Oversight Board

Public Comment Appendix for

2023-023-FB-UA

Case number

Case description

In April 2023, a Facebook user in Poland posted an image of a striped curtain in the blue, pink and white colors of the transgender flag. On the image, there is text overlay that says in Polish: “New technology. Curtains that hang themselves.” Above it, other words in Polish appear that translate into English as “spring cleaning <3.” Reactions from other users to the post were positive for the most part.

Between April and May 2023, 11 different users reported the content a total of 12 times. Of these, 10 reports were not prioritized for human review by Meta’s automated systems for a variety of reasons, including “low severity and virality scores.” Only two of the reports, falling under the Facebook Community Standard on Suicide and Self-Injury, resulted in the content being sent for human review. Both reviewers assessed it as non-violating and did not escalate it further. None of the reports based on the Hate Speech policy were sent for human review.

Three users appealed Meta’s decision to keep the content on Facebook. One appeal resulted in a human reviewer upholding Meta’s original decision that the content did not violate its Suicide and Self-Injury policy. The other two appeals, made under Facebook’s Hate Speech policy, were not sent for human review.
One of the users who originally reported the content then appealed to the Board. In their statement to the Board, the user noted that the person who posted the image had previously harassed members of the trans community online and had created a new account after being suspended from Facebook in the past. As a result of the Board selecting this case, Meta determined that the content did violate its Hate Speech policy and removed the post.

The Board selected this case to assess the accuracy of Meta’s enforcement of its Hate Speech policy, as well as to better understand how Meta approaches

content that falls between hate speech and the promotion of suicide or self-injury. This case falls within the Board's seven strategic priorities, both "Hate speech against marginalized groups" and "Gender."

The Board would appreciate public comments that address:

- Speech, whether in spoken, written or visual form, that may be described by users as "humorous" or "satirical," but which may spread hate speech or other forms of inflammatory rhetoric.
- The risks associated with widespread hate speech targeting LGBTQI+ people on social media and Meta's human rights responsibilities in this context.
- The state of anti-LGBTQI+ commentary on social media and in public discourse in Poland.
- Statements that encourage or applaud death by suicide as a form of hate speech, and whether Meta's policies and enforcement practices are sufficiently adequate to address them.
- Meta's policies and practices for reviewing multiple user reports involving the same piece of content.
- Meta's account-level enforcement practices for users who repeatedly engage in anti-trans hate speech and harassment.

As part of its decisions, the Board can issue policy recommendations to Meta. While recommendations are not binding, Meta must respond to them within 60 days. As such, the Board welcomes public comments proposing recommendations that are relevant to this case.

Oversight Board

Public Comment Appendix for

2023-023-FB-UA

Case number

The Oversight Board is committed to bringing diverse perspectives from third parties into the case review process. To that end, the Oversight Board has established a public comment process.

Public comments respond to case descriptions based on the information provided to the Board by users and Facebook as part of the appeals process. These case descriptions are posted before panels begin deliberation to provide time for public comment. As such, case descriptions reflect neither the Board's assessment of the case, nor the full array of policy issues that a panel might consider to be implicated by each case.

To protect the privacy and security of commenters, comments are only viewed by the Oversight Board and as detailed in the Operational Privacy Notice. All commenters included in this appendix gave consent to the Oversight Board to publish their comments. For commenters who did not consent to attribute their comments publicly, names have been redacted. To withdraw your comment, please email contact@osbadmin.com.

To reflect the wide range of views on cases, the Oversight Board has included all comments received except those clearly irrelevant, abusive or disrespectful of the human and fundamental rights of any person or group of persons and therefore violating the Terms for Public Comment. Inclusion of a comment in this appendix is not an endorsement by the Oversight Board of the views expressed in the comment. The Oversight Board is committed to transparency and this appendix is meant to accurately reflect the input we received.

Oversight Board

Public Comment Appendix for

2023-023-FB-UA
Case number

35
Number of Comments

Regional Breakdown

3	7	25	0
Asia Pacific & Oceania	Europe	United States & Canada	Middle East and North Africa
0	0	0	
Sub-Saharan Africa	Latin America & Caribbean	Central & South Asia	

2023-023-FB-UA	PC-17003	Europe
Case number	Public comment number	Region
Jacob	Scott	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

I am nearly 30 and gay which means I grew up under very widespread online homophobia, and for the entire time meta has existed it has failed to act sufficiently to combat hatred. They and other corporate empires have established control over the online world and now tell us they can't do any better, all the while undercutting any competitor that incurs a higher cost from moderation. These spaces are the common areas we exist in now, it has a real impact on real people for life if they experience targetted hatred, but the cost of that hatred is not shown on any balance sheet, which means amorality rises to the top.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17004	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

This is all too typical of how Musk's X rebrand operates. Hatred and bigotry walks naked in the streets, while individuals who don't fit in are forced to hide in the dark.

Bring back justice.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17005	United States & Canada
Case number	Public comment number	Region
Kestrel	Perron	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Meta does not take appropriate action when hate speech is directed at the LGBT community. In fact, Meta takes no action at all. Hate speech and harassment is allowed on Meta's platforms, despite their policies to the contrary.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17006	Europe
Case number	Public comment number	Region
James	Norrington	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Meta failed to act on posts showing curtains with trans pride flag colors that said "curtains that hang themselves." This is not by any means humorous or satire, and it is by no means isolated. LGBTQ+ people are under escalating pressure and hatred at the moment and online harassment and hatred is a big part of it. Abuse and misinformation are rife and are being used both to abuse LGBTQ+ people are to create a climate of fear around them. Frankly, to put it in more personal words, things are incredibly frightening. When posts like that go unchallenged it shows bigots and abusers that there are no consequences to escalating their abuse. It teaches young and impressionable people that the murder and deaths of LGBTQ+ people is something to laugh at.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17007	United States & Canada
Case number	Public comment number	Region
Zoe Jane	Halo	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Making a joke out of a marginalized communities propensity to end their own lives is only going to make the problem worse. The content posters intent was to encourage more suicidal behaviors in the trans community by deliberately upsetting them.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17009	United States & Canada
Case number	Public comment number	Region
Jacob	Calder	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

I expect that you're all smart enough to understand WHY this is hate speech and patently disgusting, but just in case. This is a variation on a genre of posts that talk about trans people having a much higher than average rate of attempted and successful suicide. Another common variation is a vague comment about "41%", the amount of trans people who attempt suicide at one point in their life.

The point of these posts is to both mock this and to encourage trans and queer people to kill themselves.

It's fundamentally repugnant and meta's refusal to stick to their community guidelines is both a moral and bureaucratic failing.

With that out of the way -- I want to address WHY Meta fails to properly moderate these posts. I was an engineer at Meta for nearly 5 years. I left in good standing and could probably return at any point. I say this to make it clear that I do not see everyone associated with the company as fundamentally evil and to

illustrate the fact that I have a nuanced understanding of what goes into running a social network.

Meta fails to moderate these posts for several reasons but the largest is moral cowardice. Over the 5 years I was there I saw the company back down from protecting vulnerable people when confronted with pressure from right wing politicians, hate groups, and from internal far right employees. Time and time again the company erred on the side of doing nothing because they are fundamentally afraid of scrutiny from the press. A negative outcome from inaction is fine -- that can be written off as a failing of systems or a single human error. However, when an action is taken they open themselves to controversy and criticism.

They'll never be perfect, they'll never catch everything or do everything "right". "Right" isn't even a real thing, these are messy hard issues. The problem is that Meta, particularly under Clegg and Kaplan has no interest in trying to engage with those sticky questions, instead they look to maximize their deniability.

At best they lack moral courage. At worst they lack morals all together. It doesn't really matter which one is true because the end result is the same: Meta allows hate speech against vulnerable groups to be spread on their platforms until mass violence or genocide occurs and then they retroactively make an apology and then do it again.We saw it in myanmar, we saw it in Kenya, and it feels inevitable that queer people in Poland, Hungary, and even the United States are next.

I left Meta for a number of reasons but one of them was that it felt like there was no way to make things better from the inside. All avenues for change were closed in the name of plausible deniability and quashing controversy. I still believe that internal change is impossible.
For that reason I'm skeptical that the oversight board will be able to change things. That said you're in the best position of anyone to do so. I hope you have more moral courage than Meta's leadership.

Link to Attachment
No Attachment

2023-023-FB-UA	PC-17011	United States & Canada
Case number	Public comment number	Region
nan	nan	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

who cares !!!!!!!!!!! what about the rest of the world who don't agree with this agenda at all ,all i have seen myself is a biased judgement against everyone in the world because of a small group whose feelings are hurt this makes no logical sense at all nor will i acknowledge this type of delusion .if they mind their own business and just live their lives without trying to push this on everyone then im sure there wouldnt be an issue its a issue because of the constant bias and favoritism .

