Protecting the Freedom of Expression in an Era of "Platformization:" Paving a Road to Censorship?

Jacob Mchangama, Natalie Alkiviadou

Abstract: To tackle the problems that arise with the horizontalization of content moderation and the resulting ramifications on free speech, this chapter proposes International Human Rights Law (IHRL) as a framework of first reference to re-imagine the current process of moderating contentious speech such as hate speech. Further, it looks at South African jurisprudence which adopts a nuanced and substantiated approach to the free speech – hate speech question, jurisprudence which can serve as an interpretational aide for IHRL provisions. Whilst the chapter recognizes the weakness of marrying IHRL with practices of private companies which are not bound by it, the chapter explains and concludes that IHRL can and should be developed into a workable solution for private companies in the ambit of content moderation of contentious speech.

Keywords: human rights, freedom of expression, hate speech, South Africa, global platforms.

Chapter 1. Introduction

In the 1990s, the Internet was seen as an unstoppable force for the globalization of freedom of expression. As Stanford professor Lawrence Lessig put it: "Nations wake up to find that their telephone lines are tools of free expression, that e-mail carries news of their repression far beyond their borders, that images are no longer the monopoly of state-run television stations but can be transmitted from a simple modem."[1]

1 Lawrence Lessig, "Code: version 2.0", Basic Books, 2006, 236.

Today nearly 60% of the global population – 4,66 billion people - are online and 4,20 billion are active social media users.[2] The transformation of social media platforms into the central agora where ideas are imparted and received has indeed given an unprecedented number of people a voice in local and global affairs. Yet, in tandem with the ability to organize protests, scrutinize the actions of decision makers and make visible marginalized minorities, social media has provided a platform to extremism, terrorist content, disinformation at scale, and hate speech.

But for governments alarmed about the corrosive effects of social media, the centralized amplification of hate, harm, and hoaxes comes with a silver lining. If major platforms like Facebook, YouTube, and Twitter can be forced or persuaded into purging illegal and lawful but awful content, they can become digital chokepoints, with the visibility of illegal content dropping exponentially. Potentially, centralized platforms could even end up serving as the private enforcers of government censorship, entirely inverting the initial promise of egalitarian and unmediated free speech. The most extreme examples of this development can be seen in countries like India, Russia and Turkey[3] where intense pressure is being brought on platforms to remove speech deemed illegal or undesirable by the respective governments. A less draconian – but highly influential - version of this strategy can be seen in, *inter alia*, the pioneering German Network Enforcement Act 2017 (NetzDG) and non-binding measures such as the Christchurch Call for Action.

These initiatives combined with the sheer scale of user generated content have arguably contributed to platforms significantly expanding their efforts to police and purge hate speech. The NetzDG obligates social media platforms with a minimum of 2 million users to remove illegal content –

---

2 Datareportal: "Digital 2021: Global Overview Report", 27 January 2021. https://datareportal.com/reports/digital-2021-global-overview-report#:~:text=Internet%3A%204.66%20billion%20people%20around,now%20stands%20at%2059.5%20percent.

3 96% of the total global volume of demands originated from only five countries (including Russia, Turkey and India) Twitter removal requests. https://transparency.twitter.com/en/reports/removal-requests.html#2020-jan-jun; Karan Deep Singh & Paul Mozur, "As Outbreak Rages, India Orders Critical Social Media Posts to be Taken Down", *New York Times* 25 April 2021. https://www.nytimes.com/2021/04/25/business/india-covid19-twitter-facebook.html; Human Rights Watch,"Russia: Social Media Pressured to Censor Posts: Fines, Smear Campaigns, Potential Blocking for Non-Compliance", 5 February 2021. https://www.hrw.org/news/2021/02/05/russia-social-media-pressured-censor-posts>; Human Rights Watch, "Turkey: Social Media Law will Increase Censorship" 27 July 2020, https://www.hrw.org/news/2020/07/27/turkey-social-media-law-will-increase-censorship.

including hate speech – within 24 hours, or risk large fines of up to 50 million euros. In the first quarter of 2018 (when the NetzDG had entered into force) Facebook removed 2,5 million pieces of content for violating its Community Standards on hate speech. This rose to 4 million in the first quarter of 2019 and 9,5 million in the first quarter of 2020. By the first quarter of 2021, Facebook purged 25.2 million pieces of 'hate speech' content.[4] This development also reflects that platforms increasingly rely on artificial intelligence to proactively identify and even remove content violating national laws and/or their terms of service. Their rate of content proactively identified by Facebook increased from 38% in the first quarter of 2018 to 96.8% in the first quarter of 2021.[5] While states impose intermediary liability to counter online harms, 'outsourcing' government mandated content regulation to private actors raises serious questions about the consequences on online freedom of expression.

