For the Attention of The Oversight Board

September 20, 2024

The Future of Free Speech at Vanderbilt University

The Future of Free Speech (FoFS)[1] is an independent, nonpartisan think tank located at Vanderbilt University. We work to restore a resilient global culture of free speech in the digital age through knowledge, research, and advocacy.

FoFS' comment focuses on the following issues identified by the Oversight Board:

The impacts of Meta's Hate Speech and Bullying and Harassment policies on freedom of expression around gender identity issues, and the rights of transgender people, including minors.

In light of the above question, we have sought to put forth ideas/good practices/material/arguments for purposes of assessing the impact of the above looking at issues such as (i) International Human Rights Law (ii) Best practices from national courts and their relevance in the development and application of these policies in the ambit of gender identity and rights of transgender purposes. It looks at broader issues that need to be taken into account when measuring the impact of such policies in the ambit of gender identity issues and the rights of transgender people such as (i) the impact of removal of controversial trans-related issues on public debate and social cohesion (ii) hate speech removal and free speech ramifications and (iii) automated content moderation and its effect on the LGBTQ+ community

Key Takeaways:

- The FoFS agrees with Meta's decision that neither case violated its Hate Speech and Bullying and Harassment policies. The FoFS holds that Meta's decision is in line with International Human Rights Law.

[1] https://futurefreespeech.org/

- The FoFS proposes that Meta and the Oversight Board look to IHRL and specifically Article 20(2) of the International Covenant on Civil and Political Rights and the Rabat Plan of Action as well as good practices from national courts such as higher courts of South Africa to guide the assessment of community standards around gender identity issues and the rights of transgender people, including minors.

- The FoFS proposes that Meta and the Oversight Board take into consideration the paramount significance of open public debate when it comes to current controversial issues and the social and democratic impact of silencing such debates in whole or in part.

- The FoFS suggest that Meta considers aligning content moderation documents of Instagram and Facebook to ensure certainty and legitimacy.

- The FoFS recommends that META considers alternative methods to deal with hate speech that may not meet international thresholds of prohibition. These include distributed content moderation and counterspeech.

Case Description:

These two cases concern content decisions made by Meta, on Facebook and Instagram, which the Oversight Board intends to address together.

In the first case, a Facebook user in the United States posted a video of a woman confronting a transgender woman for using the women's bathroom. The post refers to the person being confronted as a man and asks why it is permitted for them to use a women's bathroom.

In the second case, an Instagram account posted a video of a transgender girl winning a female sports competition in the United States, with some spectators vocally disapproving of the result. The post refers to the athlete as a boy, questioning whether they are female.

International Human Rights Law

In 2018, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (SRFOE) stated that the "scale and complexity of addressing hateful expression presents long-term challenges and may lead companies to restrict such expression even if it is not clearly linked to adverse outcomes (as hateful advocacy is connected to incitement in Article 20(2) of the ICCPR)."[2] However, the SRFOE clarified that "companies should articulate the bases for such restrictions, however, and demonstrate the necessity and proportionality of any content actions."[3] Importantly, in 2019, the SRFOE added that "when company rules differ from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation."[4] In this light, and beyond the internal rules of Facebook, the Oversight Board must take into account the present case in light of Article 19 (the right to freedom of expression) and Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR) as well as the threshold test set out by the Rabat Plan of Action (RPA). Article 19 provides for the right to freedom of expression, including the right to access information. Article 20 (2) provides that 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.'

The RPA which guides the interpretation and application of Article 20(2) provides that restrictions to free speech should be "the least intrusive measure available; are not overly broad,

[2] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) A/HRC/38/35
<https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf> p.11
[3] Ibid p. 3
[4] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2019) A/74/486
< https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/A_74_486.pdf> p. 19

so that they do not restrict speech in a wide or untargeted way."[5] The RPA includes a six-part threshold test to be used when applying Article 20(2), which incorporates a consideration of:

1. The social and political context
2. Status of the speaker
3. Intent to incite the audience against a target group
4. Content and form of the speech
5. Extent of its dissemination
6. Likelihood of harm, including imminence

RPA "factors should have weight in the context of company actions against speech," as "they offer a valuable framework for examining when the specifically defined content – the posts or the words or images that comprise the post – merits a restriction."[6] As noted by the SRFOE, restrictions to freedom of expression must be "narrowly defined."