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17012	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-17012

2023-023-FB-UA	PC-17013	Europe
Case number	Public comment number	Region
Caroline	Sinders	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Convocation Design + Research		Yes
Organization		Response on behalf of organization

Full Comment

Not to just echo the words of Jenni Olson, Senior Director of Social Media Safety at GLAAD, but they say it best: “Meta’s content moderators should have enforced their policies in the first place. It is a very serious problem that the post was only removed after the Oversight Board alerted them. The post is clearly asserting the horrific sentiment that trans people should kill themselves. While Meta eventually removed the post, this case powerfully illuminates highly consequential systemic failures with the company’s moderation practices that have broad implications in relation to all anti-trans and anti-LGBTQ hate content, as well as even larger implications for such coded hate content that is all too common and targets all historically marginalized groups. Such moderation may be more complex than recognizing basic slurs,” Olson added. “But this is why trust and safety teams must provide adequate training and guidance to their moderators on recognizing anti-trans hate. Meta is fully capable of implementing such training and yet continues to fail to prioritize it, resulting in epidemic levels of overt and coded anti-trans, and anti-LGBTQ hate across their Facebook, Instagram, and Threads platforms. As highlighted in

GLAAD’s 2023 Social Media Safety Index (SMSI) report, Meta’s Facebook and Instagram are largely failing to mitigate dangerous anti-trans and anti-LGBTQ hate and disinformation, despite such content conflicting with their own policies. The June 2023 SMSI also made the specific recommendation to Meta and others that they better train moderators on the needs of LGBTQ users, and enforce policies around anti-LGBTQ content across all languages, cultural contexts, and regions."

As a nonbinary and trans executive director, anti-trans rhetoric and violence is on the rise. Social media plays a part within that-- and in normalizing violence. It's deeply important for Meta, and the Oversight Board, to stand firm in this kind of violence, particularly violence that urges trans people to kill themselves. This is violence, and Meta needs to enforce their policies at scale, particularly to protect marginalized communities, like trans people.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17015	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

This comment focuses on three core matters: The risks associated with the extreme climate of hate present on Meta platforms, Meta's role in perpetuating that harm, and the specific matter of treating this "joke" as anything but encouraging self-harm:

1. The extreme climate of hate and abuse present on Meta Platforms:

Meta continues to allow a number of hate groups, including both "Gays" against "Groomers", an anti-trans hate group, "Libs of TikTok", an anti-trans hate account, and others. Moreover, it has permitted advertising for hateful content produced by Matt Walsh, a major figure in the anti-trans hate movement, on Facebook. While this is hardly the first time that Meta has looked the other way while genocidal lies have been told on its platforms, the preponderance of English-language hate material on Meta platforms eliminates its typical excuse that it simply doesn't have moderators. Hateful conduct, genocidal lies, and stochastic terrorism are laundered through Meta products with some regularity.

In a world in which legal consequences existed for a firm like Meta, it would face significant legal exposure in its general inaction in the face of this content. Sadly, this is not that world. That is not to say that Meta faces no such risk, merely that the risk it faces is far less than the risk it should face. Nevertheless, in permitting this hateful climate, Meta endangers the lives of trans people worldwide. Regardless of potential legal consequences, it is morally deficient in permitting this climate to remain. Meta may feel that permitting hateful conduct drives engagement, and that may even be true. Despite that assertion, the engagement being driven is enabling a moral panic aimed at vulnerable people.

2. Meta's role in the perpetuation of harm:
Meta chooses to treat all discourse on the rights of transgender people as a matter of politics as usual, and thus to protect the political speech of those engaged in the anti-trans hate movement. This, then, leads to a great deal of hateful content being shared on Meta platforms and protected by Meta as "political speech"' This structure of "political speech" is, itself, a harmful construct, because it sets aside the moral obligation of Meta to protect vulnerable people and allows Meta to in effect not consider the harms of such speech, purely because it is labeled political. Given that genocidal rhetoric is, in fact, a call for political action, this taxonomy of speech leads to any genocidal rhetoric that manages to give itself disclaimability being considered protected speech. Thus, in permitting anti-trans hate groups and anti-trans rhetoric on its website without an actual critical analysis of its aims and consequences, Meta serves as an enabler of these genocidal views.

I do not use any form of Facebook groups any more. I do not use Threads or Instragram. The reason for that is I have no desire to be exposed to constant abuse by bigots and have that paired with the rage-inducing indifference of Meta to that abuse. It is quite simply not safe for trans people to use Meta products. Given Meta's stated goal to connect people, it would appear that either we are not people to Meta or Meta believes that abuse is a form of connection. Either conclusion is deeply troubling and it is one we receive with some regularity from social media companies. That Meta is less egregious in this than its main competitor makes the tolerance for abuse on Meta platforms no less execrable.

3. The "joke" and the human decision made upon appeal:
"haha, trans people kill themselves" is not a joke. It is the cruel taunt of an abuser. Wrapping it in the paper-thin pretense of a pun about curtains does not change that fact. Moreover, making a joke of suicide at all is never appropriate and is always a form of suicide baiting. Human review failed entirely in this case. The human reviewing the "joke" in question either received bad instructions or did not follow those instructions. Regardless of the reason, Meta is morally culpable for the fact that the person who made the complaint received the news that a human at Meta had reviewed the "joke" and decided that suicide baiting aimed at trans people (and, quite probably, suicide baiting aimed at them or their loved one(s)) was perfectly okay. The message that sends is that power does not care what harm is done to trans people and that we have no moral worth of our own.

That this was not immediately obvious to the person who performed this review is concerning in and of itself, because either something is deeply wrong with that person or something is deeply wrong with the process that person uses. Either way, Meta is obligated to both take responsibility for this behavior and correct it. Trans people should not be receiving the message that Meta thinks suicide baiting us is okay.

That it took this board choosing to take the matter up for Meta to take any action suggests that the Board's model of considering individual cases is inadequate. Rather, it appears that a broader consideration of the practical realities of enforcement must be taken up. Depending on a process of appeal and selection to hope for a less unjust outcome is not scalable or practical. This board must be able to act proactively and to have real oversight power if any kind of fair enforcement is to be achieved. In the absence of such a thing, the Board serves as a sop to criticism with no actual substance.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17016	United States & Canada
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

Transgender people are continually under attack in this world and Meta's lack of action in addressing hate speech and harassment against this community sorely hurts the population. Content like this reinforces why social media continues to be such a dangerous and unwelcoming environment for LGBTQ+ individuals. This is a case of hate speech masquerading as a "joke" but targets a vulnerable marginalized group. Even blatant hate speech with NO allegedly humorous pretext routinely appears on Meta and other social media platforms, and is allowed to remain online due to poor moderation, alongside poor and inadequate policies that allow hate speech to thrive and minorities to be targeted. This sustained hate, repeatedly allowed to remain online by social media companies, has detrimental effects on minority groups and contribute to a world in which hate crimes flourish and where simply existing as a minority is hazardous to our health. If Meta seeks to call itself inclusive and dedicated to social responsibility, the platform must not just devise policies to more effectively combat hate speech, but to enforce and remove violative content when it appears.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17017	United States & Canada
Case number	Public comment number	Region
Vanessa	Teeter	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

The trans community has been targeted by bullies on Facebook to the point of ruining lives, causing suicides, and driving people to psychiatric holds. I know cases were hundreds of users reported the same offenders, were told they weren't doing anything wrong. This is for things like photoshopping a noose around a tans person's neck, or making entire pages designed to make fun of trans people. False allegations of grooming, etc. Also go unaddressed. We, the global queer community, know that Facebook does not have our backs. In fact we will regularly have screenshots of the hate speech flagged as hate speech when it was fine when it was directed at us, but not when we were pointing it out. I have thousands of names of people involved in these bullying groups, and hundreds of screenshots of facebook dping nothing. Facebook needs to be sued for the LGBTQ blood on their hands.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17019	United States & Canada
Case number	Public comment number	Region
Amy	Lancaster	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Anti trans hate speech uses a number of methods to obfuscate their message behind dog whistles. Because anyone at facebook refuses to listen when the trans community is telling them that content contains hate speech the platform has become overrun with messages telling trans people to kill themselves among many other horrible things. Facebook has never once removed content I've reported with anti trans hate speech and my only takeaway can be that they support that messaging on their platform.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17020	Europe
Case number	Public comment number	Region
Ryan	McLeod	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

I am a Pole by origin, currently employed by an LGBT centre in the UK (Scotland). I welcome this process and ability to submit my comment. In my opinion the post in question should have been removed as it promotes hate crime.

