Oversight Board Comment, Dia Kayyali

Summary: The Board should overturn Meta's decision on these posts because they should come down under the current language of both Meta's Hate Speech and Bullying and Harassment policies. The Board should also consider recommending that Meta change the wording of the Hate Speech policy to ensure users understand that it covers misgendering and deadnaming, and should reconsider self-reporting and public figure aspects of Meta's Bullying and Harassment policy. Finally, the Board should consider asking Meta to commission a thorough review of transphobic content on its platform.

The content in question and a word about technical challenges

As a preliminary note, the announcement for this case seems to indicate that the Board didn't consider the applicability of the prohibition on claims about gender identity for Private Minors, Private Adults, and Minor Involuntary Public Figures in Meta's Bullying and Harassment Policy. Both videos in this case make claims about gender identity, and even worse the second video targets a minor, who did not seek notoriety but is instead an "involuntary public figure."

Context is key when it comes to most hate speech, transphobic content included. The Board has taken enough cases now to know that content moderation at scale, in particular when it uses automation, faces challenges understanding context. In fact, Meta already makes a significant number of errors at the aggregate level as part of its operations. Meta's most recent Community Standards Enforcement Report shows that under its Hate Speech policy, Meta actioned 7.2 million pieces of content and 1.2 million (16%) were appealed. Out of those appeals, Meta restored 157k pieces of content, 13% of appeals, or 2% of its total removals. 2% may seem like a small number, but that's 2 out of 100 posts, and it adds up quickly. Meta does make the disclaimer that not every restored post was taken down in error, but it could better make its case by complying with the many recommendations made by both the Board and civil society regarding transparency, especially around automation.

The Board's question about technical challenges calls attention to the potential for both under- and overenforcement in relation to transphobic content. Overenforcement could occur if Meta removes legitimate political debate related to transgender people, whereas underenforcement could occur if Meta leaves up content targeting and potentially endangering transgender people on the basis of their gender identity. However, allowing content that does not only state a political opinion, or even question the science of medical transition, but additionally targets specific trans people is a far cry from "political debate," and there are no indications that Meta is over enforcing. On the other hand, there are indications of underenforcement. In its Post in Polish Targeting Trans People decision, the Board called on Meta to "improve the accuracy of its enforcement on hate speech towards the LGBTQIA+ community, either through automation or human review..."

At the end of the day, given how content moderation is done at scale, including through the widespread use of automation, if measures that try to address transphobic content lead to overenforcement, it is not likely to significantly impact political debates, whereas underenforcement has far more documented and significant, impacts on trans people.

None of this is meant to argue that moderation at scale can't be better. Perfect moderation at scale certainly is impossible, and some mistakes are unavoidable. Part of deciding how aggressively to action content, i.e. how many mistakes are acceptable, is determining how important it is that the content comes down or stays up. It is Meta's responsibility to responsibly make this calculation. Meta's current calculation is almost certainly incorrect based on the available evidence.

Context

The sociopolitical context for transgender people in the United States has changed dramatically in recent years. Although visibility of trans people has increased, both violent crime targeting trans people and anti-trans legislation have also increased. Transgender people experience incredibly high rates of violent crime compared to the general public- a 2021 study from UCLA puts it at 4 times the rate. The most recent annual Human Rights Campaign (HRC) report on the topic noted that hate crimes based on gender identity increased over 32% from 2021-22. HRC has documented 24 murders of trans people in the US as of August 2024, and in 2023 the Trans Rememembrance project documented 53 violent deaths. It should be noted that all of this research faces serious methodological difficulties that lead to undercounting. In a recent report that encompasses hate crimes based on gender identity, the United States Government Accountability Office emphasized that better measurement of measure bias-related criminal victimization on the internet [would] help DOJ identify and provide assistance to communities affected by hate. "

The legislative attack on transgender rights has been extensive. 26 states have passed bans on gender-affirming care. Six of those make "it a felony crime to provide certain forms of best practice medical care for transgender youth." At least thirteen states also have laws that ban transgender people from using bathrooms consistent with their gender identity. Some of these are focused only on public buildings. Two states, Florida and Utah, go so far as to make it a criminal offense for transgender people to use bathrooms or facilities consistent with their gender identity in some circumstances.

26 states also have laws and regulations that ban transgender students from participating in sports consistent with their gender identity, leaving around 37% of trans students living in states that restrict their ability to participate in sports. Even trans friendly states aren't exempt; Missouri and Texas have tried to force gender-affirming care providers in Washington state to turn over medical records of trans patients, including kids.