The global nature of social media platforms used by people in almost all countries around the world create significant problems when it comes to determining where to draw the line on various categories of content. In the abstract, large majorities across the globe find it very important that people can speak their mind and use the Internet without censorship. However, once moving from the abstract to specific categories of speech, there are marked variations of tolerance within and between populations of countries as well as between various governments. There is, for instance, no universal agreement on whether statements offensive to minorities should be tolerated. In the Scandinavian countries and the US around 65 % of the populations believe that free speech should extend to statements offensive to minority groups while around 80%. Conversely in Kenya, Indonesia, Turkey and Tunisia, only between 18 and 27% of the populations favor tolerating such statements.[6]

One proposed remedy to bridge the gap between the conflicting attitudes and legal regimes which global social media platforms are forced to navigate is for these private actors to rely on International Human Rights Law (IHRL) when adopting their terms of service and moderating

4 Facebook Transparency Center, "Hate Speech". https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook.

5 Facebook Transparency Center, "Proactive Rate". https://transparency.fb.com/policies/community-standards/hate-speech/.

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

content, even if not formally bound by such legal instruments. Placing content moderation in the framework of IHRL has been discussed by scholars such as Aswad[7] and Benesch,[8] who argue that IHRL, with some modification, can be used by social media companies to moderate online content. Dvoskin takes a different approach, highlighting that adopting IHRL "might not lead to more legitimate content moderation" since this area of law "leaves many speech questions unanswered."[9] It is important to note that there are crucial differences between criminal law and private content moderation. The former involves the threat of criminal sanctions, including – ultimately - the risk of prison, whilst the latter ´merely´ results in the removal of content or, at worst, the deletion of user accounts. Moreover, when restricting freedom of expression, States must follow time consuming criminal procedures and respect legally binding human rights standards. On the other hand, private platforms are generally free to adopt terms of service and content moderation practices less protective of freedom of expression and due process than what follows under IHRL. However, when governments impose intermediary liability on private platforms through laws prescribing punishments for non-removal, platforms are essentially required to assess the legality of user content as national authorities.

In 2018, the(n) UN Special Rapporteur on the Freedom of Opinion and Expression (SRFOE), David Kaye asserted that "human rights law gives companies the tools to articulate their positions in ways that respect democratic norms and counter authoritarian demands".[10]

Given the problems with conflicting legal regimes and popular attitudes towards the limits of free speech, it is tempting to support David Kaye´s assertion that IHRL paves away ahead in the current impasse. After all, IHRL claims to be universal in nature and most states across all continents

---

7 Evelyn Mary Aswad, "The Future of Freedom of Expression Online" *Duke Law & Technology Review* 17, No.1, 2018) , 52-53.

8 Susan Benesch, "But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies" *Yale Journal on Regulation Online Bulletin 39*, No.3, 2020, 90.

9 Brenda Dvoskin, "International Human Rights Law is not Enough to Fix Content Moderation's Legitimacy Crisis", *Berkman Klein Center for Internet & Society at Harvard University*, 16 September 2020. https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd.

10 United Nations, "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression" A/HRC/38/35. 2018. https://www.undocs.org/A/HRC/38/35.

have ratified conventions such as the ICCPR (of course ratification does not necessarily entail compliance or genuine commitment). However, it should be acknowledged that there are serious challenges to adopting an IHRL approach to content moderation. First of all, IHRL is binding on states, not on private companies, and while the UN has developed Guiding Principles on Business and Human Rights, these are aspirational and not legally enforceable. Moreover, there are good reasons why social media platforms should be allowed to adopt and experiment with different models of terms of service and content moderation practices dependent on their size, architecture, content, focus etc. Whether content is lawful or not is a complex exercise that is heavily dependent on careful context-specific analysis. Under IHRL, restrictions of freedom of expression must comply with strict requirements of legality, proportionality, necessity and legitimacy. These requirements make the individual assessment of content difficult to reconcile with legally sanctioned obligations to process complaints in a matter of hours or days, not to mention automated content moderation. In a 2021 study Justitia found that the available data showed that on average Council of Europe member states used more than 775 days to process hate speech cases in their national criminal law system from the date of the alleged offending speech till the conclusion of the trial at first instance[11], a time frame wholly incommensurate with how fast platforms are required to remove illegal content under intermediary liability laws such as NetzDG. All these factors mean that a human rights approach to content moderation will necessarily have to be adapted to rather than copied from the current state centric model.

However, this chapter will narrowly focus on how IHRL can contribute to the definition and moderation of the controversial and contested category of "hate speech", which is at the centre of much debate and subject to increasing regulatory scrutiny by both social media platforms themselves as well as numerous states as shown above. This question is all the more relevant given the lack of any authoritative definition of hate speech and widely differing legal standards at both the state and international level. The authors argue that the interplay between ICCPR articles 19 and 20 forms the natural framework for defining and interpreting hate speech under IHRL. In recent years much effort has been spent by both the

11 Jacob Mchangama et al, "Rushing to Judgment: Are Short Mandatory Takedown Limits for Online Hate Speech Compatible with the Freedom of Expression" *Justitia*, January 2021. https://futurefreespeech.com/wp-content/uploads/2021/01/FFS_Rushing-to-Judgment-3.pdf.

Human Rights Committee (HRC), the SRFOE and member states on trying to clarify and strengthen the protection of freedom of expression under article 19, while simultaneously attempting to more clearly and narrowly define the categories of speech that qualify as impermissible hate speech under article 20(2), resulting in a number of soft law instruments as detailed below.