As someone who supports trans people for a living and is also trans, I am fully aware of the impact of "humorous" posts like this on the community and how the presence of posts like this enables social media users to feel like online harassment is also acceptable. I trust that the Oversight Board makes the right decision, which will result in clearer guidelines for such posts in the future.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17021	Asia Pacific & Oceania
Case number	Public comment number	Region
Zahra	Stardust	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Dear Members of the Oversight Board,

Re: Post in Polish targeting trans people (2023-0230FB-UA)

Thank you for the opportunity to make a public submission on this case.

I am a Postdoctoral Research Fellow in the Australian Research Council Centre of Excellence for Automated Decision-Making and Society, situated in the Digital Media Research Centre at the Queensland University of Technology.

I work under the supervision of Professor Nicolas Suzor, who is also a member of the Oversight Board. The contents of this submission represent my views alone and Professor Suzor has not viewed nor influenced its development or publication.

The text in this post ought to be read as a case of 'gender cleansing' which seeks the systematic eradication of trans people. Trans people remain at higher risk of suicide and suicidality, and there is currently a dangerous international trend towards removing the rights of trans people to education, bathroom access, and basic services. Attempts at 'humour', made at the expense of trans people, by denigrating trans people, founded on the premise that trans lives are not worthy, and trivialising the serious issue of trans suicide, ought to be considered hate speech. Statements that encourage or applaud death by suicide must be treated with utmost seriousness.

In 2022 my colleagues and I undertook research analysing the newsroom posts of five dominant social media platforms, including Meta, to understand how they spoke publicly about safety and harm. Hate speech was one of the most prevalent harms raised by the platforms, who promoted the speed at which they detected, suppressed and removed it. However, the kinds of hate speech discussed were selective, mostly focused upon anti-Semitism and less frequently to misogyny and racism.

At the same time, we found that the platforms took a tokenistic approach to inclusion and diversity, posting about pride, celebration and visibility of LGBTQ content creators, but inadequately addressing structural, systemic, interpersonal and platformed violence towards trans people. The lack of attention to transphobia as a form of hate speech suggests that platforms do not adequately grasp the severity of this issue or their responsibility to address it.

This research can be found at: https://doi.org/10.1177/20563051221144315

To leave this content online is to be complicit in a pandemic of transphobia. Platforms are responsible not only for removing hate speech but for proactively generating safe and inclusive environments for trans and gender diverse users.

Sincerely,

Dr Zahra Stardust

Postdoctoral Research Fellow
Digital Media Research Centre
e: zahra.stardust@qut.edu.au

Link to Attachment

PC-17021

2023-023-FB-UA	PC-17023	United States & Canada
Case number	Public comment number	Region
nan	nan	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Hate speech is anything making light of the suffering of living victims and marginalized communities. It is not something that can be regulated by algorithm, but humans who actually have a functioning sense of compassion.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17024	Europe
Case number	Public comment number	Region
Withheld	Withheld	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Withheld		No
Organization		Response on behalf of organization

Full Comment

I've grown up in Bangladesh as a Muslim. The world has no idea how much of hatred towards the trans people can a society hold together. Even deaths are trolled, mentioning that they are trolling not a human's death, but of something less worthy. When I came to Europe for study, I saw not much of a difference in people's mindsets. Disrespecting trans people has become a fashionable trend among the people now. Even, they have managed to include many women with their hatred campaign too. Posts where trans people are mocked are not considered as hateful enough to remove from platforms. Transgenderism, in their term, has become an easy target to mock and claim to be rationalists at the same time! I believe the oversight board will keep these in mind while making their decisions. I'm writing my comment during a very busy schedule, but wish to engage more in the future.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17025	United States & Canada
Case number	Public comment number	Region
David	Inserra	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

As a former member of the content policy team at Meta, my view is that the content in question likely does not clearly violate the letter of Meta's policies. In its public facing community standards, Meta says that Hate Speech is "a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics." In this case, the piece of content is described as a curtain that contains the colors of the transgender flag and text that refers to "self-hanging curtains." Nothing in this is explicitly about people. To be a violation under the hate speech policy, the target must be about people who are defined by a PC. While we (or reviewers) can imply that the meaning here is some sort of twisted praise of transgender people committing suicide, the content in both visual and text form does not clearly say that. I suspect that Meta's content policy team made that implied leap during their analysis of the content, while reviewers are prohibited from making that leap at scale.

As such, the major question with this content is whether or not reviewers at

scale should be expected to make implicit leaps, not just for hate speech but across Meta's policies. While it may be tempting to see the non removal or this content at scale by Meta as a reason to have reviewers make implied leaps, there are important reasons not to do so.

First is inaccuracies and inconsistencies in reading user intention. While the content in this case appears to have only one meaning (although a less likely but plausible reading is that the transgender movement is hanging itself aka alienating others in society by making demands that many do not agree with) that will certainly not be the case with other implied content. Users who may wish to report or condemn various harmful or evil things and do so in a way that is sarcastic or otherwise could be implied to actually be harmful and hateful. Should reviewers default to viewing that implied content as hate speech? Or the hand gesture of "ok" is apparently used as a hate symbol for white supremacy. Should reviewers be expected to interpret every use of the ok symbol as hateful? Or only the hateful uses of the ok symbol? Or what about when its just not clear? How much assuming should reviewers be allowed and expected to do? And is there any hope that this will be consistent? Or take a statement that "the Catholic Church should burn" or "be destroyed" for its history of hiding sexual assaults- should the explicit reading that it is targeting an organization hold or should we assume that the target is actually Catholics or, perhaps even more plausibly, physical church buildings?

This leads to the 2nd major problem which is not just inaccuracies but allowing and empowering bias. If reviewers are broadly expected to make assumptions and understand implied violations, then it is empowering reviewers to make biased decisions. Among even the most benign types of speech, it is possible to assume ill intent. Take the innocuous punctuation joke about "Let's eat [,] grandma." Could that not be implied to be a call to violence against one's grandmother? Now jump to controversial, political issues. The Board recently overturned several pieces of abortion related content and one can imagine all sorts of similar content. Now expect reviewers who may have strong views one way or another on abortion to make the "correct" assumptions about implied violence, hate speech, or restricted goods content, and it will only mean more of

that type of content being removed (or content that should be removed being allowed) and quite possibly accurate accusations of bias being levied against Meta. Even if we assume reviewers aren't acting maliciously, the real problem of just not understanding a perspective one does not hold will regularly mean that reviewers act out of ignorance of an alternative view and reading of content because of their bias.

One counter argument in this case might be that the transgender colors represent people and so therefore there is a direct attack against people in the form of hanging people. Once again, however, that opens a very concerning precedent to expect reviewers to action against. The description of the content does not include any person but only a flag. A flag is not a person. It may represent a nation or a set of ideas but it is not a person. If attacks on flags are considered hate speech, should burning, shooting, stepping on, or otherwise desecrating any flag be considered hate speech? And moving beyond flag to other symbols, should desecrating holy books, leaders of religions, leaders of nations, etc. also be considered hate speech? It is clear that the answer to these questions must be no. And so the transgender colors on the curtain (or any other flag or symbol) must not be conflated with people as a rule. Doing so would open a Pandora's box of blasphemy law into Meta's policies.

And so the best outcome here may simply be that on escalation, this kind of content can be reviewed more holistically for implied violations. Even here, however, a clearer understanding of when the expert teams at Meta should make assumptions and when they should not, could be helpful, especially in a manner that is public and transparent. After all, no one is immune from potential bias even if just one's own ignorance, and so these experts should also have clear standards to prevent external, internal, or personal pressures from influencing the outcome of cases involving implied violations.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17026	United States & Canada
Case number	Public comment number	Region
Paige	Collings	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Electronic Frontier Foundation		Yes
Organization		Response on behalf of organization

Full Comment

Submission to the Meta Oversight Board re: Post in Polish Targeting Trans People case
By Paige Collings (Electronic Frontier Foundation)

Introduction
Just as Facebook can be used for positive advocacy, it is also routinely used with the intention to cause harm. That was clearly the case in April 2023, when a Polish user posted an image of a curtain in the colors of the transgender flag with the text overlay stating (in Polish), “New technology. Curtains that hang themselves” and “spring cleaning <3.” The intent behind this message is clear: To encourage self-harm by and violence toward transgender individuals.