The issue of threshold is also vital when assessing the question posed by the Oversight Board. On this point, some findings from a 2019 report of the Special Rapporteur on the Freedom of Expression are relevant. For example, he noted that merely offensive speech that does not incite violence does not fall within the spectrum of International Human Rights Law. He underlined that:

"it is important to emphasize that expression that may be offensive or characterized by prejudice and that may raise serious concerns of intolerance may often not meet a threshold of severity to merit any kind of restriction. There is a range of expression of hatred, ugly as it is, that does not involve incitement or direct threat, such as declarations of prejudice against protected groups." [7]

[5] The Rabat Plan of Action <https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action> p.9
[6] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
< https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/A_74_486.pdf> A/74/486, p.5
[7] Para 10 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (2019) A/74/486, para.10

He proceeded in stating that a threshold of severity must be met if restrictions will be imposed, underlining that:

"a person who is not advocating hatred that constitutes incitement to discrimination, hostility or violence, for example, a person advocating a minority or even offensive interpretation of a religious tenet or historical event, or a person sharing examples of hatred and incitement to report on or raise awareness of the issue, is not to be silenced under article 20 (or any other provision of human rights law)."[8]

Under Article 10 of the European Convention on Human Rights (ECHR), the right to freedom of expression is protected, subject to certain limitations necessary in a democratic society. These limitations include restrictions for the protection of the rights of others, such as protection against hate speech or incitement to violence. However, the European Court of Human Rights (ECtHR) has consistently held that restrictions on freedom of expression must be narrowly construed and convincingly established. In *Handyside v. United Kingdom*, the Court underscored that freedom of expression applies not only to "information" or "ideas" that are favourably received or regarded as inoffensive but also to those that offend, shock, or disturb.[9] This principle is particularly relevant in cases involving public debate on sensitive issues.

Application to the two videos: While the videos may be offensive to some viewers, it does not meet the threshold of hate speech under Article 20(2) ICCPR as interpreted by the RPA and under International Human Rights Law more broadly. The videos, while confrontational (case 1) and controversial (case 2) appear to express personal opinions rather than incitement to violence or hatred. As such, they fall within the ambit of protected speech under international standards. A broad interpretation of hate speech and a removal of videos such as the ones in

[8] Ibid. para.24
[9] Handyside v. United Kingdom, App No 5493/72 (ECtHR, 7 December 1976)

this case essentially serves to mute public debate on a contemporary and controversial issue and does little to promote a healthy and equal marketplace of ideas.

National Courts – Best Practices:

The FoFS suggests that META and the Oversight Board take inspiration from best practices on matters pertaining to the handling of hate speech as identified in national jurisprudence. While these do not create standards to follow, they nevertheless provide insight into relevant issues under the current discussion. Examples listed below:

United States: The U.S. Supreme Court in *Snyder v. Phelps* which involved homophobic speech upheld the right to express even deeply offensive opinions in public discourse, noting that such expression is vital to democratic debate.[10] Specifically, Chief Justice Roberts noted that:

"Speech is powerful. It can stir people to action, move them to tears of both joy and sorrow and as it did here inflict great pain. On the facts before us, we cannot react to that pain by punishing the speaker. As a nation we have chosen a different course – to protect even hurtful speech on public issues to ensure that we do not stifle public debate."

South African courts offer valuable insights into how freedom of expression and limitations on hate speech can be balanced effectively, minimising the risk of misuse. These courts explicitly refer to International Human Rights Law in general and the International Covenant on Civil and Political Rights (ICCPR) in particular.

