; and "mass-based appeals" aimed at coordinating or inhibiting public behaviour through control of the narrative and manipulating mass channels of communication.[21] Social media provide a new paradigm, transforming users into active creators of content whose standard of protection is defined by private companies. This increasing degree of protection of online speech can empower users in authoritarian regimes while affecting social tensions and conflicts.[22] The disintermediation of traditional media outlets allows individuals to challenge elite-dominated discourse, especially in authoritarian regimes, which tend to exercise public control over traditional media outlets. Information spread on social media can be immediately shared with other communities of users, potentially going viral. The digital spaces provided by social media have encouraged access to diverse information online, promoting a plurality of voices and sharing of opinions. In particular, the possibility to use these channels to contest central authorities and spread disinformation has

18 Dimitra Dimitrakopoulou, Georgios Tzogopoulos and Alexandra Nikolakopoulou, *The Role of Social Media in Violent Conflict* (INFOCORE Working Paper 2014/05).

19 Philip N. Howard and Muzammil M. Hussain, *Democracy's Fourth Wave?: Digital Media and the Arab Spring* (OUP 2013).

20 Thomas Zeitzoff, "How social media is changing conflict," *Journal of Conflict Resolution* 61, no. 9 (2017): 1970-1991.

21 Philip N. Howard. *The digital origins of dictatorship and democracy: Information technology and political Islam* (Oxford University Press, 2010).

22 Peter Dahlgren, "The Internet, public spheres, and political communication: Dispersion and deliberation," *Political communication* 22, no. 2 (2005): 147-162.

encouraged governments to censor online speech or even use social media as instrument of surveillance.[23]

The use of automated technologies for moderating content also produces effects that extend beyond domestic boundaries.[24] These channels of communication allow information to be disseminated more widely and with greater speed, especially in cases involving strong messages of hate or dissent. Algorithmic content moderation contributes to driving people to online hate and disinformation,[25] which can also lead to discrimination.[26] As underlined by Tufekci, "YouTube may be one of the most powerful radicalizing instruments of the 21st century".[27] In areas characterised by tensions and conflicts, this can inflame and escalate violence and conflicts - the lack of language training in certain languages makes content moderation less effective in detecting online hate speech. The Myanmar genocide has underlined the inability of Facebook to detect and limit the spread of hate speech.[28] The spread of hate speech on Facebook supported ethnic cleansing in Myanmar, but this went mostly unchecked due to the lack of moderation tools and human moderators fluent in Burmese. While Facebook significantly expanded its team of Burmese speakers to create a data set of hate and violent expressions, the international pressure to act also led to overreactions including the ban of some armed groups.[29]

Given these challenges, the first reaction by many governments has been to criminalise the spread of online hate and disinformation by users and social media. In May 2019, Singapore adopted the Protection from

23 Evgeny Morozov, *"The net delusion: The dark side of Internet freedom," PublicAffairs*, 2012.

24 Jack M. Balkin, "Free speech in the algorithmic society: Big data, private governance, and new school speech regulation," *UCDL Rev.* 51 (2017): 1149

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

26 Safiya Umoja Noble, *Algorithms of oppression: How search engines reinforce racism* (NYU Press, 2018).

27 Zeynep Tufecki, "YouTube. The Great Radicalizer" *New York Times*, May 10, 2018 https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

28 Steve Stecklow, "Why Facebook is losing the war on hate speech in Myanmar" *Reuters*, Aug. 15 2018 https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

29 Jeffrey Sablosky, "Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar" *Media, Culture & Society* (2021).

Online Falsehoods and Manipulation Act.[30] The scope of this legislation covers content that is false or misleading, whether wholly or in part and/or there are reasons to believe it affects public interest. The prohibition of communication of "false statements of fact" in Singapore applies to both individuals and online intermediaries applying a fine from S$ 20,000 (12,000 euro) up to S$ 100,000 (62,000 euro) and/or imprisonment from 1 to 10 years, whereas for intermediaries they generally range between S$ 500,000 (310,000 euro) to S$ 1 million (622,000 euro).