While this content was recognized and reported by a number of users,

Facebook's automated systems failed to prioritize the content for human review. From our observations—and the research of many within the digital rights community—this is a common deficiency made worse during the pandemic, when Meta decreased the number of workers moderating content on its platforms. In this instance, the content was eventually sent for human review and was still assessed to be non-violating and therefore not escalated further. Facebook kept the content online despite 11 different users reporting the content 12 times and only removed the content once the Oversight Board decided to take the case for review.

This incident serves as part of the growing body of evidence that Facebook's systems are inadequate in detecting seriously harmful content, particularly that which targets marginalized and vulnerable communities. Our submission will look at the various reasons for this shortcoming and make the case that Facebook should have removed the content—and should keep it offline.

The Shortcomings of Automated Decision-Making and Poorly Trained Human Reviewers

As EFF has demonstrated, Meta has at times over-removed legal LGBTQ+ related content whilst simultaneously keeping content online that depicts hate speech toward the LGBTQ+ community. This is often because the content—as in this specific case—is not an explicit depiction of such hate speech, but rather a message that is embedded in a wider context that automated content moderation tools and inadequately trained human moderators are simply not equipped to consider. These tools do not have the ability to recognize nuance or the context of statements, and human reviewers are not provided the training to remove content that depicts hate speech beyond a basic slur.

The lack of transparency only adds to the complexity of the issues as Meta does not disclose the detailed criteria for content moderation, including enforcement guidelines related to internal policies—making it difficult to assess the scale and contours of such bias as reflected in opaque internal policies, as well as any potential built-in bias regarding the moderation of LGBTQ+ content. Additionally, because algorithms can only be trained on known examples, they

are more likely to remove similar kinds of content and can be blind to others. The challenges of content moderation enforcement in languages other than English—such as Polish—further exacerbates these issues.

In countries like Poland where anti-LGBTQ+ hate speech and harassment is so prevalent both online and offline, Meta's inconsistent and inflammatory content removal systems are even more detrimental. As highlighted in GLAAD's 2023 Social Media Safety Index (SMSI) report, Meta's Facebook and Instagram are largely failing to mitigate dangerous anti-trans and anti-LGBTQ+ hate and disinformation, despite such content conflicting with the sites' policies. The June 2023 SMSI also made the specific recommendation to Meta and others that they better train moderators on the needs of LGBTQ+ users, and enforce policies around anti-LGBTQ content across all languages, cultural contexts, and regions.

Under international human rights law, restrictions to rights such as freedom of expression (article 19 ICCPR) and freedom of assembly and association (articles 21 and 22 ICCPR) can only be justified if there's a legal basis, a legitimate aim, and if they are necessary and proportionate. Without adequately taking into consideration the context in which words and audio-visual content is used, benign content is suppressed whilst hate speech and content inciting violence is able to remain online; thereby failing to meet the conditions to restrict freedom of expression, civic engagement, and activism under international human rights law.

Recommendations

It is of vital importance that online speech is put into its appropriate context. The Rabat Plan of Action provides guidance for companies seeking to remain in compliance with the UN Guiding Principles. Meta should consider (1) the social and political context prevalent at the time the post was uploaded; (2) the user's position or status in the society, specifically the individual's or organization's standing in the context of the audience to whom the post is directed; (3) the intent of the user in relation to their audience; (4) the content of the post; (5) the extent of the post, taking into account the post's reach, its public nature, its magnitude, and size of its audience; and lastly (6) the likelihood, including

imminence, of harm to result from the post.

Meta's Trust and Safety teams at Facebook, Instagram, and Threads must also provide adequate training to human reviewers to recognize how hate speech and incitement to violence can appear in a more nuanced manner than basic slurs or hateful images. The image shared on Facebook in April 2023 was a clear illustration of anti-trans hate speech, and the arbitrary and inefficient content review—first by the automated system and second by the human reviewer that chose to keep the content online—has a particularly detrimental impact for the LGBTQ+ individuals using online platforms in countries like Poland, where hate and harassment is so prolific. The anti-trans content posted on Facebook in April 2023 must remain offline.

Link to Attachment

PC-17026

2023-023-FB-UA	PC-17027	United States & Canada
Case number	Public comment number	Region
nan	nan	English
Commenter's first name	Commenter's last name	Commenter's preferred language
GLAAD		Yes
Organization		Response on behalf of organization

Full Comment

DID NOT PROVIDE

Link to Attachment

PC-17027

2023-023-FB-UA	PC-17028	United States & Canada
Case number	Public comment number	Region
nan	nan	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Institute for Strategic Dialogue (ISD)		Yes
Organization		Response on behalf of organization

Full Comment

Please see attached for full comment with sources and references. Text version below:

Thank you to the Oversight Board for the opportunity to comment on case 2023-023-FB-UA, regarding a post in Polish targeting transgender people. Transgender rights and lives are currently under attack not just in Poland or the United States but around the world. Now more than ever, it is critical to support and implement policies and laws to protect transgender and gender-diverse people in every of facet society – including on social media platforms.

The Institute for Strategic Dialogue (ISD) is an independent, non-profit organization dedicated to safeguarding human rights and reversing the rising tide of polarization, extremism, and disinformation worldwide. Our work

includes in-depth research and analysis identifying and tracking online manipulation, mis- and disinformation, hate, and extremism in real time. We also formulate, advocate and deliver evidence-based policy approaches and programming.

Transgender and gender-diverse people around the world face an increased risk of harm due to the many forms of discrimination they face daily, both online and offline. The situation in Poland is particularly dire – it is currently ranked 42nd out of 49 countries in the IGLA-Europe's 2023 Rainbow Europe Map – but not unique. The year 2022 saw 327 reported murders of transgender and gender-diverse people across the world. There is also higher suicide risk for transgender people than cisgender people: a recent study showed that transgender people in Denmark had 7.7 times the rate of suicide attempts and 3.5 times the rate of suicide deaths compared to the rest of the population. Other studies by organizations such as The Trevor Project found that the share of LGBTQI+ youth who reported "seriously considering suicide" increased from 2020 to 2022.

These concerning figures and rise in transphobic legislation or movements in countries like Poland and the United States have led several human rights advocacy groups to take action. Social media platforms, which many use for healthy debate, information-gathering, and learning about new topics, have the duty to protect LGBTQI+ people on their platforms by not only developing comprehensive policies but also enforcing them correctly and uniformly. This is especially critical when widespread online hate speech can cause serious offline consequences, such as encouraging individuals to cause physical harm to themselves or others. In this case, the post could be considered a violation of Meta's Hate Speech Policy , Suicide and Self-Injury Policy , and Bullying and Harassment Policy .

Our submission seeks to address the Oversight Board's request for comments on Meta's policies and enforcement practices regarding hateful content targeting transgender people:

1. Speech, whether in spoken, written or visual form, that may be described by users as “humorous” or “satirical,” but which may spread hate speech or other forms of inflammatory rhetoric.

Previous ISD research has shown that internet memes or content that might be described as “humorous” or “satirical” by some users can be used strategically to obfuscate extremist narratives and convey hateful meanings through association rather than explicit argument. Additionally, extreme right-wing movements regularly use memes to condense radical ideologies into a more ‘palatable’ format that is easier to spread online and recruit and radicalize others. It also gives users an easy way to deflect any accusations of violating platform policies, spreading hateful or extremist ideologies, or targeting other users or groups online: users can claim the content was a “joke” or “satire.” However, hateful content that uses humor, whether spoken, visual, or written, is still hateful content.

Recommendations:

- Meta should clarify in its Hate Speech Policy how reviewers determine whether content was intended to be satirical or not, and what that process looks like, including by providing indicative examples.

- Meta should invest in content moderation systems – whether human or through machine learning – that can catch the spread of extremist and hateful ideology through memes and “humorous” content (especially if the user posting it has strikes for other policy violating content or actions in the past) and curb inflammatory rhetoric.

- Meta should proactively invest in content moderation systems that operate in local languages, which are more adept at capturing the levels of nuance required to better identify content that propagates extremist and hateful ideology.

2. The risks associated with widespread hate speech targeting LGBTQI+

people on social media and Meta's human rights responsibilities in this context.

Widespread hate speech targeting LGBTQI+ people online have serious offline consequences. In June 2023, ISD published a series of reports highlighting how anti-drag mobilization efforts (which frequently amplify anti-trans talking points) are organized online and carried out offline. Our US report showed how online toxicity and hateful rhetoric can lead to offline aggression and verbal or physical assault. Our UK report showed how UK-based anti-drag activists were influenced by US activists and content online. In Poland, LGBTQI+ pride parades have ended in verbal or physical assault, with government officials parroting some of the rhetoric attacking LGBTQI+ people that is popular online, such as the "groomer" slur. Meta owes its LGBTQI+ users safe platforms where they can exercise their freedom of self-expression without fear of retaliation or hate speech.