In *Islamic Unity Convention v Independent Broadcasting Authority and Others* (2002), the Constitutional Court found that an administrative clause prohibiting prejudice (and used against the impugned speech) did not meet the constitutional threshold. The Court underlined that "individuals in our society need to be able to hear, form and express opinions and views freely on a wide range of matters…." The Court also underlined that "not every expression of

[10] Snyder v. Phelps, 562 U.S. 443 (2011)

speech that is likely to prejudice relations between sections of the population would constitute 'advocacy of hatred' which also constitutes 'incitement to cause harm." [11]

In *Qwelane v South African Human Rights Commission (2021)* which involved homophobic speech, the Constitutional Court of South Africa ruled that "if speech that is merely hurtful is considered hate speech, this sets the bar rather low."[12] This case is a significant point of reference for the current discussion. The facts are as follows. The Sunday Sun newspaper published a column by Jon Qwelane titled "Call me names – but gay is NOT okay…" In the article, Qwelane criticized what he described as permissive contemporary attitudes towards homosexual relationships between men, referring to them as part of the "rapid degradation of values and traditions by the so-called liberal influences of nowadays." He urged politicians to "muster the balls to rewrite the Constitution of this country, to excise those sections which give licence to men 'marrying' other men, and ditto women." The article was accompanied by a cartoon that compared homosexuality to bestiality. The Supreme Court of Appeal found that there was no hate speech and that any restrictions to Qwelane and his speech were not legitimate. The Court mentioned, amongst others, that "one must be careful not to stifle the views of those who speak out of genuine conviction." [13] This case was subsequently heard by the country's constitutional court which discussed the constitutionality of the relevant legislative provision used against Qwelane. The Court ruled that hurtful speech is not hate speech, noting that "if speech that is merely hurtful is considered hate speech, this sets the bar rather low."[14]

So, when assessing the current question of how Meta's Hate Speech and Bullying and Harassment policies affect freedom of expression, it is central to identify the actual threshold of harm that we are faced with, with low thresholds equating to stifling of the fundamental right to the freedom of expression.

[11] Islamic Unity Convention v Independent Broadcasting Authority and Others (CCT36/01) [2002]
[12] Qwelane v South African Human Rights Commission and Another (CCT 13/20) [2021]
[13] Qwelane v South African Human Rights Commission and Another (686/2018) [2019] para.70
[14] Qwelane v South African Human Rights Commission and Another (CCT 13/20) [2021] para.139

Freedom of Speech, Public Debate and Social Cohesion

While any form of hate speech is deplorable for victims and society more broadly, 'it does not necessarily follow that restriction on free speech is an effective remedy."[15] Overly restrictive application of Meta's policies can create a chilling effect, where individuals may refrain from participating in discussions on gender identity for fear of their views being labelled as hate speech or harassment. This not only stifles debate but also marginalizes voices that seek to challenge or critique prevailing norms around gender, which is essential for a vibrant democratic society. While restricting speech such as those in the two videos may seek to curb discrimination and protect trans communities, this practice may have unintended consequences that exacerbate tensions both between different minority groups and within society as a whole.

For instance, there can be significant friction between religious communities that uphold traditional beliefs and trans rights advocates. This is particularly evident when religious groups voice their opposition to certain aspects of LGBTQ+ rights, such as the recognition of non-binary genders or the use of gender-neutral language. When such opposition is categorized as hate speech, it can lead to perceptions of censorship and marginalization among religious communities, who may feel that their right to express their beliefs is being curtailed. This can foster a sense of victimization and alienation within these groups, potentially radicalizing their views and heightening inter-group conflict rather than fostering mutual understanding and respect. Moreover, removing the videos could unintentionally lend credibility to the argument that certain views are being censored, potentially turning the speakers into perceived "martyrs" of a biased system. This could undermine public trust in social media platforms as facilitators of free expression and fuel further polarization.