Malaysia similarly followed the path towards the criminalization of online disinformation even if the government decided to repeal the legislation after its adoption.[31] Nonetheless, in March 2021, the Perikatan National Government enacted an emergency ordinance using powers conferred by a January 2021 Emergency Proclamation to face the spread of online disinformation about Covid-19 or the proclamation of the emergency.[32] This measure introduces new criminal offences relating to the creation, publication, or dissemination of so-called 'fake news' and the failure to take down publications containing content deemed as 'fake news'. This conduct is sanctioned with up to three years imprisonment. Furthermore, individuals and internet platforms which do not remove content within 24 hours based on an order coming from public officials, not necessarily courts, can be sanctioned with a fine of up to 100,000 Malaysian ringgit (20,000 euro) and, in the case of a continuing offense, up to 300,000 ringgit (60,000 euro) for every day in which the content is available.

Moving from Asia to Africa, Ethiopia passed a law sanctioning the spread of online hate by Internet users and platforms providing up to three years of imprisonment and a fine of up to 100,000 birrs (2,900 euro).[33] In justifying this legislation, reference was made to the central role of hate speech and electoral related violence in neighboring Kenya as well as the introduction of the Network Enforcement Act (NetzDG) in Germany which is regarded as an ambitious legislation requiring platforms to re-

30 Protection from Online Falsehoods and Manipulation Bill (2019), https://sso.agc.gov.sg/Bills-Supp/10-2019/Published/20190401?DocDate=20190401.

31 Anti-fake News Act 2018, https://perma.cc/Y5H3-D6G8.

32 Malaysia's king declares state of emergency to curb spread of COVID-19, ABC News, https://www.abc.net.au/news/2021-01-12/malaysia-king-declares-state-of-emergency-to-curb-covid-spread/13051642.

33 Proclamation No. 1185 /2020 Hate Speech and Disinformation Prevention and Suppression Proclamation, https://chilot.me/wp-content/uploads/2020/04/HATE-SPEECH-AND-DISINFORMATION-PREVENTION-AND-SUPPRESSION-PROCLAMATION.pdf.

move hate speech within 24 hours or face fines up to 50 million euros.[34] In the case of Nigeria, the fight against online hate speech has been even more radical. In 2019, two bills were proposed to increase government powers to shut down the internet, punish government critics and sanction hate speech with capital punishment.[35]

Among other approaches, the political choice of Uganda concerning how to restrict free speech online has been different. Since July 2018, a new Ugandan tax charges citizens 5 US cents a day for the use of 60 mobile apps, including Facebook, Twitter, Skype and WhatsApp. This social-media tax was passed as part of a bill that also includes taxes on mobile transactions and was seen as a way of attempting to reduce the use of these platforms. Many Ugandans, however, chose to access them from other internet connections (rather than mobile data) or use VPNs to get around the restrictions in place by the mobile operators.

These measures are just a small part of the array of new laws attempting to shape speech on social media but not necessarily limiting access to the internet in its entirety. The next section explores a blunter and far reaching tool of public censorship as a second reaction to the spread of online hate and disinformation. As the next section underlines, governments are increasingly relying on Internet shutdowns, thus, leading to a process of normalisation of these practices as a reaction to the spread of online hate and disinformation on social media.

3. Internet shutdowns and the control of narratives

The spread of online hate and disinformation on social media is increasingly considered by some governments to be a justification (or legitimate aim) to censor speech and shut down the Internet. This is often perceived as the only immediately effective remedy to deal with the escalation of violence in the context of company-led discretion in responding and moderating content. Even though there is very limited evidence about the effects of these practices in tackling the misinformation and hate they purport to address, shutdowns have been implemented to curtail online

34 Network Enforcement Act, 2018, https://germanlawarchive.iuscomp.org/?p=1245.
35 Nigeria bill aims at punishing hate speech with death, https://www.dw.com/en/nigeria-bill-aims-at-punishing-hate-speech-with-death/a-51419750.

