Gender Identity Debate Videos

(2024-046-FB-UA, 2024-047-IG-UA)

Summary

In two posts that include videos in which a transgender woman is confronted for using a women's bathroom and a transgender athlete wins a track race, the majority of the Board has upheld Meta's decisions to leave up the content. The Board notes that public debate on policies around transgender peoples' rights and inclusion is permitted, with offensive viewpoints protected under international human rights law on freedom of expression. In these cases, the majority of the Board found there was not enough of a link between restricting these posts and preventing harm to transgender people, with neither creating a likely or imminent risk of incitement to violence. Nor did the posts represent bullying or harassment. Transgender women and girls' access to women's bathrooms and participation in sports are the subjects of ongoing public debate that involves various human rights concerns. It is appropriate that a high threshold be required to suppress such speech. Beyond the content in these cases, the Board has made recommendations to address how Meta's January 7, 2025, revisions to the renamed Hateful Conduct Policy may adversely impact LGBTQIA+ people, including minors.

Additional Note: Meta's January revisions did not change the outcome in these cases, though the Board took the rules at the time of posting and the updates into account during deliberation. On the broader policy and enforcement changes hastily announced by Meta in January, the Board is concerned that Meta has not publicly shared what, if any, prior human rights due diligence it performed, in line with its commitments under the UN Guiding Principles on Business and Human Rights. It is vital Meta ensures adverse impacts on human rights globally are identified and prevented.

About the Cases

The first case involves a Facebook video in which an identifiable transgender woman is confronted for using the women's bathroom at a university in the United States. The woman who films the encounter asks the transgender woman why she is using the women's bathroom, also stating she is concerned for her safety. The post's caption describes the transgender woman as a "male student who thinks he's a girl," and asks why "this" is tolerated. This post has been viewed more than 43,000 times. Nine users reported the content, but Meta found no violations. One of those users then appealed to the Board.

In the second case, a video shared on Instagram shows a transgender girl winning a track race, with some spectators disapproving of the result. The caption names the athlete, who is a minor (under 18), refers to her as a "boy who thinks he's a girl" and uses male pronouns. This content, which has been viewed about 140,000 times, was reported by one user but Meta decided there was no violation. The user appealed to the Board.

Key Findings

The full Board has found neither post violates the updated Hateful Conduct policy. Considering the policy prior to Meta's January 7 changes, the majority did not find a violation under this version either because neither post contained a "direct attack" against people based on their gender identity, which is a protected characteristic. A minority, on the other hand, has found that both posts would have violated the policy's pre-January 7 version.

For the majority of the Board, neither post would have broken the rule against "statements denying existence," under the previous version of the policy. This rule was deleted in Meta's January update. Nor do the posts represent a "call for exclusion" because there are no calls for the transgender woman to leave the bathroom or for the transgender athlete to be ejected, disqualified from competition or otherwise left out. Prior to January 7, there were exceptions under Meta's internal guidance (not available publicly) to specifically allow calls for gender-based exclusion from sporting activities or specific sports, as well as from bathrooms. Since January 7, these exceptions are now made clear publicly in the Hateful Conduct rules, making these rules more transparent and accessible.

A minority of the Board disagrees, finding that both posts violated the pre-January 7 Hate Speech policy, including on "calls for exclusion" based on gender identity

and the (now deleted) rule on “statements denying existence.” The overall intent of these posts would have been clear: as direct and violating attacks that call for exclusion of transgender women and girls from access to bathrooms, participation in sports and inclusion in society, solely based on denying their gender identity.

On Bullying and Harassment, the Board finds by consensus no violation for the bathroom post since the adult transgender woman would have had to self-report the content for it to be assessed under the rules prohibiting “claims about gender identity” and “calls for … exclusion.” This type of self-reporting is not required for minors (aged between 13 and 18) unless they are considered by Meta to be a “voluntary public figure.” The majority of the Board agrees with Meta that the transgender athlete, who is a minor, is a voluntary public figure who has engaged with their fame, although for different reasons. For these Board Members, the athlete voluntarily chose to compete in a state-level athletics championship, in front of large crowds and attracting media attention, having already been the focus of such attention for earlier athletic participation. Therefore, additional protections under Tier 3 of the policy, including the rule that does not permit “claims about gender identity,” do not apply, and the majority finds no violation in the athletics post.

A minority disagrees, finding that the transgender athlete should not be treated as a voluntary public figure. Such public figure status should not be applied to a child because they have chosen to participate in an athletics competition that created media attention driven by their gender identity, which is not within their control. This should not equate to voluntarily engaging with celebrity. Therefore, this post violates the rule against “claims about gender identity,” as well as “calls for exclusion” under the Bullying and Harassment policy and should have been removed.