Relevant to this point are several examples of judicial resistance to the restriction of homophobic/transphobic speech. Examples include Finland and specifically the case of Räsänen, who previously served as a government minister and is a current member of parliament. She faces criminal charges for expressing her faith-based views on marriage and

[15] Jacob Mchangama & Natalie Alkiviadou, 'Hate Speech and the European Court of Human Rights: Whatever happened to the Right to Offend, Shock or Disturb?' (2021) 21 *Humam Rights Law Review,* 1018

sexual ethics in a 2019 tweet and a 2004 pamphlet she authored for her church, which focused on the Biblical text, "male and female he created them." The case against her has been rejected by two courts already. In April 2024, the Supreme Court of Finland confirmed that Räsänen will stand trial for the third time over her Bible-verse tweet. Despite being unanimously acquitted of "hate speech" charges by both the Helsinki District Court and the Court of Appeal, the State Prosecutor appealed the case. The Supreme Court has agreed to hear the appeal, with the trial date yet to be determined.[16]

The judiciary in England and Wales has also shown reluctance to restrict this type of speech. In 2020, the High Court ruled that the police's response to allegedly transphobic tweets was unlawful. Miller, the defendant and ex-officer was visited by the police at his workplace following a complaint about tweets he posted between November 2018 and January 2019, discussing transgender issues as part of the debate on reforming the Gender Recognition Act 2004.

In one tweet, Miller wrote: "I was assigned mammal at birth, but my orientation is fish. Don't mis-species me."

Mr. Justice Julian Knowles emphasized the impact of the police visiting Miller at work "because of his political opinions." He stated, "To do so would be to undervalue a cardinal democratic freedom." He further remarked, "In this country, we have never had a Cheka, a Gestapo, or a Stasi. We have never lived in an Orwellian society."[17]

Turning to Scotland, J K Rowling has regularly misgendered trans women. Reacting to the 2021 Scottish Hate Crime Bill which criminalizes, amongst others "stirring up hatred" related to protected characteristics including transgender identity. The Police Force issued a statement

[16] Bible-tweet case to be heard at Finnish Supreme Court' (April 19 2024) *ADF International* <https://adfinternational.org/news/bible-tweet-case-to-be-heard-at-finnish-supreme-court>
[17] 'Harry Miller: Police probe into 'transphobic' tweets unlawful' (February 14, 2024) *BBC News* <https://www.bbc.com/news/uk-england-lincolnshire-51501202>

that complaints were received about Rowling's statements, but no action would be taken against her. [18]

In light of the above-described resistance, it seems out of place and disproportionate to remove videos involving misgendering/discussion of use of public toilets and participation of trans women in female sports. To add to this position, particularly in relation to the video involving sports activity, it must be underlined that several sports associations have banned the participation of trans women in female sports. Examples include World Athletics, World Aquatics, World Rugby, The International Rugby League, FINA (swimming) and the National Association of Intercollegiate Athletics (NAIA).[19]

Further, the chilling effect of restricting speech such as misgendering or debates on transgender persons and sport does not only impact those with critical views but also transgender individuals themselves, who may hesitate to share their experiences and perspectives in a hostile or overly regulated environment. The suppression of discourse around gender identity can have a detrimental impact on transgender individuals, including minors. Transgender youth, who are already vulnerable to discrimination and social exclusion, may find it even harder to access supportive communities and resources if discussions around their identities are suppressed. This could exacerbate feelings of isolation and invisibility. Conversely, providing a platform for informed debate can help raise awareness and foster understanding of transgender issues. As such, restricting free speech does not help minorities. Recent studies

