Response to: Call for Input to the High Commissioner Report on the Practical Application of the United Nations Guiding Principles on Business and Human Rights to Activities of Technology Companies

21 February 2022

Justitia

Justitia is Denmark's first judicial think-tank. Justitia aims to promote the rule of law, human rights and fundamental freedoms, both within Denmark and abroad, by educating and influencing policy experts, decision-makers, and the public.

Future of Free Speech Project

The Future of Free Speech is a collaboration between Justitia, Columbia University's Global Freedom of Expression and Aarhus University's Department of Political Science. We believe that a robust and resilient culture of free speech must be the foundation for the future of any free, democratic society. Even as rapid technological change brings new challenges and threats, free speech must continue to serve as an essential ideal and a fundamental right for all people.

Introduction

Technology companies offer people around the world endless possibilities to share and impart information and express their opinions and beliefs quickly and without cost on social media platforms (SMPs). There are 4.62 billion users (58.4% of the global population) spending an average of 2h 27m per day on social media. This has revolutionized human communication and freedom of expression and has allowed people to bypass gatekeepers of information streams and overcome traditional censorship. However, this 'platformization'of communication is not without its issues as it has provided a vehicle for phenomena, such as extremism, terrorist content, disinformation, and hate speech, to gain wide and rapid dissemination. Our input to this consultation looks particularly at theme 4 – the State's duty to protect regulatory and policy responses. In 2021, Justitia published a report which sets out International Human Rights Law (IHRL) as a 'framework of first reference' for moderating online hate speech and disinformation. It decodes relevant IHRL principles, applies them to

hate speech and disinformation, and offers recommendations on their adoption by SMPs. As the consultation requests the sharing of views on the practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, we would be pleased if the OHCHR could refer to this report in addition to our consultation input.

Regulatory Practices

Encroaching role of States vis-à-vis the functioning of social media platforms.

We are witnessing a global trend of States increasing regulatory pressures on internet intermediaries, such as SMPs, by imposing upon them the duty to remove content quickly which is broadly defined as, for example, hate speech or disinformation. For instance, the 2017 German Network Enforcement Act (NetzDG) imposes a legal obligation on platforms to remove content, such as insult, incitement, and religious defamation, within short time limits of 24 hours for 'manifestly illegal' content or risk a fine of up to 50 million EUR.As demonstrated in two of Justitia's reports in 2019 and 2020, over 20 countries have adopted or propose to adopt laws similar to the NetzDG. As such, despite Germany´s supposed efforts of good faith in adopting a tool to ensure legitimate platform regulation based on democratic methods, its pioneering efforts have legitimized and created a prototype for much more speech restrictive measures by authoritarian and semi-authoritarian regimes. It could be argued that, in such States, SMPs are, in fact, the only voice for minorities and those silenced, rendering the NetzDG template even more dangerous and vulnerable to exploitation for purposes of, among others, silencing critics, marginalizing minority groups and controlling the spectrum on social media.

The implications of free speech for intermediary liability laws were addressed by the French Constitutional Council in 2020, when it deliberated on the legitimacy of the Avia Law. The Council found the provisions which required platforms to remove hate speech within 24 hours or face large fines unconstitutional, as they were disproportionately harmful to the freedom of expression.

The problems inherent in *short time removal periods* have been highlighted in a 2021 report, issued by Justitia, which compares the duration of national legal proceedings in hate speech

cases in Denmark, Germany, the United Kingdom, France, and Austria. Overall, data extracted from all European Court of Human Rights' hate speech cases, pertaining to the five countries, reveals that domestic legal authorities took 778.47 days, on average, from the date of the alleged offending speech until the conclusion of the trial at first instance, to determine the unlawfulness of certain online content. Against the backdrop, putting platforms at risk of hefty fines to evaluate flagged content in a matter of hours poses a strong risk of over-removal of online speech.

Rising Removal Rate of Allegedly Hateful Content

i. The Harmsof Hate Speech

It cannot be disputed that certain types of hate speech which, for example, call for violence, should be removed. In the aforementioned 'Framework of First Reference' report, we dissect the threshold set out by Article 20(2) of the International Covenant on Civil and Political Rights and provide practical guidance for SMPs to incorporate this into their regulatory practices. Studies have demonstrated that online hate speech can result in fear, trauma and self-censorship, predominantly amongst minorities. We acknowledge that even speech which does not directly call for violence may lead to other harmful effects on individuals and groups. This does not, by default, imply that curtailing freedom of expression is an appropriate answer.[1]Further, we must be wary of the fact that such curtailment may also lead to other woes. In 2017, Ravndal found that the rise of far-right extremism in Western Europe is fuelled by a mix of factors including high immigration rates, low electoral support for radical right political parties and the 'extensive public repression of radical right actors and opinions.' Whilst noting that such repression may, in fact, discourage people from joining extreme groups, it may also prompt people to more violent pathways. Further, by de-platforming extremists from mainstream social media for violating terms is not the end of the road for them. Many simply