The Board is concerned about the self-reporting requirement under the Bullying and Harassment policy and its impact on victims of targeted abuse, making related recommendations.

For a minority of Board Members, the more troubling aspect is that both these posts meet the threshold of imminent risk of “discrimination, hostility or violence” against transgender people, under international human rights law, which requires

that this content be removed. The videos were posted against a backdrop of worsening violence and discrimination against LGBTQIA+ people, including in the United States. They deliberately attack and misgender specific transgender individuals as well as transgender people as a group, and in one case, involve the safety of a child.

Finally, the Board is concerned that Meta has incorporated the term "transgenderism" into its revised Hateful Conduct policy. For rules to be legitimate, Meta must frame them neutrally.

The Oversight Board's Decision

The Oversight Board upholds Meta's decisions to leave up the content in both cases.

The Board also recommends that Meta:

- In respect of the January 7, 2025, updates to the Hateful Conduct Community Standard, Meta should identify how the policy and enforcement updates may adversely impact the rights of LGBTQIA+ people, including minors, especially where these populations are at heightened risk. It should adopt measures to prevent and/or mitigate these risks and monitor their effectiveness. Finally, Meta should update the Board every six months on its progress, reporting on this publicly at the earliest opportunity.
- Remove the term "transgenderism" from the Hateful Conduct policy and corresponding implementation guidance.
- Allow users to designate connected accounts, which are able to flag potential Bullying and Harassment violations requiring self-reporting, on their behalf.
- Ensure the one report representing multiple reports on the same content is chosen based on the highest likelihood of a match between the person appealing and the content's target, guaranteeing that any technological solutions account for potential adverse impacts on at-risk groups.

*Case summaries provide an overview of cases and do not have precedential value.

Full Case Decision

1. Case Description and Background

These cases concern two posts containing videos shared on Facebook and Instagram in the United States in 2024.

The first case involves a video with a caption, shared on Facebook. A woman films an encounter in which she confronts an identifiable transgender woman for using the women's bathroom at a university. The caption refers to the transgender woman as a "male student who thinks he's a girl," while asking why "this" is tolerated. In the video, the woman asks the transgender woman why she is using the women's bathroom, challenges her on her gender and states that she "pay[s] a lot of money to be safe in the bathroom." The transgender woman responds that she is a "trans girl" and that safety in the bathroom is important to her too. The post has been viewed about 43,000 times. Nine users reported the post for hate speech and bullying and harassment, but Meta found the content was not violating. One of those users appealed to the Board.

In the second case, a video shared on Instagram shows a transgender girl winning a girls' state-level track championship race, with some spectators disapproving of the result. The caption identifies the teenage athlete by name, referring to her as a "boy who thinks he's a girl," as well as using male pronouns. The post has been viewed about 140,000 times. One user reported the content for hate speech and bullying and harassment, but Meta determined the content was not violating. The user appealed Meta's decision to the Board.

The Board's review of these cases comes at a time of significant public debate in certain parts of the world about the rights of transgender women and girls. In the United States, these debates intensified during the 2024 Presidential Election. The new U.S. administration is enacting policy changes directly affecting the rights of transgender people. Those who support broader freedom of expression for debate around these issues do not necessarily support the policy changes being enacted, many of which are also adversely impacting freedom of expression and access to information.

On January 7, 2025, Meta announced revisions to its Hate Speech policy, renaming it the Hateful Conduct policy. These changes, to the extent relevant to these cases, will be described in Section 3 and analyzed in Section 5. The Board notes content is accessible on Meta's platforms on a continuing basis, and updated policies are applied to all content present on the platform, regardless of when it was posted. The Board therefore assesses the application of policies as they were at the time of posting and, where applicable, as since revised (see also the approach in Holocaust Denial).

2. User Submissions

The user who appealed the content (bathroom post) in the first case to the Board explained that Meta is allowing what is, in their view, a transphobic post to stay on its platform. The user who appealed the athletics post in the second case said it attacks and harasses the athlete who is a minor and violates Meta's Community Standards. Neither of the users who appealed to the Board appear in either post under review. The users who shared both posts were notified of the Board's review and invited to submit a statement, but none were received.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Hateful Conduct (previously named Hate Speech) Community Standard

According to the Hateful Conduct policy rationale, Meta doesn't allow hateful conduct (previously hate speech) on its platforms because the company "believe[s] that people use their voice and connect more freely when they don't feel attacked on the basis of who they are." Meta defines "hateful conduct" in the same way it previously defined "hate speech" as "direct attacks against people" on the basis of protected characteristics, including sex and gender identity. It does not generally prohibit attacks against "concepts or institutions."