Recommendations:

- Meta's policy teams need to be responsive to these kinds of trends and adapt policies and enforcement accordingly.

- Meta should also regularly brief moderators on emerging or spiking forms of hate and potential content violations.

3. Statements that encourage or applaud death by suicide as a form of hate speech, and whether Meta's policies and enforcement practices are sufficiently adequate to address them.

Currently, Meta's Hate Speech Policy does not sufficiently address statements encouraging or applauding death by suicide of people with certain protected characteristics. The closest the policy gets to doing so is by prohibiting "expressions that a protected characteristic shouldn't exist." Similarly, the Suicide and Self-Injury Policy does not once refer to the Hate Speech Policy or address protected characteristics. While case 2023-023-FB-UA could technically fall under Hate Speech or Suicide and Self-Injury, the most relevant policy that

was overlooked by Meta and reviewers was the Bullying and Harassment Policy, which states that "everyone is protected from [...] calls for self-injury or suicide of a specific person, or a group of individuals."

Recommendations:

- Meta should bridge the gap between its Suicide and Self-Injury Policy and Hate Speech Policy by adding a clause in its Hate Speech Policy prohibiting posts alluding to, suggesting, or even outright stating that people with a protected characteristic(s) should die by suicide.

- Meta should give users the option to report a post for multiple violations. In this case, it might have been hard for a user to decide between which policy to prioritize using the existing user reporter tools: Hate Speech, Bullying and Harassment, or Suicide and Self-Injury. It would have also allowed Meta's reviewers to understand that this case was at an intersection of, and likely violating, multiple Meta policies.

4. Meta's policies and practices for reviewing multiple user reports involving the same piece of content.

In the past, ISD has documented how Meta's "delays or mistakes in policy enforcement" have allowed for hateful and harmful content to spread through paid targeted ads. In the past couple of years, researchers and organizations have noted that these repeated delays or mistakes have extended beyond just ads, and that Meta's practices for reviewing violative content are not always entirely accurate. While 100% accuracy is unrealistic, in this case, there seemed to yet again be an inconsistency in the flagging of the content to human reviewers and the amount of information sent to human reviewers to help inform their decision.

Recommendations:

- Meta should set a pre-determined number or percent of reports (no matter what policy they fall under) over a certain number of impressions or

views, when reached, the content is automatically sent to a human reviewer. With this set policy in place, for example, if an Instagram post were to be reported 10 times (2 times for Harassment and Bullying, 5 times for Hate Speech, 3 times for Suicide and Self-Injury) for 100 views or impressions, Meta would automatically send it to a human reviewer – regardless of the virality or severity.

- Meta should invest in more human expertise to continue to finetune the balance between human moderation and automated moderation in its cross-check program.

- Meta should be transparent about how they uniformly and unbiasedly determine "high-impact content" in their cross-check system and how they choose what gets sent to human reviewers. Meta should provide a breakdown every quarter of the themes of cases that were sent to human reviewers.

- Meta should inform human reviewers that receive cases that have been reported under multiple policies which policies were selected by the users reporting. It is unclear whether the human reviewers knew that users also reported the post in case 2023-023-FB-UA for Hate Speech or just Suicide and Self-Injury.

5. Meta's account-level enforcement practices for users who repeatedly engage in anti-trans hate speech and harassment.

Meta should have zero tolerance for users who repeatedly engage in anti-trans hate speech and harassment, especially if the user has been banned or suspended from Meta platforms before. Meta should ban IP addresses, phone numbers, or emails of repeat offenders to dissuade them from rejoining the platform.

Link to Attachment

PC-17028

2023-023-FB-UA	PC-17029	United States & Canada
Case number	Public comment number	Region
Shoshana	Goldberg	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Human Rights Campaign Foundation		Yes
Organization		Response on behalf of organization

Full Comment

Shoshana Goldberg, PhD MPH
Director of Public Education and Research
Human Rights Campaign Foundation
1640 Rhode Island Avenue, NW
Washington, DC 20036

The Human Rights Campaign Foundation welcomes the opportunity to provide public comment in response to Meta's Oversight Board Case 2023-023-FB-UA ("Case") concerning "Post in Polish Targeting Trans People." Below, we respond to five of the six questions posed by the Oversight Board, calling on our expertise as the largest civil rights organization working to achieve lesbian, gay, bisexual, transgender, and queer (LGBTQ+) equality.

We argue that neither Meta's Suicide and Self-Harm policy, nor its Hate Speech policy sufficiently capture the nuance and breadth of harmful speech on their platforms, with both written too narrowly to allow for appropriate identification and removal. This is particularly true for posts involving innuendo, coded, and so-called 'humorous' language. Given the grievous risk to mental health and well-being that can result from exposure to hate speech, we recommend that Meta revise their content policies and content moderation practices. This recommendation is both timely and urgent in light of growing anti-trans sentiment which often begins online, but, increasingly, finds its way to legislative action and even physical violence throughout the world. Finally, we recommend that Meta move beyond moderating content on a post-by-post basis to also consider a user's history, including developing more stringent and transparent practices for handling repeat offenders.

Question 2: The risks associated with widespread hate speech targeting LGBTQI+ people on social media and Meta's human rights responsibilities in this context.

In including "hate speech against marginalized groups" as one of the Oversight Boards strategic priorities "to reshape Meta's approach to content moderation," the Board acknowledges the substantial risks of such content, and the urgency of addressing it.

Hate speech which starts online rarely stays there. In December 2022, the Human Rights Campaign Foundation tracked coordinated social media attacks on Meta and X (formerly known as Twitter) by a few select users against doctors and providers of gender-affirming care in 24 children's hospitals and clinics in 21 U.S. states, which were followed by bomb threats, doxing, and threatening, violent, and hate-filled offline harassment.

Exposure to hateful and harassing content online can increase risk for anxiety, depression, and suicidality: A 2021 report from the Anti-Defamation league found that 11% of American adults surveyed who had encountered online hate/harassment had "depressive or suicidal thoughts" as a result. A meta-analysis of 20 studies of cyberbullying and mental health among children and

adolescents found that those who experienced cyberbullying were over twice as likely to self-harm, have suicidal thoughts, and/or attempt suicide. Simple exposure to online hate speech and harassment targeting one's can severely harm mental health and well-being: HRC and the University of Connecticut's 2022 LGBTQ+ Youth Study found that 96% of all LGBTQ+ youth, including 97% of transgender and gender-expansive youth, had seen hateful or offensive anti-LGBTQ+ content, memes, or posts on social media, with those who encountered this content significantly more likely to screen positive for depression and anxiety. This risk was highest for trans and gender-expansive youth, with over two-thirds of those who encountered online hate speech screening positive for anxiety (68.6%), compared with half (51.9%) of cisgender LGBQ+ youth. A report from UltraViolet, GLAAD, Kairos, and Women's March found that the majority of LGBTQ+ adults feel personally attacked simply after encountering online harassment of other LGBTQ+ figures.

The negative impact to mental health that can result from exposure to online hate speech is particularly serious for LGBTQ+ and transgender people, who, as a group, are already at heightened risk for suicidality and self-harm. The 2015 U.S. Transgender Survey found that 40% of respondents have attempted suicide in their lifetime—nearly nine times the attempted suicide rate in the U.S. population (4.6%). The Trevor Project's 2022 National Survey on LGBTQ Youth Mental Health found that more than half of transgender and nonbinary youth considered attempting suicide in the previous year.

When a population group has an increased risk of suicidality, it does not come out of nowhere. Rather, it is tied to social, cultural, and structural factors in their environment which perpetrate isolation, exclusion, and stigma. Globally, and in Poland in particular, the transgender, non-binary, and gender-expansive community continues to face heightened stigma, discrimination, and violence throughout their daily lives, perpetuated through hate speech such as that reflected in the post at the heart of this Case.

Thus, Meta and the Oversight Board, in their desire to achieve their strategic priority of "protecting marginalized groups," must ensure that hateful and

violent speech that targets these groups is not allowed to remain on their platforms.

Question 3: The state of anti-LGBTQI+ commentary on social media and in public discourse in Poland.