However, given the non-binding nature of these soft law instruments and the paucity of relevant decisions in actual hate speech cases from the HRC, the chapter will do a comparative analysis of two other sources of hate speech jurisprudence, that might be used as an interpretive guide for identifying the obligations under ICCPR articles 19 and 20, when applied in practice. First, the chapter will examine hate speech case law of the European Court of Human Rights (ECtHR) and subsequently relevant hate speech jurisprudence from the South African Constitutional Court and Supreme Court of Appeal. It will be argued that the South African model provides a more convincing, consistent and robust approach to balancing speech protected by freedom of expression against speech which falls afoul of the ban against hate speech as per the dichotomy of ICCPR articles 19 and 20. Conversely it will be argued that the jurisprudence of the ECtHR suffers from serious shortcomings that would add more confusion than clarity and weaken rather than strengthen the protection of freedom of expression if forming the basis of a human rights approach to online content moderation.

Chapter 2. International Human Rights Law: A Framework of First Reference?

1. Pros and Cons to an IHRL approach to Online Content Moderation

There are currently 173 state parties to the ICCPR, making it the most widely accepted convention regulating civil and political rights, including freedom of expression. Accordingly, ICCPR forms the natural focus point of an IHRL approach to content moderation. Article 19 (2) guarantees that "everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice". The fact that article 19 ensures the right to both receive and impart information regardless of frontiers and choice of media, is highly relevant to the Internet and social media, suggesting a positive obligation to facilitate access to information.

Article 19(3) sets out a number of permissible restrictions to freedom of expression as well as procedural and substantive safeguards that must accompany any such restrictions.

Article 19 (3) incorporates a three-part test for limiting freedom of expression. Restrictions must be "provided by law" and are "necessary" for, amongst others, "the respect of the rights or reputations of others" which for hate speech cases is the most relevant of grounds. When it comes to proportionality, the HRC notes that restrictions must be "appropriate to achieve their protective function",[12] and "must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interests to be protected."[13]

In 2011 the HRC published General Comment 34 (GC 34), which constitutes the most authoritative guidance to the obligations under article 19. According to GC 34 the ICCPR protects "even expression that may be regarded as deeply offensive."[14] This seems to entail a heightening of the threshold, which must be met before speech – including hate speech - can be restricted under article 19. For instance, in GC 34 the HRC has held that "Laws that penalize the expression of opinions about historical facts are incompatible with the obligations that the Covenant imposes on States parties in relation to the respect for freedom of opinion and expression. The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events". [15]

This holding can be contrasted with the HRC´s decision in *Faurisson v France*, in which an academic challenged the use of gas for extermination at Nazi concentration camps. Faurisson was convicted for contesting crimes against humanity, with the HRC finding no violation of the freedom of expression as provided for by article 19. It held that "the restrictions placed on the author did not curb the core of his right to freedom of expression, nor did they in any way affect his freedom of research; they were intimately linked to the value they were meant to protect - the right to be free from incitement to racism or anti-Semitism; protecting that value could not have been achieved in the circumstances by less drastic means."

Accordingly, it would seem that post-GC 34 article 19 now prohibits so-called "memorial laws" criminalizing the denial of historical events such

12 HRC General Comment 34, para. 34.
13 HRC General Comment 34, para. 34.
14 HRC General Comment 34, para. 11.
15 HRC General Comment 34, para. 49.

as the Holocaust, which as we shall see also marks a decisive difference between the HRC and the ECtHR.

2. *Article 20(2): An Analysis*

Article 20 (2) not only permits restrictions of freedom of expression, but states that "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law".

As with 4 of the International Convention on the Elimination of All Forms of Racial Discrimination which prohibits, amongst others, the dissemination of racist ideas, 20 differs to other articles in the ICCPR since it imposes a positive obligation on states to prohibit certain types of speech. However, the HRC holds that "articles 19 and 20 are compatible with and complement each other. The acts that are addressed in article 20 are all subject to restriction pursuant to article 19, paragraph 3."[16] In *Ross v Canada*, the HRC underlined that "restrictions on expression which may fall within the scope of article 20 must also be permissible under article 19, paragraph 3."[17]

The 2012 report of the SRFOE underlined that "the threshold of the types of expression that would fall under the provisions of article 20(2) should be high and solid."[18] In 2011, the Office of the United Nations High Commissioner for Human Rights organised a series of expert workshops on incitement to national, racial or religious hatred, as reflected in IHRL.[19] The workshops resulted in the Rabat Plan of Action (RPA) which was launched in 2013.[20] It provides that there must be a high threshold when applying article 20 of the ICCPR.[21] To achieve this, the RPA sets

16 HRC General Comment 34: para. 50.
17 Ross v Canada Communication no 736/1997 (18 October 2000) CCPR/C/70/D/736/1997, para. 10.6.
18 Ross v Canada, para.45
19 International Justice Resource Center, "UN Launches the Rabat Plan of Action", 25 February 201. https://ijrcenter.org/2013/02/25/un-launches-the-rabat-plan-of-action/.
20 Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (launched in 2013) para. 6.
21 Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22.