Following Meta's January 7, 2025, update, the policy rationale states that Meta's policies are designed to "allow room" for various types of speech, including for people to use "sex- or gender-exclusive language" when discussing "access to

spaces often limited by sex or gender, such as access to bathrooms, specific schools, specific military, law enforcement or teaching roles, and health or support groups." It recognizes that people "call for exclusion or use insulting language in the context of discussing political or religious topics, such as ... transgender rights, immigration or homosexuality."

In the same update to the Hateful Conduct policy, Meta removed various Tier 1 prohibitions (the violations considered most severe), including the rule against "statements denying existence, including but not limited to claims that protected characteristic(s) do not or should not exist, or that there is no such thing as a protected characteristic."

Under Tier 2 of the Hateful Conduct policy, Meta continues to prohibit "calls or support for exclusion or segregation or statements of intent to exclude or segregate" on the basis of protected characteristics, including sex or gender identity, unless otherwise specified. Meta prohibits "social exclusion," defined as "denying access to spaces (physical and online) and social services, except for sex or gender-based exclusion from spaces commonly limited by sex or gender, such as restrooms, sports and sports leagues, health and support groups, and specific schools." Prior to the January 7 update, this exemption was narrower, specifying only "gender-based exclusion in health and positive support groups." At the time the posts were first reviewed, Meta's internal guidance to reviewers specified that calls for exclusion from sporting activities or specific sports were permitted. However, calls for exclusion from bathrooms were permitted only on escalation. When content is escalated, it is sent to additional teams within Meta for policy and safety review. Meta's January 7 changes have made both of these previously unpublished exceptions public and turned the bathroom exception from escalation-only to the default at-scale meaning that all human reviewers are instructed to leave content up, without requiring escalation to an internal team at Meta.

The updated Hateful Conduct policy also now exempts from its prohibition on "insults" (described under the previous policy as "generalizations that state inferiority") any "allegations of mental illness or abnormality when based on gender or sexual orientation, given political and religious discourse about transgenderism and homosexuality and common non-serious usage of words like 'weird.'"

Bullying and Harassment Community Standard

The rationale for the Bullying and Harassment policy states that "bullying and harassment happen in many places and come in many different forms from making threats and releasing personally identifiable information to sending threatening messages and making unwanted malicious contact." The Bullying and Harassment Community Standard is split into four tiers, with Tier 1 providing "universal protections for everyone," and Tiers 2 – 4 providing additional protections, limited according to the status of the targeted person. Meta distinguishes between public figures and private individuals "to allow discussion, which often includes critical commentary of people who are featured in news or who have a large public audience." For private individuals, the company removes "content that's meant to degrade or shame." In certain instances, self-reporting is required because it helps the company understand whether the person targeted actually feels bullied or harassed. The policy rationale also states that Meta recognizes "bullying and harassment can have more of an emotional impact on minors, which is why the policies provide heightened protection for anyone under the age of 18, regardless of user status."

Tier 3 of the policy prohibits "claims about ... gender identity" and "calls for ... exclusion." Private adults who are targeted by such claims must report the violating content themselves for it to be removed. Self-reporting is not required for private minors and minors who are considered involuntary public figures. A minor who is a voluntary public figure and all adult public figures are not protected under Tier 3 of the Bullying and Harassment policy, even if they self-report.

The policy rationale defines public figures, among others, as "people with over one million fans or followers on social media and people who receive substantial news coverage," as well as government officials and candidates for office. Meta's internal guidelines define "involuntary public figures" as: "Individuals who technically qualify as public figures but have not engaged with their fame."

II. Meta's Submissions

Meta kept both posts on Facebook and Instagram, finding neither post violated its Hateful Conduct (previously named Hate Speech) or Bullying and Harassment policies. It confirmed that this outcome was not impacted by its January 7 policy

changes. The Board asked questions on the scope and application of these policies and Meta responded to all of them.

Bathroom Post

Meta determined the bathroom post in the first case did not violate the Hateful Conduct policy.

First, it did not constitute a "call for exclusion" under the Hate Speech policy because it was ambiguous whether it was questioning the transgender woman's presence in the specific bathroom or the broader policy of allowing transgender women in women's bathrooms. Meta noted that "removing indirect, implicit, or ambiguous attacks would interfere with people's ability to discuss concepts or ideas on its platforms," in this case the concept of transgender women using women's bathrooms. Meta explained that, following the January 7 update, it now considers calls for exclusion from bathrooms on the basis of sex or gender to be permissible. In its view, this update to the public-facing language improved transparency and simplified enforcement of this rule. Second, the post did not violate the (now deleted, and no longer applicable) Tier 1 rule on denying the existence of a protected characteristic group. Meta does not consider the post describing the depicted transgender woman as male (i.e., misgendering) to deny the existence of transgender people. Meta stated that it did not equate a statement denying that a person belongs to a protected characteristic group with denying the existence of that group.