The Case must be considered against the backdrop of public discourse and anti-LGBTQI+ sentiment in Poland. In their 2023 Annual Review of the Human Rights Situation of (LGBTQI+) People in Europe and Central Asia, ILGA Europe ranked Poland as the lowest country across the entire European Union in terms of achieving LGBTQI+ rights (and 42nd out of 49th across the European continent), a position it has held consistently since 2020. In recent years, Poland has emerged as a hotbed of state-sponsored /supported anti-LGBTQ+ stigma and discrimination: in 2019, various municipalities and voivodeships in Poland infamously began establishing themselves as 'anti-LGBTQ+'/'LGBTQ+ free' zones in response to the Warsaw LGBT declaration; by June 2020, almost 30% of the country was living in one of these zones. It was only in late 2021, after the European Commission threatened to block funds to Poland because of these declarations, that they were withdrawn.

At the same time, anti-LGBTQ+ biases and hate speech continue to proliferate in Poland. In their full report, ILGA Europe highlighted Poland as a country that has seen a "continuing trend of rising hate speech," and a rise in anti-trans hate speech specifically, both online and espoused by politicians and state representatives. As noted in the State Department's 2022 Country Report on Poland, Jaroslaw Kaczynski, the leader of the ruling PiS party, made transphobic statements, resulting in a reprimand from the Sejm Ethics Committee who deemed his words a "disgraceful mocking of transgender people," and Education and Science Minister Przemyslaw Czarnek retracted his own social media post in which he stated that LGBTQI+ persons were "not equal to normal people." These statements, among others, perpetuate and legitimize what Amnesty International has deemed "an atmosphere of hostility" toward LGBTQI+ people in Poland.

Question 4: Statements that encourage or applaud death by suicide as a form of hate speech, and whether Meta's policies and enforcement practices are sufficiently adequate to address them.

That the post in the current Case—which was described by the Oversight Board as "an illustrated meme format to mock transgender victims of suicide and to promote suicide in the transgender community"—was found to not be volitive suggests current Meta policies and enforcement practices around suicide and self-injury are NOT sufficiently adequate to address their desired aims.

Meta's Suicide and Self Injury policy states it does not allow content which "intentionally or unintentionally celebrate[s] or promote[s] suicide or self-injury" or which "mocks victims or survivors of suicide, self-injury." Thus, in using the language they did to describe the Case, the Board highlights how the post should have been viewed as a violation of Meta's Suicide and Self-Harm policy, both with regards to the spirit and the letter of the policy.

One potential issue may be that the Suicide and Self-Injury policy language focuses on content that attacks individuals, yet the post in question focuses on an attack against an entire community of people (in this case: transgender and non-binary people). This contrasts with Meta's Hate Speech policy, which explicitly covers "content targeting a person or group of people." The complications and difficulty for enforcement that arise from this inconsistency is clearly on display in the Case at hand. Initially the post was sent for human review under Meta's Suicide and Self-Injury policy and found to be non-violative. It was not escalated for human review under Meta's Hate Speech policy. Despite this, the post was ultimately removed based on its violation of Meta's Hate Speech policy.

In the absence of language that explicitly states how all of Meta's Community Standards, and Meta's Suicide and Self-Harm policy specifically, extends to posts/content targeting both individuals and/or communities, there is the potential for ambiguity that could impede enforcement, as this Case so clearly demonstrates. Revision of Community Standards to make this interpretation

explicit will hopefully be successful in reducing confusion and ensuring sufficient content moderation of undesired content.

Question 1: Speech, whether in spoken, written or visual form, that may be described by users as “humorous” or “satirical,” but which may spread hate speech or other forms of inflammatory rhetoric.

It is clear that content which encourages or celebrates suicide among a specific group meets Meta’s threshold for hate speech—including content which does so satirically, as seen in the present Case.

One of the Oversight Board’s stated seven strategic priorities focuses on Hate Speech Against Marginalized Groups, noting that “hate speech creates an environment of discrimination and hostility toward marginalized groups. It is often context-specific, coded, and with harm resulting from effects which gradually build up over time."

The post in this Case is a prime example of this coded nature, using innuendo and so-called ‘humor’ to celebrate the high rate of suicide in the transgender community. Allowing such images to accumulate creates a discriminatory environment for transgender people—much as, as the Oversight Board notes, “allowing images of blackface to accumulate would create a discriminatory environment for Black people.”

Meta’s Hate Speech policy further specifies multiple categories of banned content, including:

Dehumanizing speech or imagery in the form of comparisons[or] generalizations…[to] certain objects”

In this Case, comparing trans people to curtains, to mock the idea that trans people, like curtains, “hang themselves.”

“Harmful stereotypes historically linked to intimidation, exclusion, or violence”

In this Case, harmful stereotypes that trans people are mentally ill, which have been used by anti-trans hate groups to justify violence and discrimination against trans people, including by Polish political leaders.

“Exclusion in the form of calls to action, statements of intent, aspirational or conditional statements, or statements advocating or supporting…explicit exclusion”

In this Case, praising the exclusion and isolation of trans people out of Polish society (through their deaths) by equating this to “spring cleaning.”

It is not immediately clear why the post was not initially removed, given that its content violates so many of the stated tenants of Meta’s policy, but that it was not suggests that either Meta’s policy or its training of content moderators is insufficient.

Ironically, the issue may lie in the fact that the hateful conduct of this post was based on innuendo and ‘humor’ rather than explicit language and/or images: Whereas the lack of explicit clarification of included groups in Meta’s Suicide and Self-Harm policy created ambiguity, here, the lack of explicit guidance that Meta’s Hate Speech policy extends to coded language and dog whistles may have created confusion. Thus, we recommend that Meta revisit its Hate Speech policy to make clear that all hateful speech—whether innuendo or explicitly stated—will not be tolerated on Meta’s platforms.

Question 6: Meta’s account-level enforcement practices for users who repeatedly engage in anti-trans hate speech and harassment.

Currently, while posts are reviewed on an individual basis, a user’s posting history is not taken into account, resulting in violative posts (sometimes) being removed while the user remains on the platform to harass another day. This emerges in the present Case where, as the Oversight Board acknowledges, the

content in question was posted by a user who had previously been suspended from Facebook in response to their harassment of "members of the trans community online."

That this person was able to create another account—despite being removed for prior violations—highlights that enforcement is insufficient to remove known antagonizers. That the user's current account was allowed to stay on the platform, even after this specific post was removed, and despite their previous account being removed, further shows the limits of Meta's enforcement policies.

This is a feature, not a bug, of Meta's current policy. It is not until their seventh 'strike' that a user is restricted from posting content, and at that point, the restriction is for a single day. More than 10 violations result in only a 30-day restriction. Strikes themselves only count if a post is initially found to be in violation—meaning that content which is initially flagged but deemed non-violative due to moderator error and/or unclear policies will not count toward restriction.

And while a user may ultimately have their account disabled "after repeated warnings and restrictions," there is no clear definition for "repeated," nor is there any indication that this disabling results in a lifetime ban; the latter issue in particular is relevant if, as in the Case of the post in question, the user is able to create a new account and resume harassment with their strike count resetting to zero. Thus, it remains clear that Meta's enforcement practices are insufficient to address users/accounts who repeatedly engage in anti-trans hate speech and harassment.

Link to Attachment

PC-17029

2023-023-FB-UA	PC-17030	United States & Canada
Case number	Public comment number	Region
Subramaniam	Vincent	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

My brief comments are multifold, laid out here in parts.

A. On the content itself. "New technology. Curtains that hang themselves" on the transgender flag backdrop.

If this comment was meant to be satirical, it has to be "known" to the platform as a signal from the user, even if not marked on the content. In digital media, satire is already a problematic genre in terms of the impression it leaves on users. There are well-known cases of satire in news like content that has been misunderstood to be news. There are standardization initiatives in journalistic transparency ongoing, asking news publishers to label satire in their metadata for machines to know this. In the absence of any specific mechanisms to discover the intent to make a satirical or humorous post for UGC, this kind of post would have to be taken at face value for what it means.

B. On hate speech and suicide/self-injury policy: Did Meta consider this post as a possible case of dehumanizing speech?

The transgender community is a stakeholder in adjudicating the complaints made on this post. It is likely - I am not transgender and I cannot speak for them - that they will see this as dehumanizing speech, even if not hate. Dehumanizing speech is a more useful distinction to make here because the poster (given the account background - prior behavioral context) is likely to be signaling metaphorically using the "curtains hang themselves" analog that trans people ought to hang themselves. This is a signal to them that they do not have human worth and dignity; to live with their heads held high as equal humans like everyone else. That in turn is a form of judgment on whose humanity counts and whose does not. In that sense, it is dehumanizing.

C. Mechanisms: The mechanisms Meta has followed in handling the complaints and appeals are all fraught with one aspect being unclear. Did Facebook bring representative voices of the impacted community (stakeholder: transgender people) into the human review process?