out a six-part threshold test to be referred to when applying article 20(2) and includes the assessment of the (i) social and political context (ii) status of the speaker, (iii) intent to incite the audience against a target group (iv) content and form of the speech (iv) extent of its dissemination and (vi) likelihood of harm, including imminence. Since its adoption, the RPA been referred to in several documents, such as in Human Rights Council Resolution 16/18 and the United Nations Strategy and Plan of Action on Hate Speech (2020).[22] The SRFOE has also referenced the RPA extensively including in the 2019 report on online hate speech.[23]

There has been relatively little case law before the HRC on article 20(2) and the degree the six-part test of the RPA has been adopted by the HRC. As such, how this test might apply to real cases cannot be readily discerned. However, in *Mohamed Rabbae, A.B.S and N.A v The Netherlands* from 2016, the HRC gave a relatively extensive overview of article 20(2). Here, the authors claimed to be victims of a violation of their rights under article 20(2) due to allegedly racist statements made by Geert Wilders, leader of the far-right Dutch Freedom Party and his subsequent acquittal by the domestic court. The HRC found that article 20(2) secures the right of persons to be free from hatred and discrimination, but that it is "crafted narrowly" to ensure the protection of freedom of expression. It recalled that this freedom may include "deeply offensive" speech and speech which is disrespectful for a religion, unless the strict threshold of article 20(2) is met.[24] The HRC found no violation of article 20(2) since the Netherlands had developed a suitable legislative framework which victims could reach out to, thereby ensuring that it took the necessary and proportionate measures to prohibit statements made in violation of article 20(2).[25] Relevant to the high threshold attached to article 20(2) is also the concurring individual opinion of Cleveland (Vice Chair of the HRC at the material time) and Politi, in which they noted, amongst others, that "hate speech and similar laws ironically are often employed to suppress the

22 United Nations "Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences", September 2020. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf.

23 Report of the Special Rapporteur on the Freedom of Opinion and Expression, "Online Hate Speech" A/74/486, 9 October 2019. https://www.undocs.org/A/74/486.

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

25 Mohamed Rabbae, A.B.S and N.A v The Netherlands, para. 10(7).

very minorities they purportedly are designed to protect."[26] Importantly, their concurring opinion noted the uniqueness in article 20, insofar as it requires the restriction of the "highly protected freedom of expression." This, they argue, means that article 20(2) is "narrowly circumscribed and sets the bar high for the expression that must be prohibited," [27]demonstrating just how narrowly they have construed article 20 to be. Moreover, the finding in favour of the Netherlands is also reflective of this narrow construction.

Despite the lack of binding case law and the paucity of decisions by the HRC, it is submitted that the ICCPR provides a suitable "framework of first reference" for the determination of hate speech by private social media companies, even if not formally bound by this convention.

It would also be a suitable compass for states who are seeking to impose more and more moderation duties at risk of penalties and in short time frames. The post-GC 34 cases on hate speech, the RPA and the guidance and opinions of the SRFOE can guide private companies along the path of adequately protecting the fundamental freedom of expression whilst simultaneously ensuring the safety and dignity of their users. However, the lack of a substantial body of case law applying these principles to specific instances of controversial speech, means that additional sources of hate speech jurisprudence might be needed to help interpretate the relationship between articles 19 and 20.

Chapter 3. The European Court of Human Rights: A Template to Avoid?

No other human rights court has made more decisions in general or on hate speech specifically than the ECtHR. Given that the ECtHR has jurisdiction over 47 member states ranging from Ireland to Azerbaijan and Iceland to Turkey, and that the majority of these states are democracies, it might be tempting to use ECtHR case law on hate speech as a guide to the interpretation of ICCPR article 20(2).

However, there are fundamental differences in the way the ECtHR and the HRC approaches the question of hate speech, and the amount of

---

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

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

weight these two bodies attach to freedom of expression in such cases. To commence our discussion, we turn to an analysis conducted by Justitia on a total of 60 identified cases of the former European Commission of Human Rights and the ECtHR, decided upon between 1979-2020.[28] 57 of those cases were brought by speakers and 3 by the targets/victims. *61% of cases brought by the speakers resulted in the applicant's loss* through a finding of non-violation of article 10 (on freedom of opinion and expression): (21%), incompatible *ratione materiae* (9%) and manifestly ill-founded (32%). Only 39% of cases brought by the speakers on the grounds of an article 10 violation have resulted in a finding in favour of the applicant. Thus, on average, free speech restrictions have been upheld in just over one out of three hate speech cases before the ECtHR.

To demonstrate this in a qualitative manner, we will turn to some indicative (by no means exhaustive) case-law, which will cover a range of protected characteristics.