Meta also concluded the bathroom post did not violate the Bullying and Harassment policy because the transgender woman targeted in the post did not report the content herself. Meta clarified that the prohibition on "claims about gender identity" prohibits misgendering, and had the targeted person self-reported, it would have been found violating. However, even if the user had self-reported, Meta would have found the rule against "calls for exclusion" not violated, as there was no explicit call for exclusion.

In response to the Board's questions, Meta stated that it has considered alternatives to the self-reporting requirement, but they present risks of overenforcement. Meta explained it would be difficult to define the appropriate level of relationship between a targeted person and a third-party reporting on

their behalf. It added it would be challenging to validate the accuracy of the information provided.

In response to the Board's questions, Meta explained that the company does not remove content solely because it contains footage of an identifiable person without consent in a private setting, as an additional violating element is required. This is because, "while private settings present different risks from public ones, many non-private activities and speech occur in private settings."

Athletics Post

Meta concluded the athletics post in the second case did not violate the Hate Speech (now Hateful Conduct) policy.

First, Meta found there was no prohibited call for exclusion. For Meta, the way the post draws attention to the spectators' disapproval of the transgender girl's victory may be directed at the "concept" of allowing transgender girls and women to compete in sporting events consistent with their gender identity. Meta explained the updated Hateful Conduct policy now publicly clarifies that social exclusion does not include "sex or gender-based exclusion from spaces commonly limited by sex or gender, such as … sports and sports leagues," which was previously enforced through an exception in the internal guidance to reviewers.

Second, for the same reasons as the bathroom post, Meta found this post did not violate the (now deleted) Tier 1 rule on denying the existence of a protected characteristic group.

Meta also concluded that this post did not violate the Bullying and Harassment Community Standard. Meta found it did not contain a "call for exclusion" and that although the athlete was a minor (aged between 13 and 18), she was a "voluntary public figure" because she had engaged with her celebrity. She was therefore not protected from the Tier 3 prohibition on "claims about gender identity" (which prohibits misgendering an individual). Had she not been classified as a voluntary public figure, the content would have violated the rule on "claims about gender identity." In that instance, as she is a minor, she would not have had to self-report the content for a violation to be found.

In Meta's analysis, the company considered the targeted minor a "public figure," given the significant news coverage about her as an athlete, and that she "may have capacity to influence or communicate to large groups of individuals." Meta explained that the company allows "more discussion and debate around public figures in part because – as here – these conversations are often part of social and political debates and the subject of news reporting." They said that "athletes who enter competitions and generate news coverage, for reasons positive or negative, automatically become public figures when they appear in a specified number of news articles." Meta also clarified that minors under the age of 13 cannot qualify as public figures. The transgender athlete in this case, who was not under 13 but is a minor, was a "voluntary public figure" because she had in Meta's view, "to some extent," engaged with her fame, "speaking publicly about" their transition, to a school newspaper in 2023. Through the distinction between minors who are "voluntary" or "involuntary" public figures, Meta "seeks to balance the safety of minors with their right to agency, expression, and dignity through, for example, choosing to engage with their celebrity, including the notoriety that may come with it." The company explained "this approach respects the rights of minors by allowing the public to discuss minors who have voluntarily engaged with their fame while restricting potentially harmful negative attention directed toward[s] minors who have become famous because they are victims of crime or abuse."

Meta added that, even if either post had violated its content policies, they would still have been kept up under the newsworthiness allowance, upon escalated review. This is because both posts relate to topics of considerable political debate in the United States, and the facts underpinning the post about the transgender athlete who is a minor were subject to significant news coverage.

4. Public Comments

The Oversight Board received 658 public comments that met the terms for submission. Of these comments, 53 were submitted from Asia Pacific and Oceania, 174 from Europe, 8 from Latin America and the Caribbean, one from Sub-Saharan Africa and 422 from the United States and Canada. Because the public comments period closed before January 7, 2025, none of the comments address the policy changes Meta made on that date. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: immutability of biological traits; research into harms of misgendering or exclusion of transgender people; risks of under and overenforcement of content involving transgender people; the self-reporting requirement and the status of the involuntary public figure, who is a minor, under Meta's Bullying and Harassment policy; and, the impact of the participation of transgender women and girls in sports and women's bathrooms on women's rights.