The case says this: "One appeal resulted in a human reviewer upholding Meta's original decision that the content did not violate its Suicide and Self-Injury policy." To me, the human review seems to have been overly simplistic in its targeting the post for policy scrutiny. Suicide and Self-Injury policy and Hate Speech were not the only policies this post needs to be reviewed over. Dehumanizing speech, as noted in (B) is a more matched category for review.

Summary: Meta needs to explain how its human review process works if a marginalized community is the target of a post called into review. Were people representing that community included in the human review, before the decision was taken? I would request the Board to look into this aspect of how comprehensive the human review at the company is, w.r.t. to input from marginalized groups., before a piece of content is appealed to the board itself. If that was done is this case, was the category of dehumanizing speech considered? If so, what did the human reviewer write down as their justification? All these

things matter.

For instance, if transgender community members had given written input about this post to Meta's human review team, how would Meta have factored that input? These questions are not clear.

Thank you for posting this case. Transgender people are facing an increasingly hostile climate worldwide, not just in Poland or the United States. So reviewing this case as an example will help.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17031	United States & Canada
Case number	Public comment number	Region
Rasha	Younes	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Human Rights Watch		Yes
Organization		Response on behalf of organization

Full Comment

This case, which contains a clear example of hate speech, highlights the importance of reliable, transparent, and proactive content moderation practices that are consistently applied. Widespread hate speech targeting LGBT people on social media is not transient but has real offline consequences that reverberate throughout victims' lives.

In February 2023, Human Rights Watch published a report on the digital targeting of LGBT people in five countries across the Middle East and North Africa region. The report details how government officials across the MENA region are targeting LGBT people based on their online activity on social media, including on Meta platforms. In cases of online harassment, which took place predominantly in public posts on Facebook, affected individuals faced horrific offline consequences, which often ruined their lives.

In Poland, since the Law and Justice (PiS) party came to power in 2015, the

government has persistently attacked the rights of LGBT people in the context of its broader attacks on the rule of law. The government has deliberately undermined the independence of the judiciary and media freedom and sought to silence independent civil society groups, activists, and those who protest against its policies, including through the courts.

Hostile attitudes toward LGBT people found full expression in 2019 when regions and municipalities began to declare themselves "LGBT Ideology Free" or joined a government-supported Family Charter, calling for the exclusion of LGBT people from Polish society. More than 90 regional and municipal authorities have now declared themselves "LGBT ideology free" or signed the charter.

Since PiS came into power in Poland, LGBT activists have faced pressure and interference from the authorities over their peaceful activism, including arrests and criminal prosecutions, some under blasphemy laws. LGBT activists also reported the use by local authorities of what is known as Strategic Lawsuits Against Public Participation ("SLAPPs") to interfere with, and silence their work.

In addition to undermining the independent functioning of civil society, a clear rule of law violation, these measures have helped contribute to a hostile climate for LGBT people and activism in Poland.

Human rights apply online just as they do offline. Companies, including Meta, have a responsibility to respect human rights—including the rights to nondiscrimination, privacy, and freedom of expression—under the United Nations Guiding Principles on Business and Human Rights. The Guiding Principles require companies to "[a]void causing or contributing to adverse human rights impacts" and "[s]eek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services." The Guiding Principles further require businesses to be transparent about their policies, practices, and steps they take to identify, prevent, and mitigate human rights abuses.

International law permits legitimate restrictions on freedom of expression, including to ensure speech does not infringe on other people's rights. Content moderation should be carried out in a transparent, accountable and consistent manner.

Over-reliance on automation undermines Meta's ability to meet these requirements. In this instance, Meta's automated system failed to capture the severity of the complaint, rejecting all 12 complaints and 2 out of the 3 appeals without any human review.

In contexts where LGBT people face violence and discrimination, such as Poland, part of applying these human rights principles should compel Meta to invest in content moderation. Underinvesting in content moderation is especially detrimental to people who are marginalized, including LGBT people, who are disproportionately affected by the risks and harms stemming from content moderation.

The Santa Clara Principles on Transparency and Accountability in Content Moderation, which Meta has endorsed, provide helpful guidance. Meta should follow the Santa Clara Principles, including on training human content moderators on human rights and the adverse impacts for users of these platforms, including those that disproportionately affect LGBT people. Meta should also reassess its over-reliance on automation in assessing complaints.

Link to Attachment

No attachment

2023-023-FB-UA	PC-17032	United States & Canada
Case number	Public comment number	Region
Subramaniam	Vincent	English
Commenter's first name	Commenter's last name	Commenter's preferred language
DID NOT PROVIDE		No
Organization		Response on behalf of organization

Full Comment

Addendum to comments filed earlier today Sep 27, 2023.

1. Mechanisms/Due Process: There is a problem with Meta's due process in giving marginalized community stakeholders a chance to weigh in during human review. For instance, even if a group representing the transgender community submits their view internally that the Polish post carried dehumanizing speech, a fair process ought to offer the poster a chance to defend themselves or justify the meaning behind their post. Is there a policy to ask a poster (only in cases where content is flagged for human review) what they "meant" in the post, especially given any potential claims from petitioners about prior account history?

2. One of the problems here undoubtedly is scale. But any system that needs to take into account the efficacy of content moderation policy around human rights standards cannot operate at scale only for content distribution and throw

in the towel on human review. In this case, what if the mere provision of asking the poster to supply an intended meaning and justification results in a) their response text becoming more evidentiary around hate or further dehumanizing context OR b) what if the poster decided to take down the post themselves, reluctant to offer a justification at all? Is that a better outcome?

It would be helpful if the Board could a) find out what prior attempts have been made to complicate the human review process using some democratic/fair hearing principle, beyond strict application of policy and deciding one way or other, internally b) weigh in these approaches as the Board itself and make recommendations to Meta.

Link to Attachment

No Attachment

2023-023-FB-UA	PC-17033	United States & Canada
Case number	Public comment number	Region
Kayla	Gogarty	English
Commenter's first name	Commenter's last name	Commenter's preferred language
Media Matters For America		Yes
Organization		Response on behalf of organization

Full Comment

The Board has asked respondents for comments and recommendations on Meta's content moderation policies and enforcement practices against hate speech, and specifically that which targets LGBTQI+ people. The prevalence of anti-LGBTQI+ hate on Meta's platforms has remained a recurring problem and contributed to real-world harm, as the company has failed to consistently and adequately enforce its policies against such hate. Meta must better enforce and bolster these current policies to directly address anti-LGBTQI+ hate and harassment.

The case being deliberated by the Board — involving a post that falls between hate speech and the promotion of suicide or self-injury — speaks to Meta's broader content moderation issues around anti-LGBTQI+ hate, which have largely stemmed from the platform's inability to consistently and adequately enforce its policies. New Media Matters data reveals the extent to which anti-LGBTQI+ hate proliferates on the platform amid Meta's failures, which are

threefold: (1) There are loopholes that have exempted key purveyors of anti-LGBTQI+ hate from moderation; (2) the company allows networks of pages to amplify anti-LGBTQI+ hate with impunity; and (3) Meta fails to moderate accounts that repeatedly engage in anti-LGBTQI+ hate speech and harassment. These failures have contributed to real-world harm of LGBTQI+ people.

Meta has repeatedly chosen profit and positive press over the safety of its users, partially out of fear of relentless yet false claims from conservatives that they're being censored. As a result, Meta has repeatedly bent its rules, giving preferential treatment to right-wing media and politicians and carving out exemptions while inaccurate and harmful content — typically from right-leaning pages — dominates on the platform.

The company's failures, including with regard to anti-LGBTQI+ hate, have had real-world implications. Violence against trans people — especially trans women — has steadily worsened in the era of social media, with 2021 marking the deadliest year on record for trans people in the United States. In June 2023 alone, there were over 145 incidents of anti-LGBTQI+ hate and extremism in the United States. The most commonly cited trope in these incidents was the right-wing myth that LGBTQI+ people "groom" children. This trope has spread widely across Meta platforms, with the company even profiting from over 200 advertisements that use the "groomer" slur and push the anti-LGBTQI+ trope.

As the Board considers this case specifically and the broader harms of anti-LGBTQI+ hate and the promotion of self-harm on Meta's platforms, it should also take into account the prevalence of suicide among LGBTQI+ youth: Half of all trans and nonbinary young people considered suicide in the past year, according to a study from The Trevor Project, with 20% making an attempt to end their own lives.

Right-leaning pages dominate trans-related discussions on Facebook with content denying the existence of trans people and praising their exclusion.