The 2009 case of *Féret v. Belgium* was brought by the leader of a nationalist Belgian party who had been ceased from office for a period of ten years for, amongst others, the preparation and dissemination of publications which included statements of the following sort: "Stop the Islamization of Belgium," "Save our people from the risk posed by Islam, the conqueror." The Court did not succumb to the Belgian government's request for an invocation of article 17 of the European Convention on Human Rights (ECHR) which prohibits the abuse of Convention rights but ruled that there was no violation of article 10. Importantly for the threshold discussion, the Court noted that:

> incitement to hatred did not necessarily call for specific acts of violence or other offences. Insults, ridicule or defamation aimed at specific population groups or incitation to discrimination, as in this case, sufficed for the authorities to give priority to fighting hate speech when confronted by the irresponsible use of freedom of expression which undermined people's dignity, or even their safety.[29]

Therefore, hate speech was deemed to include even insults and ridicules. This line of reasoning was continued in *Vejdeland v Sweden (2012)* which involved the dissemination of homophobic leaflets in school lockers. The Court found no violation of article 10 and reiterated its findings in *Féret*,

28 For the full database and quantitative illustrations visit: https://futurefreespeech.com/hate-speech-case-database/.

29 Féret v Belgium, Application no. 15615/07 (ECHR 16 July 2009) para. 73.

noting that "although these statements did not directly recommend individuals to commit hateful acts, they are serious and prejudicial allegations."[30]

The low threshold was further embedded in the Court's approach in *Lilliendahl v Iceland*, which involved an applicant who wrote comments below an online news article reporting a municipal decision to strengthen education and counselling in schools for pupils identifying as lesbian, gay, bisexual or transgender. The applicant used derogatory comments such as "sexual deviation" when referring to homosexuality and said that this is "disgusting. To indoctrinate children with how sexual deviants copulate in bed." This was the first time that the Court took the question of hate speech and what it means on a more conceptual level (the term yet remains undefined by the Court). It put forth two categories of hate speech, the first being the "gravest forms of hate speech…which fall under Article 17"[31] and the second being the "less grave forms of hate speech" which include "attacks on persons committed by insulting, holding up to ridicule or slandering specific groups of the population [and which] can be sufficient for allowing the authorities to favour combating prejudicial speech…"[32]

The ECtHR has also put forth conflicting positions when it has come to insult in other cases. For example, *Ibragim Ibragimov and Others v Russia* (2018) was a case which involved the banning of Muslim scholar Said Nrusi's book, as it was extremist. Here, the ECtHR found a violation of article 10, noting that:

> 'merely because a remark may be perceived as offensive or insulting by particular individuals or groups does not mean that it constitutes "hate speech." Whilst such sentiments are understandable, they alone cannot set the limits of freedom of expression. The key issue in the present case is thus whether the statements in question, when read as a whole and in their context, could be seen as promoting violence, hatred or intolerance.'[33]

30 Vejdeland and Others v Sweden, Application No. 1813/07 (ECHR 9 February 2012) para. 54.
31 Lilliendahl v Iceland (2020) Application No.29297/18 (ECHR 12 May 2020) para. 34.
32 Lilliendahl v Iceland (2020) para.36.
33 Ibragim Ibragimov and Others v Russia (Application nos. 1413/08 and 28621/11) para 115 .

However, in a case two years later, namely *Atamanchuk v Russia* (2020), the Court took a different approach. Here, the applicant, a journalist/politician was convicted of making statements against non-Russians, referring to them as criminals (without calling for violence). The Court found no violation of article 10, underlining that:

> 'inciting hatred does not necessarily involve an explicit call for an act of violence, or other criminal acts. Attacks on persons committed by insulting, holding up to ridicule or slandering specific groups of the population can be sufficient for the authorities to favour combating xenophobic or otherwise discriminatory speech in the face of freedom of expression exercised in an irresponsible manner.'[34]

Therefore, in the 2018 case, insult was not considered to be sufficient to allow for a restriction to article 10 whereas in the latter it was. Noteworthy is the fact that the 2020 case involved speech directed towards an ethnic group, which the Court appears to have a lower tolerance towards.

Another illustration of the inconsistency in the ECtHR's approach is the manner in which it deals with historical events. For example, the Court systematically finds negationist or revisionist speech in relation to the Holocaust[35] to constitute hate speech, sometimes ousted through the application of the so-called abuse clause in article 17. However, in a case involving the denial of the Armenian genocide,[36] it ruled that this fell within the framework of protected speech.

The treatment of totalitarian symbols is yet another indication of the contradictions found in the Court's approach to alleged hate speech. In *Fáber v Hungary (2012)*,[37] the Court found that article 10 protected an applicant who held a striped Árpád flag[38] less than 100 metres away from a demonstration against racism and hatred. In *Vajnai v Hungary* (2008),[39] during a demonstration, the applicant wore a red communist star and was convicted of the offence of using a *totalitarian symbol which the ECtHR*

---

34 Atamanchuk v Russia, Application no. 4493/11 (ECHR 11 February 2020) para.52.

35 See, inter alia, Williamson v Germany, Application No. 64496/17 (ECHR 8 January 2019), Pastörs *v. Germany, Application No. 55225/14 (ECHR 3 January 2020), Garaudy v France, Application No. 64496/17 (ECHR 7 July 2003).*

36 Perinçek v Switzerland, Application No. 27510/08 (ECHR 15 October 2015).

37 Fáber v Hungary, Application No.40721/08, ECHR 24 October 2012.

38 Used by the Hungarian Fascist Arrow Cross party, responsible for crimes against Jews during World War II.

39 Vajnai v Hungary, Application No.33629/06, ECHR 8 July 2008.

found to be a violation of the applicant's freedom of expression. However, in the recent case of *Nix v Germany* (2018) – which a German blogger was convicted for using symbols of a banned organization after posting a picture of Heinrich Himmler wearing a swastika armband and likening him to the officers of an employment office which he alleged discriminated against his mixed-race daughter. Despite the fact that the applicant neither advocated nor defended Nazism, the Court found the conviction justified.[40]