One thread runs through this legislation: politicians and radical activists are now specifically targeting transgender youth. Some of the aforementioned bathroom bans are focused only on K-12 schools, and some schools in states without relevant legislation. This focus on kids is particularly troubling given that LGBTQ+ young people are more than four times as likely to attempt suicide than their peers and roughly half of transgender and nonbinary youth considered attempting suicide in 2023.

The global context for transgender people in sports that was on display during the Paris Olympics is also relevant. Recent research sponsored by the International Olympic Committee suggests that trans women and men are at a disadvantage relative to their cis counterparts in almost every measure, and trans athletes did compete at the Paris Olympics. However, it was a cis athlete who was presumed to be trans, boxer Imane Khelif, who has endured vitriolic transphobic hatred on social media. Khelif has since filed a cyberbullying complaint against several high profile individuals with the Paris Prosecutor's Office, and the Office confirmed to the press on August 14 that it was opening an investigation.

As described in greater detail in the next section, Khelif is not alone in her experience. Online harassment and bullying directed at trans people, and even cis people perceived to be trans, is key to understanding the context of this video.

Impact

Although the research done by civil society organizations and some academics has been groundbreaking and informative, there is still insufficient academic, peer-reviewed research related to

the prevalence and impact of transphobic content, including how it may be linked to offline violence and how it may impact trans people's mental health. As explained in greater detail, the Board could make a recommendation that could help address this problem

That being said, the high level takeaway from the current research is that:

- transphobic content is prevalent on a variety of social media platforms, including Meta
- this content, along with overmoderation of content posted BY trans people, both negatively impacts the ability of trans people to use social media platforms and negatively impacts their mental health
- transphobic content is linked with real-world violence
- trans kids often rely on the Internet to access supportive communities and trans and gender diverse information

It important to accurately represent research, so to be clear: there's an obvious correlative link between transphobic content online and real-world violence. While the causative link needs more research, that's the case with almost all major content moderation issues- and the evidence in this area is significant if not conclusive. What's more, researchers have clearly documented the link between transphobic and antisemitic content, which Meta prohibits and the removal of which the Board has repeatedly and appropriately upheld. The 2022 Bratislava nightclub shooting explicitly targeted LGBTQIA+ people and the manifesto included several pages dedicated to mocking trans people. The Buffalo Shooter's manifesto largely replicated the manifesto of the Christchurch Call shooter, but it specifically included transphobic comments. It and claimed that the rise in "transgenderism" can be attributed to a Jewish conspiracy to undermine the West. Similar ideas have been widely shared by white supremacist groups. As one report notes, "[t[his anxiety surrounding queer and trans children is a continuation of antisemitic discourses that frame Jewish people as predatory."

The increase in threats against gender affirming care providers, and the link to transphobic content is more clearly causative. This content does not even need to directly call on followers to commit violence. The account "Libs of Tik Tok", which exists on several platforms, exemplifies this. The account is run by Chaya Raichik, who recently stated that "she's proud of being called a stochastic terrorist — someone who inspires supporters to commit violence by demonizing a person or group. " Media Matters documented " at least 48 instances of threats or harassment" against individuals and institutions targeted by Libs of Tik Tok post. In 2022, after the account falsely claimed on Twitter that Boston Children's Hospital performs hysterectomies on children, the hospital received a barrage of harassment including threats of violence. Later, when a woman was charged with making a bomb threat, both her lawyer and Boston Children's Hospital argued that she was directly influenced by Libs of Tik Tok while in a vulnerable mental state. Similarly, in May 2022, " FBI agents arrested a California man who had threatened to kill a staff member of a Wisconsin school district that was shamed by Libs of TikTok." And particularly relevant to this case, the day after the Libs of Tik Tok account reposted a video of a physical assault that allegedly took place in the women's bathroom at a high school in upstate New York, and claimed that the perpetrator was "a male student who identifies as a girl," the school received bomb threats.

It should be noted that Meta has suspended Libs of Tik Tok multiple times, in particular when media attention was on the account, but reinstated it quietly after that attention waned. Given the constant targeting of private adults, it's unclear how Meta justifies the presence of the account on its platforms.

Potential recommendations

Below are some specific recommendations the Board could consider, keeping in mind that prioritizing freedom of expression means limiting egregious forms of content that make the platform hostile to marginalized users.