[1] As discussed by scholars such as Eric Heinze, "Hate Speech and the Normative Foundations of Regulation'" *International Journal of Law in Context 9* No.4 (2013) 599; Eric Bleich in "Hate Crime Policy in Western Europe: Responding to Racist Violence in Britain, Germany, and France" *American Behavioral Scientist 51* No 2, (2007) 149–165 and Joost Van Spanje and Woost Van Der Brug, "The Party as Pariah: The Exclusion of Anti-Immigration Parties and its Effect on their Ideological Positions" (2007) *Western European Politics 30,* No5 (2007) 1022-1040. For an overview of several positions regarding harms of repression see Jacob Mchangama, "How Censorship Crosses Borders" *Cato: A Journal of Debate* (2018) <https://www.cato-unbound.org/2018/06/11/jacob-mchangama/how-censorship-crosses-borders>

migrate to other platforms which use encrypted messaging services like Telegram.[2]This may subsequently render law enforcement tasks more complex but also reduces prospects of counter-speech, which has been proved effective as a response to hate speech.

ii. The Prevalence and Illegality of Online Hate Speech

Moreover, when dealing with the future of moderating online hate speech, it is imperative to take a step back and look at its prevalence. There has been a certain narrative put forth by civil society, academia and States themselves on the very high alleged prevalence of hate speech online. However, the empirical backdrop of this position is rather weak. A 2019 study, led by Siegel *et al,* looked at whether Trump's 2016 election campaign and its immediate aftermath (6 months) contributed to the rise in hate speech or white nationalist language. The study analyzed 1.2 billion tweets, 750 million of which were election-related, and nearly 400 million were random samples. On any given day, between 0.001% and 0.003% of the tweets contained hate speech. A study on Ethiopia demonstrated a similarly low prevalence of hate speech on Facebook.

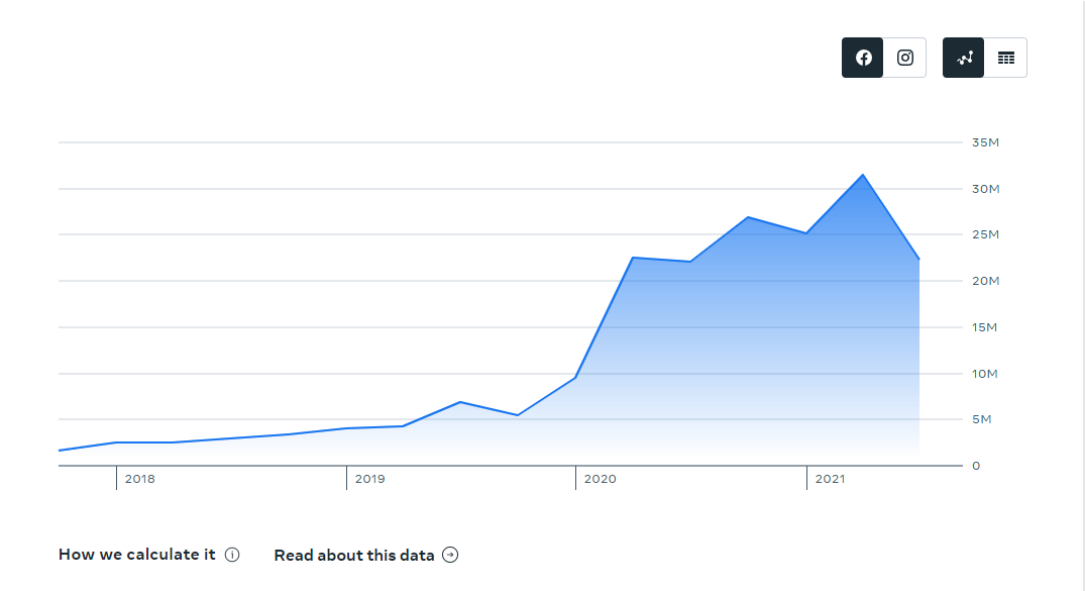
In a 2022 report, published by Justitia, which investigated the extent of illegal comments on 676 Danish Facebook pages through a representative segment of 63 million comments, it was determined that only 0.0066% of these violated provisions of the Danish Criminal Code, including threats, deriding and demeaning remarks against certain groups, incitement to criminal action and the express support of terror.