Meta's hate speech policies governing Facebook and Instagram prohibit "violent

or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation." And yet, Media Matters has found that right-leaning pages that post anti-LGBTQI+ content — such as content denying the existence of trans people — remain on the platform and even dominate discussion related to trans people.

To assess the prevalence of anti-LGBTQI+ content on Facebook, Media Matters compiled and analyzed over 112,500 trans-related posts from U.S. news and politics pages since January 1, 2023, and found that right-leaning pages have dominated the conversation, accounting for nearly half (49%) of trans-related posts and earning 66% of total interactions on related posts. By comparison, left-leaning pages accounted for 13% of related posts and earned 20% of interactions while ideologically nonaligned pages accounted for 38% of related posts and earned only 14% of interactions.

The new Media Matters analysis also found that these posts from right-leaning pages often push anti-trans rhetoric. Nine of the 10 posts earning the greatest number of interactions came from right-leaning pages; these posts dismissed the existence of trans people, promoted the boycott of a company for affiliating with a trans woman, and praised the exclusion of trans people.

Previous research by Media Matters has also repeatedly shown that trans-related Facebook posts from right-leaning pages have overshadowed related posts from left-leaning and ideologically nonaligned pages.

A Media Matters study of 225 articles, blog posts, and videos about trans topics that earned 100,000 or more Facebook interactions and were posted from February 15, 2019, through February 15, 2020, found that two-thirds of the 66 million total Facebook interactions were earned by right-leaning sources such as right-wing anti-LGBTQI+ outlets The Daily Wire and LifeSiteNews.

Another study found that from October 2020 through September 2021, right-leaning pages posted about trans issues more often, earned more total

interactions, and earned more average interactions than either ideologically nonaligned or left-leaning pages. These posts were about trans athletes, state legislation, students and schools (including pronoun and bathroom use), and health care for trans youth.

Loopholes, exemptions, and a lack of consistent enforcement allows anti-LGBTQI+ hate to thrive on Meta's platforms.

Meta has a history of repeatedly bending its rules and giving preferential treatment to right-wing media and politicians. This preferential treatment has resulted in loopholes, exemptions, and enforcement failures that right-wing accounts on Meta's platforms have exploited to push anti-LGBTQI+ hate.

Loopholes have exempted key purveyors of anti-LGBTQI+ hate from moderation

Meta has specifically carved out exemptions for right-wing politicians: Politicians (and content quoting or showing political speech) are exempt from Meta's fact-checking program and are thereby able to spread harmful misinformation with no recourse. This allows right-wing politicians to post anti-LGBTQI+ hate and misinformation.

Media Matters' latest study on the prevalence of anti-LGBTQI+ hate on Facebook also found that at least 1,500 posts came from right-wing politicians, including posts misgendering trans people and denying their existence.

Additionally, Facebook's "cross check" or "XCheck" program exempted high-profile accounts from content moderation, including accounts that used the platform for harassment or incitement of violence. The program — which was "initially intended as a quality-control measure for actions taken against high-profile accounts," but which ultimately grew to include at least 5.8 million users in 2020 and covered "pretty much anyone regularly in the media or who has a substantial online following, including film stars, cable talk-show hosts, academics and online personalities with large followings" — likely covered the

10 right-leaning pages that Media Matters identified as earning the most interactions on posts related to LGBTQI+ issues since January 1. These pages included Ben Shapiro, Fox News, Matt Walsh, Breitbart, PragerU, Young America's Foundation, Newsmax, Daily Wire, Michael Knowles, and The Western Journal.

Media Matters has documented a plethora of instances in which right-wing media and personalities, including PragerU and Ben Shapiro's Daily Wire, benefited from Facebook algorithms or skirted Meta's policies.

Networks of Facebook pages amplify anti-LGBTQI+ hate with impunity

For years, right-wing media outlets have exploited Facebook's algorithms, promoting sensational content and amplifying it with networks of pages. Right-wing media have used these tactics to spread anti-LGBTQI+ hate.

Media Matters' latest study on the prevalence of anti-LGBTQI+ hate on Facebook demonstrates how these networks contribute to the proliferation of hate. Facebook pages affiliated with right-wing outlets the Daily Wire, TheBlaze, and the Western Journal (a combined total of 45 pages) were responsible for over 22,000 of the trans-related posts from U.S. news and politics pages, or 20%. These posts earned over 22.5 million interactions, accounting for nearly a quarter of all interactions earned on trans-related posts.

In one example, Media Matters found that a January 2023 article from the Western Journal's "news" section denied the existence of a transgender woman, claiming, "There's only one problem: Jakrajutatip is not a woman, but rather a 'transgender woman,' aka a man." Western Journal-affiliated pages posted this article on Facebook at least 27 times.

The Daily Wire's network of pages has similarly amplified anti-LGBTQI+ content. Media Matters found that a May 2023 article from the Daily Wire — which falsely claimed being transgender is a "social contagion" and compared it to other supposed "social contagions," such as people allegedly committing

suicide in a manner similar to characters in a novel, supposed uncontrollable dance fits in the 14th and 16th centuries, and an alleged "craze" of false allegations of sexual assault — was shared at least 9 times by the Daily Wire's network of pages.

In a previous analysis of the Daily Wire's trans-related posts from January 1, 2021, to March 13, 2023, Media Matters found that the Daily Wire's network of Facebook pages earned over 17 million interactions from nearly 13,000 posts — many of which amplified attacks on trans athletes, criticism of gender-affirming care, and praise for government officials restricting trans rights.

Accounts have repeatedly engaged in anti-LGBTQI+ hate speech and harassment

Meta has allowed numerous Instagram accounts with tens of thousands of followers to repeatedly target the LGBTQI+ community, despite the company holding policies against such content and publicly promoting its platforms as a safe space for LGBTQI+ users.

In a recent study, Media Matters found that the experienced right-wing personalities behind the anti-LGBTQI+ account "Gays Against Groomers" have repeatedly violated Instagram's policies by promoting the anti-LGBTQI+ "groomer" slur, claiming trans people have mental and moral deficiencies, and spreading related misinformation that's been debunked by Meta's third-party fact-checkers. But Meta's interpretation of its policies appears to be extremely narrow: In response to the study, Meta said that examples from the report that the company reviewed were "non-violating," even while claiming that "if someone were to use the term 'groomer' as an attack against someone based on being part of the LGBTQI+ community, it violates our hate speech policies." (Examples from the study did, in fact, use the term as an attack on members of the LGBTQI+ community.)

Another extreme anti-LGBTQI+ account, "Libs of TikTok," earned notoriety pushing the anti-LGBTQI+ "groomer" narrative in early 2022 and targeting schools, Pride events, and individuals on social media. On Instagram, Libs of

TikTok targeted a teacher and called them "sickening" for supposed "grooming behavior," and the account called a public library drag queen event an example of "tax dollars … funding the grooming of children."

Libs of TikTok's spread of "groomer" rhetoric contributed to the proliferation of these right-wing attacks, including on Facebook, where Media Matters found that between March 1 and April 18, 2022, U.S. news and politics pages posted at least 1,100 times mentioning "groomer," "grooming," or other related language, earning over 1.7 million interactions on these posts.

The cross-platform activity of these accounts, particularly Libs of TikTok, has been linked to real-world harm. Media Matters has documented the extensive harm connected to Libs of TikTok, which has baselessly maligned and targeted people, going so far as to reveal the names and locations of teachers, LGBTQI+ people, and others. This targeting and promotion of dangerous "groomer" rhetoric has resulted in harassment, threats, and lost livelihoods for private individuals. In fact, there have been a plethora of instances in which people, events, and places faced violent threats after being targeted in Libs of TikTok's social media posts. In August 2022, Facebook suspended Libs of TikTok for less than 24 hours following an anti-trans harassment campaign against Boston Children's Hospital which resulted in violent threats against health care providers. When the suspension was lifted, Libs of TikTok immediately returned to targeting children's hospitals that provide gender-affirming care for trans youth.

Despite the repeated violations, ties to real-world harm, and several previous suspensions, Gays Against Groomers and Libs of TikTok remain on Meta's platforms. Meta even claimed a recent suspension of one of the accounts was "platform error."

As the Board considers recommendations on Meta's moderation of hate speech, it must consider the prevalence of anti-LGBTQI+ hate and harassment on Meta's platforms, which has contributed to real-world harms for the LGBTQI+ community. These harms require Meta to more consistently enforce its current

policies, including by closing loopholes, removing exemptions, and taking action against accounts that repeatedly push anti-LGBTQI+ hate.

Link to Attachment

PC-17033