In sum, these cases reflect that the ECtHR attaches a low threshold to freedom of expression when it comes to hate speech. This has led to an inconsistent and incoherent case law, with no proper demarcation between freedom of expression and hate speech, resulting in the permissible restriction of speech deemed merely "offensive" or "prejudicial", but with no clear nexus to any harm, speech which included no hateful intent and the selective restriction of the denial of historical events. The ECtHR case law thus fails to satisfy several of the elements of the RPA, and the higher thresholds for restricting hate speech developed by the HRC. Accordingly, using the ECtHR´s case law as a guide to interpreting ICCPR articles 19 and 20 would result in increased confusion, less clarity and a lower degree of protection of freedom of expression.

Chapter 4. South Africa: A Good Practice Template

As noted in the section on the ECtHR, social media companies may look at sources such as Court judgements for inspiration on their content moderation practices. For purposes of providing a well-rounded overview of what is out there in terms of good practices in the ambit of handling hate speech, this chapters offers an overview of key (but not exhaustive) hate speech cases that were heard before the highest courts of South Africa. We choose this country as South Africa has only relatively recently become a liberal democracy after emerging from a long period of white supremacy, which systematically denied both the equality, dignity and the freedom of expression of its non-white population. Accordingly, South Africa is perhaps uniquely suited to act as a "laboratory" when it comes to safeguarding the values of freedom, equality and dignity. Moreover, the South African Constitution is explicitly founded on the values of, inter alia, human rights, and obliges South Africa to "consider international

40 Nix v Germany, Application No. 35285/16, ECHR 13 March 2018 Para. 47.

law" – including IHRL – when interpreting the constitution´s bill of rights. South African courts frequently rely on international precedents, including the ICCPR, when interpreting the South African constitution´s bill of rights. These factors have, we submit, contributed to South African courts developing a nuanced and substantiated approach to the treatment of hate speech, taking into consideration both the fundamental nature of free speech but also the importance of maintaining dignity and equality.

Section 16 of the South African constitution provides for the freedom of expression. Part 2 therein notes that this freedom does not extend to, *inter alia*, "the advocacy of hatred that is based on race, ethnicity, gender or religion, and that constitutes incitement to cause harm." This provision differs from article 20(2) ICCPR since it is not a positive obligation to prohibit hate speech but, instead, means that hate speech (which meets a certain threshold) is exempt from constitutional protection.

The case *Islamic Unity Convention v Independent Broadcasting Authority and Others* involved statements made on a radio show by a historian who denied the legitimacy of Israel and argued that Jews were not gassed during WWII. The South African Jewish Board of Deputies claimed that the broadcast contravened the Code of Conduct for Broadcasting Services since it was "likely to prejudice relations between sections of the population."

In its judgment, the Court pointed out that freedom of expression:

> …. lies at the heart of a democracy. It is valuable for many reasons, including its instrumental functions as a guarantor of democracy, its implicit recognition and protection of the moral agency of individuals in our society and its facilitation of the search for truth by individuals and society generally. The constitution recognizes that individuals in our society need to be able to hear, form and express opinions and views freely on a wide range of matters….[41]

The Court placed its analysis of expression within a historical context, reiterating the country's recent restrictive past and noting that restrictions would be incompatible with a "constitutionally protected culture of openness and democracy and universal human rights for South Africans of all ages, classes and colours."[42]

---

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

42 Islamic Unity Convention v Independent Broadcasting Authority and Others para. 25.

The Court further explained the hate speech threshold and the requirement of its real life impact by noting that "not every expression of speech that is likely to prejudice relations between sections of the population would be 'propaganda for war', or 'incitement of imminent violence' or 'advocacy of hatred' which also constitutes 'incitement to cause harm'."[43] This was reiterated in subsequent case-law, such as *Qwelane* discussed below.

The Court ruled that the Code's section prohibiting the impugned speech was broader than what was permissible under the Constitution as it referred to "a section of the population" and not a specific group. It further noted that the reference to "prejudice" did not meet the harm requirement needed for satisfying section 16 of the constitution. Comparatively, two points can be made. Firstly, that, by protecting prejudicial speech, the Court's decision is in line with the high threshold set out by article 20(2) of the ICCPR and further assessed by the RPA as well as HRC case law (see for example *Rabbae* and the extension of the freedom of expression to 'deeply offensive' speech). In addition, the test developed by the Constitutional Court in the Islamic Unity Convention case is more speech protective than the ECtHR which has permitted the restriction of prejudicial speech (see, for example, *Vejdeland*). Moreover, the decision sides with GC 34 over the case law of the ECtHR when it comes to the controversial question of whether to protect even the denial of historical crimes such as the Holocaust.