5. Oversight Board Analysis

The Board selected these cases to assess whether Meta's approach to moderating discussions about gender identity respects the human rights, including freedom of expression, of all people. The Board analyzed Meta's decisions in these cases against Meta's content policies, values and human rights responsibilities. The Board also assessed the implications of these cases for Meta's broader approach to content governance.

5.1 Compliance With Meta's Content Policies

I. Content Rules

Hateful Conduct (previously named Hate Speech) policy

Following the January 7 policy changes, the Board finds neither post violates Meta's Hateful Conduct policy. A violation consists of two elements: (i.) a "direct attack" in the form of prohibitions listed under the "Do not post" section of the policy; (ii.) that targets a person or group on the basis of a listed protected characteristic. For both posts, the absence of a "direct attack" under the revised rules means there is no violation. The Board notes that "gender identity" remains a protected characteristic under Meta's Hateful Conduct policy.

Prior to the January 7 policy changes, the Board assessed both posts against two prohibitions (i.e., "direct attacks") within the Hate Speech policy: (i.) statements denying the existence of transgender people or identities; (ii.) calls for social exclusion of transgender people.

The majority of the Board found that neither post violated Meta's rule (now deleted and no longer enforced) on "statements denying existence." For this rule to have been violated, the content would have needed to include a more categorical statement that: transgender people or transgender identities do not exist; that no one is transgender; or, that anyone who identifies as transgender is not. Both posts refer to the biological sex of the individuals in the videos to say they "think" they are female. While this may show disregard for these individuals' gender identities and may be rude or offensive to many, it does not amount, even by inference, to a statement that transgender people or identities do not exist. One might infer from the posts a rejection of the idea that gender identity, rather than biological sex, should determine who can participate in women's and girls' sports or access women's bathrooms. The expression of this opinion, however controversial, did not violate this rule in the Hate Speech policy.

A minority of the Board found that both posts violated Meta's previous rule on "statements denying existence." For a minority, the assertions in both video captions that the depicted people are males "who think they are females," without explanation or qualification, categorically reject the possibility that transgender women and girls are or can be anything other than male. The language and tone, while implicit, seek to characterize all transgender identities as a delusion, rather than as an identity. For this minority, finding a violation would be consistent with Board precedent recognizing how indirect narratives or "malign creativity" in statements can constitute hate speech (see Holocaust Denial and Post in Polish Targeting Trans People).

The Board notes that Meta's prohibition on calls for social exclusion is retained in the January 7 policy update, but in addition to allowing gender-based exclusion from "health and support groups," the policy now allows exclusion based on sex or gender from "spaces commonly limited by sex or gender, such as restrooms, sports and sport leagues." The policy rationale was also updated to recognize that Meta seeks to permit sex- or gender-exclusive language on these issues.

For the majority of the Board, neither post constituted a call for social exclusion under the Hate Speech policy prior to these changes. In the bathroom post, there is no call for the transgender woman to leave the facility, be involuntarily removed, or be excluded in future. Rather the person recording asks the

transgender woman, “Do you think that’s OK?” While the conversation may have been unwelcome and rude, it does not meet the plain definition of a “call for exclusion.” In the athletics post, there is no call for the transgender athlete to be ejected, disqualified from competition or otherwise left out. The post depicts her participation and victory, implicitly elevating a question as to whether it is fair. Debating the validity of various approaches to transgender athletic participation or questioning the eligibility of a single athlete does not amount to a call for social exclusion, in violation of Meta’s policy. The majority of the Board notes that prior to January 7, Meta’s internal guidance to reviewers included instructions to allow calls for gender-based exclusion from sporting activities or specific sports, and for decisions made by Meta’s internal policy teams, to allow calls for gender-based exclusion from bathrooms. Making Meta’s rules more transparent and accessible, as the January 7 amendments do in this area, is generally welcome.

For a minority, both posts, understood in context (see the minority’s human rights analysis in Section 5.2), constituted prohibited “calls for exclusion” based on gender identity. That context, taken together with the statements denying the existence of transgender identity by characterizing it as a delusion, makes the overarching intent of these posts as a direct and violating attack clear: the exclusion of transgender women and girls from access to bathrooms, participation in sports and inclusion in society, solely based on denying their gender identity. Finding a violation of this rule was consistent with Meta’s Hate Speech policy rationale, which previously stated that hate speech was not permitted because “it creates an environment of intimidation and exclusion, and in some cases may promote offline violence.” For the minority, the January 7 policy changes are not in line with Meta’s human rights responsibilities, which require the removal of both posts (see Section 5.2).