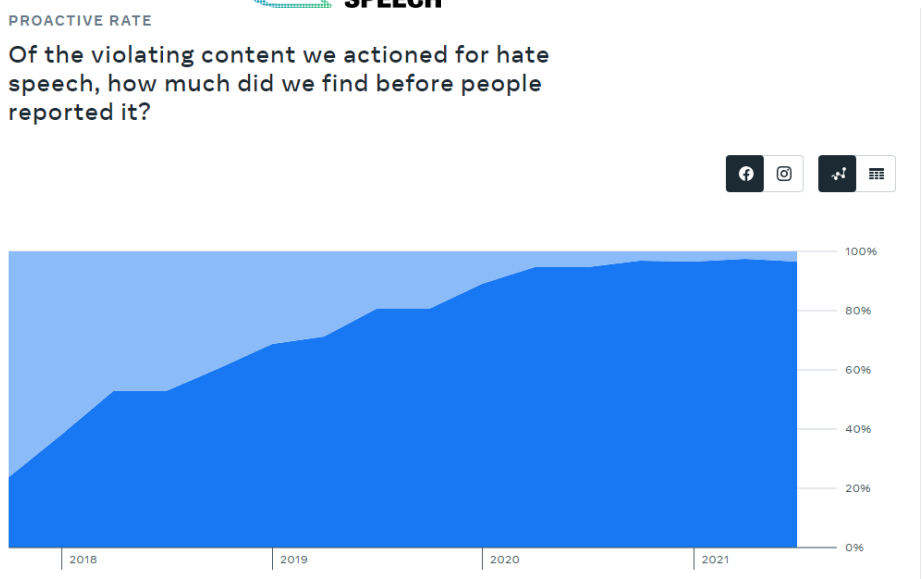
iii. State Pressure, Increased Removal Rates and Artificial Intelligence

A profit-driven private company, depending on good relations with governments in the countries in which it operates, will thus be incentivized to adopt a better safe than sorry approach. The German government relies on statistics which shows that the NetzDG has not resulted in dramatic purging or "over-blocking" of content in Germany. However, the removal rates under the NetzDG regime cannot be viewed in isolation, since most of the content deleted by social media platforms is removed pursuant to the platforms' Terms of Service/Community

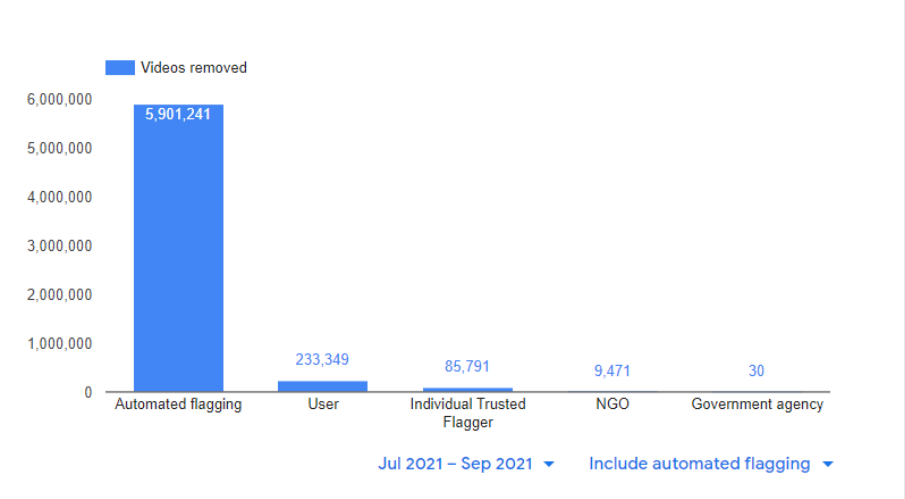
[2] Aleksandra Urman and Stefan Katz "What They Do in the Shadows: Examining the Far-Right Networks on Telegram, *Information, Communication & Society* (2020) DOI: 10.1080/1369118X.2020.1803946

Standards, rather than national laws. The possibility of developments, such as the NetzDG, 'indirectly' affecting the moderation practices of private companies must also be considered. We have seen a'drastic increase in content removal over the last few years.' For example, the first graph (as obtained from Facebook's Community Standards Enforcement Report) demonstrates the massive rise in the removal of the broadly conceptualized notion of hate speech between 2018 and 2021. This depicts the current *status quo* whereby private social media companies, not bound by IHRL, have become the 'ultimate arbiters of harm, truth and the practical limits of the fundamental rights to freedom of expression.' The second graph demonstrates a respective rise in relation to proactive removal due to advances in the use of Artificial Intelligence (AI).