In relation to incitement, the Constitutional Court recently held that a law criminalizing incitement to "any offence" was "unquestionably overbroad and its inhibition of free expression is markedly disproportionate to its conceivable benefit to society."[44] The case revolved around statements made by the president of the political party "Economic Freedom Fighters," who called his supporters to illegally occupy land. In his majority decision Chief Justice Mogoeng Mogoeng noted that freedom of expression is the 'lifeblood of constitutional democracy' and that '[w]hen citizens are very angry or frustrated, it serves as the virtual exhaust pipe through which even the most venomous of toxicities within may be let out to help them calm down, heal, focus and move on.'

---

43 Islamic Unity Convention v Independent Broadcasting Authority and Others para. 34.

44 Economic Freedom Fighters, Julios Selo Malema v Minister of Justice and Correctional Services, National Director of Public Prosecutions, Case CCT 201/19, Para. 61.

The Court's position was, once again, informed by the country's apartheid history. The judgement referred to the fact that the right to freedom of expression was violated during the "highly intolerant and suppressive past"[45] and, it "thus has to be treasured, celebrated, promoted and even restrained with a deeper sense of purpose and appreciation of what it represents." Although the Court also emphasized that freedom of expression is not absolute, nor more important than other rights, it stressed that limitations can only occur in specific circumstances, such as when national interest, dignity, physical integrity or democracy is threatened. The Court noted that this complied with the country's international obligations in respect of limitations to free expression making a specific reference to article 19 ICCPR.[46]The Supreme Court´s view of free speech as a vital democratic exhaust pipe and the country´s history of white supremacy as a caution *against* censorship, marks a stark difference to the ECtHR. The Strasbourg court tends to stress the (supposed) capability of controversial speech to cause harm and danger – even absent any direct incitement to harm - and sees European history as offering a compelling argument *in favour* of restricting extreme speech.

The Supreme Court of Appeal (SCA) has also developed a high threshold in relation to the restriction of hate speech, as witnessed in the case of *Qwelane v South African Human Rights Commission*. This case involved a 2008 publication by Jon Qwelane, a well-known anti-apartheid activist and journalist in the Sunday Sun. The article was titled "Call me names but gay is not okay…" and used homophobic language and was accompanied by a cartoon comparing homosexuality to bestiality. The article stated, *inter alia*, that:

> The real problem, as I see it, is the rapid degradation of values and traditions by the so-called liberal influences of nowadays; you regularly see men kissing other men in public, walking holding hands and shamelessly flaunting what are misleadingly termed their 'lifestyle' and 'sexual preferences…. At this rate how soon before some idiot demands to 'marry' an animal and argues that this constitution 'allows it'?

45 Economic Freedom Fighters, Para. 2.
46 It does so in footnote 51.

In 2017, the Johannesburg High Court decided that certain statements were "hurtful, harmful, incite[d] harm and propagate[d] hatred"[47] thereby violating Section 10(1) of the Equality Act. Qwelane appealed the case to the SCA on the grounds that the Equality Act's definition of hate speech was unconstitutional since it prohibited more speech than provided for in section 16(2) of the Constitution. The SCA referred to the freedom of expression as the "lifeblood of a democratic society." It noted that section 10 of the Equality Act did, in fact, go beyond what was constitutionally permissible under section 16(2) and warned that "one must be careful not to stifle the views of those who speak out of genuine conviction."[48] It placed its assessment within a historical framework, holding that "given our history...freedom of expression must also be prized."[49] As such, it found section 10 of the Equality Act to be unconstitutional and gave Parliament 18 months (as of November 2019) to remedy the current content of the said section. The high threshold adopted in *Qwelane* by the SCA particularly was based on two cases which were heard together in 2018, namely *Moyo v Minister of Justice and Constitutional Development* and *Sonti v Minister of Justice and Correctional Services and Others* (2017). The SCA noted that, for restrictions to speech to be legitimate, there must be a nexus between the speech and actual harm (not merely perceived harm) and, as such, "no one is entitled to be insulated from opinions and ideas that they do not like even if those ideas are expressed in ways that place them in fear…."[50]

The SCA maintained the high threshold to hate speech after *Qwelane.* In December 2018, it ruled on *Masuku and Another v South African Human Rights Commission obo South African Jewish Board of Deputies* (2018). The cases involved statements made by Masuku, the secretary of the International Relations arm of the Congress of South African Trade Unions. In the framework of the Israel-Palestine conflict, Masuku made statements such as:

> "Let us bombard the COSATU offices with phone calls to let them know our anger. It is hard[er] to ignore phone calls than email.

47 Qwelane v South African Human Rights Commission and Another, Case 686/2018, 29 November 2019, para. 10.
48 Qwelane, para. 70.
49 Qwelane para. 84.
50 Moyo v Minister of Justice and Constitutional Development and Others; Sonti v Minister of Justice and Correctional Services and Others, Cases 287/2017; 286/2017, para. 31.

> Maybe we should start a policy that Israel-loyal Jews refuse to employ COSATU members in retaliation to COSATU's evil actions."