Bullying and Harassment Policy

The Bullying and Harassment Community Standard was not revised on January 7.

In the first case, the bathroom post, the Board finds by consensus that, since the transgender woman is an adult and not a public figure, she would have had to self-report the content for Tier 3 of the Bullying and Harassment policy to be assessed, including the rules on “claims about gender identity” and “calls for

exclusion." As the transgender woman in the video did not report the content herself an analysis of Tier 3 of the policy is not necessary.

While the Board acknowledges that self-reporting may assist Meta in ascertaining if a targeted person feels bullied or harassed, the Board is concerned about the practical challenges and additional burden on users to report harassing content under the Bullying and Harassment policy. Public comments submitted to the Board (see PC-30418 and PC-30167) and various reports highlight the shortcomings of the self-reporting requirement and its impact on victims of targeted abuse. Moreover, the changes that Meta announced on January 7, which were explicitly designed to reduce automated detection of "less severe policy violations," could increase this burden. In this regard, Meta should continue to explore how to reduce this burden on targets of bullying and harassment, for example by allowing trusted representatives to report with their agreement and on their behalf.

Relatedly, when Meta requires users to self-report under certain policy lines, these reports should be effectively prioritized for review to ensure accurate enforcement of these policies. As the Board previously explained in the Post in Polish Targeting Trans People decision, Meta's automated systems monitor and deduplicate multiple reports on the same piece of content to "ensure consistency in reviewer decisions and enforcement actions." In the Board's understanding, this may result in omissions of self-reports where there are multiple user reports. The Board, therefore, recommends that Meta should ensure that self-reports from users are prioritized for review, guaranteeing that any technological solutions implemented account for potential adverse impacts on at-risk groups (see Section 7).

While the minority acknowledges that Meta's rules require private adults to self-report bullying and harassment violations, these Board members are concerned about Meta's general analysis of the setting of the bathroom post. Confronting a transgender woman in a bathroom is an invasive act that should be considered a form of "harassment." This was not a "non-private activity," but an invasion of a person's privacy.

In relation to the athletics post in the second case, the Board notes that Tier 3 of the Bullying and Harassment policy does not protect people between the ages of

13 and 18, who are public figures, and who have "engaged with their fame." According to Meta, this engagement distinguishes voluntary public figure status from involuntary status. The Board agrees that Meta was wrong to categorize the minor transgender athlete as having "engaged with" her fame (and therefore as a *voluntary* public figure) solely on the basis that she participated in an interview with a school newspaper a year before the athletics competition shown in the video took place. This was not a sufficient basis for Meta to demonstrate agency on the part of the child for voluntarily becoming a public figure.

The majority finds that the depicted athlete qualifies as a voluntary public figure who is a minor by virtue of her choice to compete in a state-level athletics championship. Such state-level competitions garner wide attention, take place in front of large spectator crowds and are often covered by the media to generate attention. The choice to perform in a high-profile sporting event, particularly after already being the focus of media reporting for her earlier athletic participation, is a voluntary decision by the transgender athlete. For the majority, Meta properly recognizes "minor voluntary public figures" on the basis that they are exercising agency, expression and dignity through their choice to shape a public identity. With older children participating in high-level sporting competitions, active in the entertainment industry, influential on social media and occupying other prominent public roles, such recognition of personal agency and expressive rights is appropriate.

A minority finds that the transgender athlete should not be considered a voluntary public figure. At most, she should be treated as an involuntary public figure and be afforded all the protections of the Bullying and Harassment policy, including Tier 3. These Board Members disagree with basing a "public figure" status, especially of a child, solely on an arbitrary number of online media references to them. Such media coverage does not, in itself, turn a child into a public figure, nor should it be the basis for a reduction in the protections she receives. Endorsing this approach is inconsistent with the Sharing Private Residential Information policy advisory opinion and is especially concerning when applied to a minor. A child's choice to participate in a state-level athletics competition should not be equated to *voluntarily* engaging with their apparent celebrity, especially when media coverage has been driven by the minor's gender identity, which is not within their control. While the athlete participated in this event knowing she may attract attention, that is not the same as having agency

and the freedom of expression to engage with the media attention that followed. There is no indication that the minor sought to engage with this apparent fame or actively participated in the media attention she received.

Under the Tier 3 Bullying and Harassment rules, a minority finds that the athletics post violates the prohibition on "claims about gender identity." These Board Members agree with Meta that "claims about gender identity" include misgendering. This post directly states that the transgender athlete is a "boy who thinks he's a girl" and uses male pronouns. In the minority's view, these are claims about gender identity targeting an identifiable child to harass and bully them, and as such violate the policy.