speech, and particularly content that is seen to be provoking violence or promoting dissent.[36]

Internet shutdowns have increased in scale and scope over several years[37], particularly in Asia and Africa.[38] From India, where there have been many localized Internet shutdowns,[39] to Cameroon, a country that brazenly blocked access in half of the country for more than 230 days between 2017 and 2018,[40] shutting down the Internet (either partially or entirely) appears to be used by governments when they want to act quickly, particularly to quell perceived or potential civil unrest, and might have limited capacity for other mechanisms of online control. The rise of internet shutdowns also reflects a frustration on the part of some governments with their inability to intervene in the governance of the digital platforms that are often controlled by businesses in another continent. In the absence of concerted cooperation with companies, shutting down the entire network or specific digital spaces has become increasingly popular. While the ire and frustration coming from countries such as New Zealand, Germany, or France toward Facebook or Twitter's inability to control disinformation and hate speech has been significant, they have also found more engagement at company headquarters. This may be because poorer countries and those that typically resort to Internet shutdowns have far less leverage over the large American companies. It is helpful to keep in mind that the GDP of a country like Burundi is approximately 3 billion USD while the value of Facebook is roughly 240 times that at 720 billion

36 Statista. "Government Justifications for Internet Shutdowns Worldwide 2019." Accessed March 25, 2021, https://www.statista.com/statistics/1096316/government-justifications-for-internet-shutdowns/#:~:text=Official%20government%20justifications%20for%20internet%20shutdowns%20worldwide%202019&text=Fake%20news%20and%20hate%20speech.

37 In 2020 it was estimated that there were at least 155 shutdowns in 29 countries, down from 213 incidents in 2019 (https://www.accessnow.org/keepiton/).

38 For an overview of trends on internet shutdowns in Africa see: Eleanor Marchant and Nicole Stremlau, "The Changing Landscape of Internet Shutdowns in Africa", International Journal of Communication, 14(2020), 4216-4223 and Eleanor Marchant and Nicole Stremlau, "A Spectrum of Shutdowns: Reframing Internet Shutdowns from Africa" International Journal of Communications 14(2020): 4327-4342.

39 Megha Bahree, "India leads the world in the number of Internet shutdowns: Report", *Forbes*, November 12, 2018 https://www.forbes.com/sites/meghabahree/2018/11/12/india-leads-the-world-in-the-number-of-internet-shutdowns-report/.

40 Abdi Latif Dahir, "Africa Internet shutdowns grow longer in Cameroon, Chad, Ethiopia," Quartz Africa, November 19, 2018. https://qz.com/africa/1468491/africa-internet-shutdowns-grow-longer-in-cameroon-chad-ethiopia/.

USD. Given these severe inequalities it is not surprising that complaints from countries in Africa have scarce reception in Silicon Valley. In fragile states, the lack of negotiating powers of governments in respect of social media underline the power that these actors can exercise, thus, making shutdowns an apparent necessity to censor online speech.

When Internet shutdowns occur, they are usually met with condemnation by free speech advocates and Internet freedom groups such as Access Now.[41] The effects of Internet shutdowns by virtue of the role of the digital environment in today's society cannot be neglected. Domestic deterrents, such as arguments around potential economic costs, appear to have little impact (particularly if governments are weighing up the comparative economic costs of protests or unrest), and advocacy groups that focus on publicly shaming governments have not reduced the use of shutdowns. The Internet is not only relevant from a technical or economic perspective,[42] but also for the exercise of democratic values such as assembly and freedom of expression and, therefore, as a crucial source of information and knowledge.