[18] Megan Bonar & Katy Scott, 'JK Rowling hate law posts not criminal, police say.' (April 3, 2024) <https://www.bbc.com/news/uk-scotland-68712471>
[19] NAIA bans transgender women from competing in women's sports' (April 9, 2024) *NBC News* <https://www.nbcnews.com/nbc-out/out-news/naia-bans-transgender-women-competing-womens-sports-rcna147017> World Athletics < https://worldathletics.org/news/press-releases/council-meeting-march-2023-russia-belarus-female-eligibility> ; Ben Church, 'World aquatics launches open category for transgender athletes at swimming world cup' (August 17, 2013) < https://edition.cnn.com/2023/08/17/sport/world-aquatics-transgender-athletes-swimming-spt-intl/index.html>; World Rugby <https://www.world.rugby/the-game/player-welfare/guidelines/transgender> 'Transgender women banned from women's international rugby league' (2022) < https://www.bbc.com/sport/rugby-league/61875651#:~:text=Transgender%20women%20have%20been%20banned,perceived%20risk%20to%20other%20participants%22.>; Ciaran Fahey, 'World Swimming Bans Transgender Athletes from Women's Events' (2022) < https://apnews.com/article/transgender-swimmers-new-rules-fina-world-governing-body-c17e99d3121fa964336458b57ae266f7>

have demonstrated that members of the LGBTQ+ community frequently encounter silencing on social media platforms. A recent GLAAD study highlights specific instances in which content promoting or discussing LGBTQ+ issues is disproportionately flagged and removed, compared to non-LGBTQ+ content.[20] This inconsistency suggests a bias in content moderation, previously documented for different markets by the Electronic Frontier Foundation, that contributes to the marginalization and erasure of LGBTQ+ voices in online spaces.[21] Furthermore, as several researchers have pointed out, the use of AI for moderating content such as hate speech poses severe challenges to the right to freedom of expression and access to information online.[22] This is particularly so in relation to the free speech of minorities such as the LGBTQ+ community. The use of AI may lead to biased enforcement of removal obligations as a result of insufficient data and biased datasets which disproportionately silence members of minority communities raising issues of violations in relation to free speech but also the right to non-discrimination.[23]

In brief, by allowing diverse perspectives, Meta can contribute to a more informed public dialogue, which is crucial for the recognition and protection of the rights of transgender people. Moreover, a restrictive free speech framework has led to mistakes and abuses of the rights of members of minority groups as can be seen in the above studies but also to trends across Central and Eastern Europe in particular where "LGBTQ propaganda" is banned under the guise of maintaining "traditional" or "family values." [24]

---

[20] Gladd: 'LGBTQ Contnet Suppression Case Study: Men Having Babies (Instagram) 2024 Social Media Index Safety < https://glaad.org/smsi/2024/lgbtq-content-suppression-case-study/>

[21] Electronic Frontier Foundation, 'Blunt Policies and Secretive Enforcement Mechanism: LGBTQ+ and Sexual Health on the Corporate Web. <https://www.eff.org/deeplinks/2018/10/blunt-policies-and-secretive-enforcement-mechanisms-lgbtq-and-sexual-health>

[22] See, amongst others, Emma Llanso *et al*., "Artificial Intelligence, Content Moderation and Freedom of Expression." Transatlantic Working Group, 2020, accessed November 23, 2022, https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf; Thiago Oliva Dias, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression," *Human Rights Law Review* 20, no. 4 (2020): 607-640.

[23] See, amongst others, Natasha Duarte and Emma J. Llansó, "Mixed Messages? The Limits of Automated Social Media Content Analysis." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81 (2018): 106-106; Thiago Oliva Dias *et al.*, "Fighting Hate Speech, Silencing Drag Queens?," (2021).

[24] Francesco Bortoletto, "Anti-LGBTQ+ Crackdowns in Bulgaria: Calls for Sanctions against Sofia Multiply" (August 21, 2024) *EU News* < https://www.eunews.it/en/2024/08/21/anti-lgbtq-crackdown-in-bulgaria-calls-for-

Hate Speech Policies Can Significant Limit Expression

Justitia has shown that hate speech policies can significantly limit expression.