For a minority, the post also violates the Tier 3 Bullying and Harassment prohibition on calls for exclusion for the same reasons it violated the similar prohibition on calls for exclusion under the previous Hate Speech policy. The transgender athlete is clearly identifiable and named in the post.

For the majority, as the athlete was a voluntary public figure, Tier 3 of the Bullying and Harassment policy does not apply, and analysis of potential violations is therefore not necessary.

5.2 Compliance With Meta's Human Rights Responsibilities

The majority of the Board finds that keeping both posts on the platforms was consistent with Meta's human rights commitments. A minority of the Board disagrees, finding that Meta has a responsibility to remove both posts.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including views about politics, public affairs and human rights (General Comment No. 34, paras. 11-12). The UN Human Rights Committee has highlighted that the value of expression is particularly high when discussing political issues (General Comment No. 34, paras. 11, 13; see also para. 17 of the 2019 report of the UN Special Rapporteur on freedom of expression, A/74/486). When restrictions on expression are imposed by a state they must meet the requirements of legality, legitimate aim,

and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test."

The Board uses this framework to interpret Meta's human rights responsibilities in line with the UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression" (A/74/486, para. 41).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and must "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (*Ibid.*). The UN Special Rapporteur on freedom of expression has stated that, when applied to private actors' governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). People using Meta's platforms should be able to access and understand the rules, and content reviewers should have clear guidance regarding their enforcement.

The Board finds that in relation to the updated Hateful Conduct rules as applied in these cases, the legality standard is satisfied, as those rules are clear and accessible.

II. Legitimate Aim

Any restriction on expression should pursue one of the legitimate aims of the ICCPR, which include protecting the "rights of others."

In several decisions, the Board has found that Meta's Hate Speech (renamed Hateful Conduct) policy aims to protect the rights of others (see Knin Cartoon.) The Hateful Conduct policy rationale still states that Meta believes "that people use their voice and connect more freely when they don't feel attacked on the basis of who they are." The Hate Speech policy previously noted that the company prohibited hate speech because "it creates an environment of intimidation and exclusion, and in some cases may promote offline violence."

The Board has previously found that the Bullying and Harassment Community Standard also aims to protect the rights of others, noting that "users' freedom of expression may be undermined if they are forced off the platform due to bullying and harassment," and that "the policy also seeks to deter behavior that can cause significant emotional distress and psychological harm, implicating users' right to health," (see Pro-Navalny Protests in Russia). In respect of children, respecting the best interests of the child (Article 3 UNCRC) is additionally important (see Iranian Make-up Video for a Child Marriage).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected," (General Comment No. 34, para. 34).

The Board notes that public debate on policy issues around the rights of transgender people and inclusion must be permissible. The Board agrees that international human rights principles on freedom of expression protect offensive viewpoints (see UN Special Rapporteur on freedom of expression, report A/74/486, at para. 24). To justify a restriction on speech, a direct and immediate connection between the speech limited and the threat should be demonstrated in a specific and individualized fashion (General Comment No. 34, op. cit., at para. 35). For these cases, Board Members disagree on the nature and degree of harm the two posts posed, and therefore what limitations were necessary and proportionate.

The majority of the Board finds that neither post creates a likely or imminent risk of incitement to violence, so there is an insufficient causal connection between restricting these posts and preventing harm to transgender people. This also means that there is no affirmative responsibility for Meta to prohibit these posts (e.g., under Article 20, para. 2 of the ICCPR).

For the majority, issues of transgender women and girls' access to women's bathrooms and participation in sports are subjects of ongoing public debates (see PC-30308) that implicate a range of human rights concerns. As The Future of Free Speech organization argues, an "overly restrictive application of Meta's policies can create a chilling effect" on individuals that "may refrain from participating in discussions on gender identity for fear of their views being labeled as hate speech or harassment," and "marginalize voices that seek to challenge or critique prevailing norms around gender, which is essential for a vibrant democratic society."

It is therefore appropriate that a high threshold be demonstrated to justify any restriction, to avoid precluding public discourse and impairing understanding of these issues. The majority acknowledges that amid the intensity of these debates, these posts have the potential to be deeply offensive and even hurtful. However, the UN Special Rapporteur has stated: "Expression that may be offensive or characterized by prejudice and that may raise serious concerns of intolerance may often not meet a threshold of severity to merit any kind of restriction. There is a range of expression of hatred, ugly as it is, that does not involve incitement or direct threat, such as declarations of prejudice against protected groups. Such sentiments would not be subject to prohibition under the International Covenant on Civil and Political Rights ... and other restrictions or adverse actions would require an analysis of the conditions provided under article 19 (3) of the Covenant," (report A/74/486, at para. 24).