A polarized debate has emerged with governments grasping for ways to control flows of misinformation and hate speech, sometimes with legitimate concerns and frustration over their inability to control the vast amount of user generated content, the tepid engagement or responses social media companies to address this issue, and the forceful (and unbending) condemnation of Internet shutdowns by advocacy groups and the human rights community. This can make it difficult to have a nuanced conversation about when and under what circumstances shutdowns might be justified. While there is a lack of transparency and accountability of states when shutting down the Internet, including justification of the reasons or the procedures on which these restrictive measures are implemented, there have been some efforts to map the reasons governments have provided. The majority of explanations reference national security, including political mobilization or protest.[43] Election periods are another

41 Access Now, "The state of Internet shutdowns around the world the 2020", *#KEEPITON Report*, 2021 https://www.accessnow.org/cms/assets/uploads/2021/03/KeepItOn-report-on-the-2020-data_Mar-2021_3.pdf.

42 The Organisation for Economic Co-operation and Development, "OECD digital economy outlook 2017," https://www.oecd.org/sti/ieconomy/oecd-digital-economy-outlook-2017-9789264276284-en.htm.

43 Lynsey Chutel, "Zimbabwe's government shut down the Internet after fuel price protests turned deadly," *Quartz Africa*, January 15, 2019 https://qz.com/africa/1524405/zimbabwe-protest-internet-shut-down-military-deployed-5-dead/; Peter Micek

highly contested period.[44] In some cases, targeted shutdowns have been regionally specific whereby governments have tried to marginalize specific groups that may, for example, be attempting publicize human rights violations or may be protesting the absence of government service delivery in peripheral regions. And Internet shutdowns have also been implemented for more benign seeming issues, such as before school exams to prevent cheating.[45]

Unlike social media which are not bound to respect human rights according to international human rights law, states have an obligation to respect human rights according to covenants and customary international law that protects the right to freedom of expression limiting the shutting down of the digital environment. In January 2020, the Supreme Court of India recognised that freedom expression online enjoys constitutional protection,[46] even if this decision has not changed the general approach to Internet shutdowns in India. In January 2019, a Zimbabwean court ruled that government's internet shutdown as an answer to protests was illegal.[47] Similarly, in June 2020, the Economic Community of West African States (ECOWAS) Community Court decided that, by shutting down the Internet during the anti-government protests in 2017, the Togolese government violated human rights.[48] According to the court, the arguments based on

---

and Deji Olukotun, "Internet disrupted in Bahrain around protests as wrestling match sparks shutdown in India," *Access Now*, 24 June 2016 https://www.accessnow.org/internet-disrupted-bahrain-around-protests-wrestling-match-sparks-shutdown-india/; Philip N. Howard, Sheetal D. Agarwal, and Muzammil M. Hussain, "When do states disconnect their digital networks? Regime responses to the political uses of social media," *The Communication Review* 14, no. 3 (2011): 216-232.

44 Hilary Matfess, "More African countries are blocking internet access during elections," *Quartz Africa*, June 1, 2016 https://qz.com/africa/696552/more-african-countries-are-blocking-internet-access-during-elections/; Deji Olukotun, Peter Micek, and Gustav Bjorksten, "Vietnam blocks Facebook and cracks down on human rights activists during Obama visit," *Access Now*, 23 May 2016. https://www.accessnow.org/vietnam-blocks-facebook-human-rights-obama/.

45 Nour Youssef, "Algeria's answer to cheating on school exams: Turn off the Internet," *The New York Times*, June 21, 2018 Retrieved from https://www.nytimes.com/2018/06/21/world/africa/algeria-exams-cheating-internet.html.

46 Anuradha Bhasin vs Union of India & Ors. Writ Petition (Civil).

47 MacDonald Dzirutwe, Zimbabwe court says internet shutdown illegal as more civilians detained https://www.reuters.com/article/us-zimbabwe-politics/zimbabwe-court-says-internet-shutdown-during-protests-was-illegal-idUSKCN1PF11M.

48 Amnesty International et al. v. The Togolese Republic, 2020, https://www.accessnow.org/cms/assets/uploads/2020/07/ECOWAS_Togo_Judgement_2020.pdf.

national security could not justify the internet shutdown according to local or international law. This, however, does not mean that states cannot rely on legitimate interests to rely on shutdowns for example in cases of self-defence. Although there are different nuances of freedom of expression in regional human rights instruments and areas of the world, the Universal Declaration of Human Rights (UNDHR) and the International Covenant on Civil and Political Rights (ICCPR) are the primary structures to take into account for the three step-test based on legality, legitimacy and proportionality of the actions public authorities may take. Together, they can have a role in mitigating the rise of Internet shutdowns.