The increased use of AI for content removal can also be seen with **YouTube:**



To be able to comply with such standards (and avoid hefty fines), companies use AI alone, or in conjunction with human moderation, to remove allegedly hateful content. As noted by Oliva, such circumstances have prompted companies to "act proactively in order to avoid liability…in an attempt to protect their business models". Gorwa, Binns and Katzenbach highlight that as "government pressure on major technology companies build, both firms and legislators are searching for technical solutions to difficult platform governance puzzles such as hate speech and misinformation".

AI provides SMPs with 'tools to police an enormous and ever-increasing flow of information.'[3] Whilst this is necessary in areas involving, for example, child abuse, the use of AI to regulate more contentious 'grey' areas of speech, such as hate speech, is complex and allows for the possibility of sweeping over removals. Technology handling content, such as hate speech, is still in its "infancy". The results of enhanced moderation of contentious areas of speech, such as 'hate speech' and the use of AI, have led to a fall in media diversity. For example, YouTube removed 6,000 videos documenting the Syrian conflict and shut down the Qasioun News Agency, an independent media group reporting on war crimes in Syria. Several videos were flagged as inappropriate by an automatic system designed to identify extremist content. Other hash matching technologies, such as PhotoDNA, also seem to operate in 'context blindness' which could be the reason for the removal of those videos. In sum, as also noted by the OSCE Representative on Freedom of the Media, the use of AI could seriously jeopardize our human rights, in particular the freedom of expression. Further, there is an effect on non-discrimination too. For example, the Centre for Democracy and Technology revealed that automated mechanisms may disproportionately impact the speech of marginalized groups. Although technologies, such as natural language processing and sentiment analysis, have been developed to detect harmful text without having to rely on specific words/phrases, research has shown that they are "still far from being able to grasp context or to detect the intent or motivation of the speaker". Such technologies are just not cut out to pick up on the language used, for example, by the LGBTQ community whose "mock impoliteness" and use of terms such as "dyke," "fag" and "tranny" occurs as a form of reclamation of power and a means of preparing members of this community to "cope with hostility".

iv. Change of Platform Policies on Misinformation/Disinformation

The pressure to act decisively on misinformation, and the resultant calls to restrict more content, may be better understood in the context of "elite panic", which sociologists Lee Clarke and Caron Chess explained as a phenomenon resulting when decision-makers are under intense media scrutiny to act decisively. They argue that, in such situations of crisis, social elites might

[3] Thiago Dias Oliva, D. Antonialli, A. Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risk to LGBTQ Voices Online?' (2020) *Sexuality and Culture* 701

make rash decisions that could potentially make things worse than the very real problems against which these actions were aimed.

In this light, we have witnessed the widespread removal of misinformation/disinformation. Platforms such as Facebook and Instagram generally downrank such content. However, following the onset of the pandemic, there has been an increasing trend towards removal. For example, in 2020, Facebookand Instagram introduced detailed sections on COVID-19 misinformation that allowed for content removal for information that could cause harm.We argue that, in relation to disinformation, platforms' terms and conditions should be tailored to protect the grounds in Article 19(3) ICCPR and Article 25 ICCPR (right to participate in voting and elections). In addition, platforms must refrain from adopting vague blanket policies for removal. Only disinformation promoting real and immediate harm should be subject to the most intrusive restrictive measures such as content removal. In determining the limits of disinformation, platforms should focus on the post's content, its context, its impact, its likelihood of causing imminent harm, and the speaker's intent.

Conclusion and Recommendations

If major SMPs are forced into purging lawful but awful content, they can become digital chokepoints. Potentially, centralized platforms could even end up serving as private enforcers of government censorship, entirely inverting the initial promise of egalitarian and unmediated free speech through the advent of the internet. Thus, we recommend that:

- Content moderation of contentious areas of speech, namely hate speech and disinformation, occur within the framework of International Human Rights Law, with removal of such content being legitimate, necessary and proportional and in line with Article 19 ICCPR.

- Only disinformation entailing real and immediate harm should be subject to removal. For other categories of disinformation, platforms may resort to less restrictive forms of moderation, such as downranking content, flagging content and providing users with access to reliable/official sources of medical information, among others.

- AI based content filters should not be used without human monitoring as they may be susceptible to biased data sets and unable to pick up on the nuance of language.

- A voluntary pledge where platforms adopt a human rights standard for disinformation and hate speech ensuring more transparency and consistency.

- The creation of a free speech framework agreement administered under the auspices of the United Nations in order to ensure compliance with the voluntary pledge.

We remain at your disposal for any clarification/further information you may require.

Yours Sincerely,

Jacob Mchangama
Executive Director, Justitia