In a 2022 report, Justitia analyzed 2,400 Facebook comments labelled as "hateful attacks"[25]– a representative sample of over 900,000 hateful attacks identified by reviewing 63 million comments on Facebook pages belonging to Danish politicians and media outlets. Justitia found that only 11 comments (0.066% of the total) could be considered illegal under Danish prohibitions on incitement and hate speech. These findings suggest that expansive interpretations of hate speech, either in the law or community standards, may lead to the mass removal of content beyond international human rights principles

Hate Speech Policies on Instagram and Facebook

In a recent report,[26] FoFS showed that since July 2020 Instagram has hyperlinks in Instagram Community Guidelines – including in the phrase "hate speech" – that direct to the Facebook Community Standards. FoFS understands that, by adding these hyperlinks in Instagram's Community Guidelines, the company implied that Facebook's Community Standards apply to content on Instagram. Meta's Transparency Center seems to confirm this: "Facebook and Instagram share content policies. Content that is considered violating on Facebook is also considered violating on Instagram." [27]As a result, it is unclear why Facebook and Instagram have two separate sets of policies. This overlap is problematic because the scope of the hate speech provision in Instagram's Community Guidelines differs from the hate speech policy in the Facebook Community Standards. Instagram's policy included 10 protected characteristics, compared to the 16 that Facebook's policy covers. This inconsistency goes against the legality

sanctions-against-sofia-multiply/>; "Russia: Expanded 'Gay Propaganda' Ban Progresses Toward Law" (November 25, 2022) < https://www.hrw.org/news/2022/11/25/russia-expanded-gay-propaganda-ban-progresses-toward-law>; "Hungary: Propaganda Law has "created cloud of fear" pushing LGBTI+ community into the shadows" < https://www.amnesty.org/en/latest/news/2024/02/hungarypropaganda-law-has-created-cloud-of-fear-pushing-lgbti-community-into-the-shadows/>

[25] "Preventing "Torrents of Hate" or Stifling Free Expression Online?" (2024) *Future of Free Speech* https://futurefreespeech.org/preventing-torrents-of-hate-or-stifling-free-expression-online/

[26] Jacob Mchangama, Abby Fanlo, and Natalie Alkiviadou, "Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies" (2023) *The Future of Free Speech* <https://futurefreespeech.org/scope-creep/>

[27] Meta, "Community Standards Enforcement Report - Q1 2023 Report," Transparency Center <https://transparency.fb.com/reports/community-standards-enforcement/>

requirement in Article 19 (3) of the ICCPR, as users cannot know which exact rules apply to Instagram.

Conclusion and Recommendations

In conclusion:

- The decision not to remove the videos is in line under IHRL.
- The videos while confrontational (video 1) and controversial (video 2) do not appear to incite hatred, violence or discrimination, thus falling within the scope of protected speech.
- Maintaining the videos on the platform supports robust public debate and avoids creating the perception of censorship, which could have counterproductive effects.

Recommendations:

- Monitoring and Contextualisation: Meta could consider adding contextual information or a disclaimer to the video, clarifying its position on the issue and the importance of respectful discourse.

- Promotion of Counterspeech: Encouraging counter-narratives and informed debate on the platform can mitigate the potential harms of the video without resorting to removal.

- Continuous Review: The platform should continuously review the video for any developments, such as new complaints or evidence of harm, ensuring that its approach remains responsive and proportionate.

- Decentralized Content Moderation as a Complement: In FoFS' view, a potential way of limiting excessive content removal, while protecting those impacted by hate speech, is relying on decentralized content moderation. Platforms could adopt a minimum set of rules applying to all the content they host. These rules could mirror ICCPR's approach or the more speech-protective U.S. Constitution First Amendment, limiting content policies' scope compared to the current situation. Then, possibly through third-party

systems, users would be able to filter out additional content to avoid speech they find problematic based on their preferences. FFS has briefly explored this possibility in a recent report and is conducting more work in this field that will be published in the coming months.

- Harmonize the policy documents of Instagram and Facebook into one single document to ensure cohesion and certainty.