Despite the potential relevance of these legal procedures, the law has limitations when applied to Internet shutdowns. The scope of applicable regulation and legitimate interests could shape this framework which can be broadly exploited for political purposes. These concerns are particularly relevant in authoritarian regimes since the limits of the law in relation to Internet shutdowns are not only about the boundaries of the three-step test but also concern the scrutiny of these practices. The challenges posed by Internet shutdowns is also due to the lack of a common international enforcement mechanism that allows for both the transparent implementation of processes and procedures for when shutdowns might be justified, as well as the scrutiny of when shutdowns might be applied inappropriately.

4. Building consensus on interventions

This challenge of international coordination and the legitimacy, or illegitimacy, of shutdowns (even in cases when online speech is connected with extreme violence such as genocide) brings us to our third area of focus. Internet shutdowns cannot be a general remedy due to the violations of international human rights law, and even if these violations were not the case, shutdowns would still not be a preferred tool. The growing prominence of social media in spreading hate and inciting violence prompts questions about whether, and to what extent, international law and cooperation can offer new options. The role of media in disseminating hate and violence has been a longstanding aspect of violent conflict.[49] In the last thirty years, such (mis)use of media has exacerbated numerous wars

49 Cees Jan. Hamelink. *Media and conflict: Escalating evil* (Routledge, 2015); Thompson and Price, *Forging peace: intervention, human rights, and the management of media space*, (2002).

and violent conflicts, and in some cases even genocide like in Rwanda and Bosnia.[50] In the past, international actors, including the United Nations, have intervened in the media environment by implementing measures under the broad umbrella of "information intervention."[51] Information interventions are strategic efforts to interfere in (whether disrupting, manipulating or altering) a communications environment within a community, region or state afflicted by mass atrocities, in order to prevent or counter the dissemination of violence-inciting speech. The intervention can take place at various stages of a conflict and it can involve subsidizing or countering messages (through, for example, counter-narratives or providing support to certain media outlets- so-called 'peace media') or it may involve the direct closure of particular outlets such as the bombing of radio towers or the shuttering of a newspaper.

Information interventions are complex and political endeavours as much as legal ones. They must navigate international law, particularly the principle of non-intervention as expression of national sovereignty, the protection of human rights (i.e. freedom of expression). Such interventions, however, would get their legitimacy from humanitarian norms advancing the responsibility to protect (R2P),[52] and Chapter VII of the United Nations Charter with respect to the threats to the peace and acts of aggression. While the boundaries of the non-intervention principle raise the question of whether information interventions can be justified when seeking to prevent mass atrocities provoked by hate speech and disinformation, international law does not preclude the UN Security Council deciding what kind of speech or incitement satisfies the threshold required to trigger the Chapter VII mechanism. Therefore, while the spread of hate speech and disinformation may lead to conflicts and mass atrocities, the degree of danger may not be considered a threat to international peace and security.

Further challenges are political – gaining consensus on an information intervention is likely to be challenging. The responsibility to protect a regime does address whether and to what extent the international community should intervene in situations where state actors fail (voluntary or involuntary) to protect their population from mass atrocities or genocide.[53] In the absence of UN authorization, interventions cannot be legally based

50 Article 19, "Broadcasting genocide Censorship, propaganda & state-sponsored violence in Rwanda, 1990-1994," *Article* 19, London (United Kingdom, 1996).
51 Metzl, *Information intervention: When switching channels isn't enough*, 1997.
52 Alex J. Bellamy, *Responsibility to protect: A defense* (OUP Oxford, 2014).
53 Ibid.

on R2P and/or humanitarian reasons thus constituting the most relevant challenge to information intervention. This authorization constitutes the ultimate safeguard to avoid that compelling reasons (or excuses) are used to interfere with states' sovereignty. But in recent years, stemming from the challenging (and ultimately failed) intervention in Libya in 2011, the responsibility to protect has been criticized as being a cover for politically motivated interventions and advocates for invoking interventions based on the responsibility to protect have struggled to get traction within the UN Security Council.