1. Ask Meta to commission and publish a third party audit of transphobic content on its platform

In other cases, the Board has called on Meta to conduct and publish research or HRIAs on its content moderation practices. The Board should make a similar recommendation in this case. If the Board considers a third party audit completely impossible to convince Meta to take on, it can at a minimum ask Meta to conduct an internal review with clear parameters that it makes available to the public in its Transparency Center.

If done well such a review could provide significant insight into the prevalence and moderation of transphobic content on Meta platforms. It is also worth noting that since Meta has shut down Crowdtangle, it is nearly impossible for anyone to conduct meaningful research. The Board recently made several very strong recommendations regarding Crowdtangle and the need for researcher data access, and hopefully thus recognizes why it would very important for any review of transphobic content on Meta platforms to be made public.

2. Make it clear that targeted deadnaming and misgendering are policy violations for everyone

Targeted deadnaming and misgendering should be specifically listed as policy violations. Tik Tok, Snap, and Discord have such policies, and Twitter had a clear prohibition on deadnaming and misgendering until Elon Musk took over and starting using the platform to himself target and harass trans people.

The Hate Speech policy rationale states that hate speech must be 1. a direct attack on 2. a protected characteristic. The policy also explicitly includes gender identity as a protected characteristic. In this case, the direct attack could be either the use of a harmful stereotype that has "historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence." It's worth noting that, while Meta lists some specific harmful stereotypes, the list is written as if it were exemplary rather than exhaustive. Deadnaming and misgendering could also fall under the current prohibition " Statements denying existence," as misgendering or deadnaming often amounts to denying the existence of transgender people.

Furthermore, and perhaps as importantly, in some contexts deadnaming can be intended to incite offline violence. This is particularly true when it is done in the context of doxxing or when a trans person draws the attention of individuals like Chaya Raichik. Dead naming allows offline identification and targeting of trans people, as well as their families.

3. Revisit the "self-reporting" and "public figure" aspect of the Bullying and Harassment policy

Where an identifiable individual is being bullied or harassed, violations should not have to be self-reported. Even if a specific individual doesn't feel targeted or isn't on a platform where they are mentioned, allowing such content intimidates all trans users. Many trans individuals who know they may be targeted also simply do not have the time nor the capacity to review hateful posts made against them. In fact, digital security experts regularly recommend people who are experiencing online harassment to have other people monitor accounts and document content for them in order to preserve

their mental health. Meta should not be forcing traumatized people to relive their trauma on the platform over and over again simply in order to be safe. Finally, the fact that a figure is "public" does not mean that they are an acceptable target for hate, and in fact trans people in the public eye are more likely to experience harassment that is seen by others- for example, assistant secretary for health Rachel Levine noted that the attacks on her hurt " hurting the thousands of LGBTQ Pennsylvanians who suffer directly from these current demonstrations of harassment." There should not be any exception for public figures.

Reports and research:

Social Media Safety Index, GLAAD, documenting safety for LGBTQ people on social media platforms, published in 2024.

Fatal Violence Against the Transgender and Gender-Expansive Community in 2024, Human Rights Campaign: documentation of fatal violence against trans people for 2024 through August

The Epidemic of Violence Against the Transgender and Gender Non-Conforming Community in the United States, Human Rights Campaign: focus on fatal violence for the year from 2022 Trans Day of Remembrance to 2023 Trans Day of Remembrance, and declares a National State of Emergency for LGBTQ+ Americans, Published November 2023

LGBTQ Policy Spotlight: Bans on Medical Care for Transgender People, Movement Advancement Project, provides a comprehensive and up to review of legislative attempts to ban and restrict medical care transgender youth and adults, updated as of April 20, 2023

Violent Victimization by Sexual Orientation and Gender Identity, 2017–2020, Bureau of Justice Statistics, published June 2022

Understanding the impact of Bell v Tavistock, Mermaids UK, A study that looked at how a UK Court ruling that set a near-impossibly high competence standard for minors to consent to medical treatment impacted trans kids, published December 2022

Sleeping with the Enemy: Sex, Sexuality and Antisemitism in the Extreme Right,
The International Centre for the Study of Radicalisation, documents the links between gender-critical theorists, radical accelerationists, transphobia, and antisemitism, published June 2022

Transphobia in the Buffalo Shooter's Manifesto, VoxPol, Documents transphobic content, and its links to antisemitism, in the manifesto of the lone-wolf shooter who livestreamed his murder of ten people in a Buffalo supermarket, published June 2022