Again, the Court highlighted that speech may be "hurtful of people's feelings or wounding, distasteful, politically inflammatory or downright offensive [but this] does not exclude it from protection."[51]

The above approach to the free speech – hate speech debate marks a stark contrast to the ECtHR´s position on homophobic speech as set out in *Vejdeland and Others v Sweden*, where merely prejudicial allegations were sufficient to constitute hate speech that a State could prohibit without violating article 10. This position was also adopted in a 2020 ECtHR case, *Lilliendahl v Iceland*, which involved homophobic and transphobic speech. Here, the Court reiterated its position in *Vejdeland,* nothing that speech which is "prejudicial" can also constitute hate speech.[52] As such, the Court "[saw] no reason to disagree with the Supreme Court's assessment that the applicant's comments were 'serious, severely hurtful and prejudicial.'"[53]

It must be noted that the South African Human Rights Commission appealed the SCA's judgement at the Constitutional Court. In July 2021, the Constitutional Court[54] found that only the inclusion of the term "hurtful" was unconstitutional, but that the elements of hate and harm were constitutional. As such, it ruled that Qwelane's article constituted hate speech in line with the other elements of Section 10(1) of the Equality Act (hateful and harmful speech). Nevertheless, it did underline that "hate speech travels beyond mere offensive expression and can be understood as extreme detestation and vilification which risks provoking discriminatory activities against that group", accordingly while the Constitutional Court appears to have modified the very speech protective direction of South African courts vis-a-vis the prohibition of hate speech, the current threshold is still significantly more speech protective than what follows under the EctHR, and arguably more in line with what follow under ICCPR Articles 19 and 20(2).

---

51 Masuku and Another v South African Human Rights Commission obo South African Jewish Board of Deputies, Case 1062/2017, 4 December 2018 , para. 31.

52 Lilliendahl v Iceland, 2020, para. 36.

53 Lilliendahl, para. 39.

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

Conclusion

The author argues that the South African approach, which emanates from recent experience with systemic speech repression is nuanced and substantiated, providing a more rigorous and convincing balancing between freedom of expression and hate speech. The highest courts of South Africa have found that speech which is merely "offensive" or "prejudicial" is protected, whereas such speech, has often fallen afoul of the ECtHR. The same is true of statements denying historical crimes such as the Holocaust. Whilst the protection of expression was impacted by the final decision in this case, the authors argue that South Africa continues to constitute a good example of a substantiated approach to hate speech. This jurisdiction marks a significant contrast to the ECtHR´s approach to hate speech, which is steeped in the doctrine of "militant democracy" according to which statements that allegedly undermine (essentially undefined) democratic values are undeserving of protection.

On global social media platforms with users from all continents and cultures with widely diverging and clashing conceptions of where free speech ends and hate speech begins, a robust, narrow and harm-based definition of hate speech is more likely to be operational than one which includes deeply subjective notions of "offense" and "prejudice". Accordingly, stakeholders such as private companies, states and international organizations could look at the judgments of the highest courts of South Africa as a guide to re-considering current approaches to the treatment of online hate speech. Moreover, this case-law is an effective ambit through which stakeholders can align content moderation requirements with the thresholds set out by IHRL and particularly article 20(2) with the HRC deciding cases such as *Rabbae* and holding that speech extends even to 'deeply offensive' speech. In brief, South African jurisprudence (from its highest courts) provides for a substantiated approach to hate speech, preventing over-restriction (for example allowing prejudicial speech) whilst placing analysis in the realm of real-life experience (its own apartheid).

The current digital era is marked by increasing pressure on social media platforms to quickly remove content such as "hate speech." The obligation to remove such a contested and poorly defined area of speech within short time spans on global platforms is ill suited to offer the necessary safeguards for freedom of expression. This approach is more likely to initiate a global censorship race to the bottom, than act as a bulwark of liberty, and indeed such a development already seems to be well under way. While no quick fix is likely to resolve this situation, IHRL, provides the best, or least bad, "framework of first reference" for both states and major platforms when

it comes to determining the relationship between protected expression and impermissible hate speech. In particular, articles 19 and 20 of the ICCPR offers a promising way ahead, which offers a more robust and speech protective way forward than the incoherent case law of the ECtHR. Yet, given the paucity of legally binding cases relating to the ICCPR, South African case law on the relationship between freedom of expression and hate speech offers a compelling interpretational aide when further defining the relationship between article 19 and 20.

Bibliography

Aswad, Evelyn Mary. "The Future of Freedom of Expression Online." *Duke Law & Technology Review* 17, no.1 (2018): 26-70.

Benesch, Susan. "But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies." *Yale Journal on Regulation Online Bulletin 39*, no.3 (2020): 86-111.

Dvoskin, Brenda. "International Human Rights Law is not Enough to Fix Content Moderation's Legitimacy Crisis." *Berkman Klein Center for Internet & Society at Harvard University*, September 16, 2020. https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd.

Lessig, Lawrence. *Code: version 2.0.* New York: Basic Books, 2006.

Mchangama, Jacob et al. "Rushing to Judgment: Are Short Mandatory Takedown Limits for Online Hate Speech Compatible with the Freedom of Expression." *Justitia*, January 2021.

Singh, Karan Deep and Mozur, Paul. "As Outbreak Rages, India Orders Critical Social Media Posts to be Taken Down." *New York Times*, April 25, 2021. https://www.nytimes.com/2021/04/25/business/india-covid19-twitter-facebook.html.