Despite these challenges, information interventions have been applied to traditional media outlets, and in the current climate is important to consider its potential relevance to online communications, and social media in particular. As already underlined, Chapter VII of the UN Charter can be used to authorise international interventions in the media environment of a target state without violating the principle of non-intervention. In cases where social media are involved in the escalation of violent conflicts, particularly mass atrocities such as genocide (as we have seen in Myanmar), the UN Security Council could, in theory, authorise an intervention under Chapter VII due to a breach of international peace and security. In this situation, an independent international body (which we will refer to as an Information Intervention Council) could be involved in limiting access to social media as part of its response to addressing mass atrocities and, as a remedy of last resort, shutting down the Internet.

At first glance, UN authorization could provide a way to extend the doctrine of information intervention to social media promoting online hate and disinformation. Nonetheless, any information intervention measure must take into consideration the network architecture and modalities through which it is possible to limit dissemination of online hate and violence with specific regard to Internet shutdowns. In this case, the cooperation between the international community and social media is critical. For example, social media could remove content or block accounts based on the recommendation of the Information Intervention Council. This would help to foster a more positive framework of content moderation, with greater safeguards to avoid arbitrary internet shutdowns as well as greater care on the part of social media actors to avoid having their activities shut down by the intervention of the external Council.

However, moving towards information interventions risks collateral censorship, particularly in conflict-affected countries where citizens may have significant needs for accurate and plural information sources. Unlike traditional media outlets, which operate within a specific region and have an important role in providing information to those in that area, inter-

national social media platforms are driven by business incentives. As a consequence, social media companies may be motivated to cease operating in riskier regions where information interventions might be enacted which may lead to financial and reputational losses.

Information interventions are political as much as they are legal. There are, of course, risks with any intervention and particularly with one interfering with a information space. The line between information intervention and censorship can become blurred, with the real test being whether or not the measures address the responsibility to protect.

5. Conclusion

The governance of online speech is increasingly being shaped by a mix of public and private policies in an ad hoc and (often) arbitrary manner. Efforts by social media platforms have demonstrated the challenges of governing speech transnationally, particularly as their approach to moderating content is driven by business purposes rather than human rights norms. This leads to a clash between private interests focusing on profit and public values and the tension between protecting free speech while balancing conflicting rights and freedoms.

The offline harms associated of hate speech are a central justification as to why governments have proposed to criminalise online hate and disinformation, and have, at times, turned to blunter mechanisms, such as internet shutdowns, to regulate content online. Escalating concerns between online content and offline harms calls for urgent action, particularly by independent bodies such as the United Nations. The doctrine of information intervention offers one starting point to think about the potential role and responsibilities of international actors to intervene and address the most severe, or egregious cases, where online speech is leading to mass atrocities and human rights abuses such as genocide.

Bibliography:

Access Now, The state of Internet shutdowns around the world the 2020 #KEEPITON Report. (2021). https://www.accessnow.org/cms/assets/uploads/2021/03/KeepItOn-report-on-the-2020-data_Mar-2021_3.pdf.

Ahmad, Sana. *COVID-19 and the Future of Content Moderation*. Coronavirus and its Societal Impact- Highlights from WZB Research, 2020. Accessed March 25, 2021. https://www.wzb.eu/en/research/corona-und-die-folgen/covid-19-and-the-future-of-content-moderation.

Article 19, London (United Kingdom); *Broadcasting genocide Censorship, propaganda & state-sponsored violence in Rwanda, 1990-1994*. 1996.

Bahree, Megha. "India leads the world in the number of Internet shutdowns: Report." *Forbes*, November 12, 2018.

Balkin, Jack M. "Free speech and hostile environments." *Columbia Law Review* (1999): 2295-2320.

Balkin, Jack M. "Free speech in the algorithmic society: Big data, private governance, and new school speech regulation." *UCDL Rev.* 51 (2017): 1149.

Barendt, Eric. "Freedom of expression in the United Kingdom under the Human Rights Act 1998." *Ind. LJ* 84 (2009): 851.

Bellamy, Alex J. *Responsibility to protect: a defense*. OUP Oxford, 2014.

Belli, Luca, David Erdos, Maryant Fernandez Perez, Pedro Augusto P. Francisco, Krzysztof Garstka, Judith Herzog, Krisztina Huszti-Orban et al. *Platform regulations: how platforms are regulated and how they regulate us*. Leeds, 2017.

Bloch-Wehba, Hannah. "Global platform governance: private power in the shadow of the state." *SMU L. Rev.* 72 (2019): 27.

Bozdag, Cigdem. "Managing Diverse Online Networks in the Context of Polarization: Understanding How We Grow Apart on and through Social Media." *Social Media+ Society* 6, no. 4 (2020): 2056305120975713.

Cheng, Sage, Felicia Anthonio and Berhan Taye. #KeepItOn: Internet shutdowns put lives at risk during COVID-19. *Access Now*, 26 May 2020. https://www.accessnow.org/keepiton-internet-shutdowns-put-lives-at-risk-during-covid-19/.

Chutel, Lynsey. "Zimbabwe's government shut down the Internet after fuel price protests turned deadly." *Quartz Africa*, January 15, 2019.

Cohen, Julie E. *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press, 2019.

Dahir, Abdi Latif. "Africa Internet shutdowns grow longer in Cameroon, Chad, Ethiopia." *Quartz Africa*, November 19, 2018.

Dahir, Abdi Latif. "Half the world is now connected to the internet–driven by a record number of Africans." *Quartz Africa*, December 11, 2018.

Dahlgren, Peter. "The Internet, public spheres, and political communication: Dispersion and deliberation." *Political communication* 22, no. 2 (2005): 147-162.

De Gregorio, Giovanni, and Nicole Stremlau. "Internet Shutdowns and the Limits of Law." *International Journal of Communication* 14 (2020): 20.

Dimitrakopoulou, Dimitra, Georgios Tzogopoulos and Alexandra Nikolakopoulou. *The Role of Social Media in Violent Conflict*, INFOCORE Working Paper 2014/05, (2014).

Donovan Joan, Brian Friedberg and Emily Dreyfuss. “The Capitol siege was the biggest media spectacle of the Trump era,” *The Guardian*, January 11 (2021) https://www.theguardian.com/commentisfree/2021/jan/11/the-capitol-siege-was-the-biggest-media-spectacle-of-the-trump-era.

Festinger, Leon. *A theory of cognitive dissonance*. Vol. 2. Stanford University Press, 1962.

Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

Hamelink, Cees Jan. *Media and conflict: Escalating evil*. Routledge, 2015.

Herzstein, Robert Edwin. *The war that Hitler won: The most infamous propaganda campaign in history*. Putnam Publishing Group, 1978.

Howard, Philip N. *The digital origins of dictatorship and democracy: Information technology and political Islam*. Oxford University Press, 2010.

Howard, Philip N., Sheetal D. Agarwal, and Muzammil M. Hussain. "When do states disconnect their digital networks? Regime responses to the political uses of social media."*The Communication Review* 14, no. 3 (2011): 216-232.

Klonick, Kate. "The new governors: The people, rules, and processes governing online speech."*Harv. L. Rev.* 131 (2017): 1598.

Marchant, Eleanor and Nicole Stremlau. “The Changing Landscape of Internet Shutdowns in Africa”, International Journal of Communication, 14(2020), 4216-4223

Marchant, Eleanor and Nicole Stremlau. “A Spectrum of Shutdowns: Reframing Internet Shutdowns from Africa” International Journal of Communications 14(2020): 4327-4342.

Matfess, Hilary. "More African countries are blocking internet access during elections”, *Quartz Africa*, June 1, 2016.

Metzl, Jamie F. "Information intervention: When switching channels isn't enough." *Foreign Affairs* (1997): 15-20.

Micek, Peter and Deji Olukotun. "Internet disrupted in Bahrain around protests as wrestling match sparks shutdown in India." *Access Now,* 24 June 2016.

Morozov, Evgeny. The net delusion: The dark side of Internet freedom. *PublicAffairs*, 2012.

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

Noble, Safiya Umoja. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

The Organisation for Economic Co-operation and Development“OECD digital economy outlook 2017, (2017). https://www.oecd.org/sti/ieconomy/oecd-digital-economy-outlook-2017-9789264276284-en.htm.

Office of the High Commissioner for Human Rights. *Preventing incitement to hatred and violence in the Central African Republic*. https://www.ohchr.org/EN/NewsEvents/Pages/PeacekeepersDay2019.aspx.

Olewe, Dickens. "Kenya, Uganda and Tanzania in “anti-fake news campaign.” *BBC News*, 16 May 2018.

Olukotun, Deji, Peter Micek and Gustav Bjorksten. "Vietnam blocks Facebook and cracks down on human rights activists during Obama visit." *Access Now*, 23 May 2016.

Potkin, Fanny and Poppy McPherson. “‘Spreading like Wildfire: Facebook Fights Hate Speech before Myanmar Poll,” *Reuters*, November 5, 2020 https://www.reuters.com/article/myanmar-election-facebook-idUSL4N2HQ3QU.

Posetti, Julie and Kalina Bontcheva. "Disinfodemic: deciphering COVID-19 disinformation. Policy brief 1," 2020. https://en.unesco.org/covid19/disinfodemic.

Price, Monroe E. and Mark Thompson, eds. *Forging peace: intervention, human rights, and the management of media space*. Indiana University Press, 2002

Purnell, Newley. "Sri Lankan Islamist Called for Violence on Facebook Before Easter Attacks."*Wall Street Journal*, April 30, 2019. https://www.wsj.com/articles/sri-lankan-islamist-called-for-violence-on-facebook-before-easter-attacks-11556650954.

Radu, Roxana. Fighting the ‘Infodemic’: Legal Responses to COVID-19 Disinformation. *Social Media+ Society*, *6*(3), 2020. 2056305120948190.

Roberts, Sarah T. *Behind the screen*. Yale University Press, 2019.

Sablosky, Jeffrey. “Dangerous organizations: Facebook’s content moderation decisions and ethnic visibility in Myanmar” *Media, Culture & Society* (2021).

Shahbaz, Adrian. "The Rise of Digital Authoritarianism: Freedom on the Net 2018." *Freedom House*, October 2018.

Statista. “Government Justifications for Internet Shutdowns Worldwide 2019.” Statista. https://www.statista.com/statistics/1096316/government-justifications-for-internet-shutdowns/#:~:text=Official%20government%20justifications%20for%20internet%20shutdowns%20worldwide%202019&text=Fake%20news%20and%20hate%20speech. Accessed March 25, 2021.

Statt, Nick. “How Facebook Is Using AI to Combat COVID Misinformation and Detect ‘Hateful Memes.” *The Verge*, May 12, 2020. Accessed March 25, 2021. https://www.theverge.com/2020/5/12/21254960/facebook-ai-moderation-covid-19-coronavirus-hateful-memes-hate-speech.

Stecklow, Steve. “Why Facebook is losing the war on hate speech in Myanmar” *Reuters*, Aug. 15 2018 https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

Stremlau, Nicole. *Media, Conflict and the State in Africa*. Cambridge University Press (2018).

Tufecki, Zeynep. “YouTube. The Great Radicalizer” *New York Times*, May 10, 2018 https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.

Youssef, Nour. "Algeria’s answer to cheating on school exams: Turn off the Internet." *The New York Times*, June 21, 2018.

Zeitzoff, Thomas. "How social media is changing conflict." *Journal of Conflict Resolution* 61, no. 9 (2017): 1970-1991.