For the majority, it follows that suppression of speech voicing viewpoints that are hateful or discriminatory, but below the incitement threshold, as in these cases, would not make any alleged underlying prejudice disappear. Rather, people with those ideas are driven to other platforms, often with like-minded people rather than a broader range of individuals, an outcome that may exacerbate intolerance, instead of enabling a more transparent and public discourse about sensitive issues. That the posts are not respectful is not grounds for suppressing speech.

The Board has often used the Rabat Plan of Action's six-factor test to assess whether content qualifies as incitement under the terms of the ICCPR and to establish a high threshold for restrictions: the social and political context; the status of the speaker; intent to incite audience against a group; the content and form of the expression; the extent of its dissemination; and, the likelihood of harm, including imminence. For the majority, the following combined factors demonstrate that Meta has no positive responsibility to remove the two posts:

- **Context:** The majority is cognizant that transgender people face discrimination, harassment and even violence in many parts of the world, including the United States. While the pitched tenor of these debates elevates risks for transgender people, it does not follow that posts discussing related policy issues, even when using coarse or insensitive language, will themselves incite discrimination or violence (see Myanmar Bot). The majority nevertheless emphasizes that it is important for Meta to be mindful of context, as changes in civic freedoms and/or policy changes impacting equality protections can create an environment where violence and discrimination may be more easily incited.

- **Content and form**: The majority acknowledges that hate speech and incitement can be implicit. However, neither post was characterized by an explicit, implied, or indirect call for others to engage in violence or discriminatory acts, such as harassment or threats, against the transgender individuals in the videos, nor against transgender people broadly.

- **Intent:** The majority find that the two posts contain no hidden or contrary indicators of intent advocating harm. To the extent the posts can be interpreted as advocating for the exclusion of transgender women and girls from women's bathrooms or certain competitive sporting events, international human rights principles on non-discrimination do not prohibit such access or participation being based on biological sex. As such, speech advocating that outcome, without more, cannot be considered impermissible incitement under the ICCPR. To find otherwise would, in the view of the majority, severely limit freedom of expression for people who believe that biological sex should remain a determinative categorization in certain contexts, notwithstanding individuals' gender identities. The ongoing policy

debates within sports leagues about how to best ensure fairness to all in terms of the participation of transgender athletes illustrate the ongoing nature of these dialogues and the impracticability of suppressing certain viewpoints, or associating them with an intent to advocate harm.

- **Speaker's status and extent of dissemination**: Additionally, in the majority's view, the speaker status and the extent of the posts' dissemination, do not change the assessment that these posts did not rise to the level of incitement. The account creator is influential in online discourse and known for sharing intentionally provocative content and for spreading harsh anti-transgender sentiment. That said, they do not occupy a position of formal authority or equivalent over others to the extent that general statements of opinion would be interpreted as an instruction or calls for others to act.

- **Likelihood and imminence of acts of violence, discrimination and hostility**: Lastly, and as a result of the assessment of each of the factors above, the majority finds that acts of discrimination or violence were not likely or imminent as a result of these two posts to the transgender people in the videos, or more broadly.

The majority notes that the Rabat Plan also calls for positive initiatives that do not infringe on freedom of speech to promote tolerance and inclusion, including encouraging counter-speech, such as the forceful condemnation of offensive or degrading speech. Education, information, dialogue and storytelling to foster dialogue can help drive forward these debates in a constructive way that avoids denigration and discrimination, and social media companies can play their part. There may also be less intrusive means available to Meta to address concerns around intolerance short of content removal such as removal of posts from recommendations or limits on interactions or shares.

In relation to Bullying and Harassment, the majority note that Meta's policies in this area pursue different objectives to the Hateful Conduct policy and are focused on reducing harms to targeted individuals.

However, the Bullying and Harassment prohibitions are potentially very broad in their application and could sweep up speech that is self-referential, satirical, or culturally specific. Meta mitigates the risk of over-enforcement by requiring self-

reporting for some violations and exempting public figures from protection against lower severity violations. While the self-reporting tools are limited, they are an appropriate mechanism for ensuring a targeted individual actually feels attacked before action on that content is taken. As noted in Section 5.1, the Board has doubts about the criteria Meta applied in designating the teen in the second case as a "voluntary public figure." However, as applied to this post, the athlete would have understood that her participation in this level of competition would attract attention because of her transgender identity. It is, for the majority, consistent with the Convention on the Rights of the Child to consider an older teen's autonomy and evolving capacity to take decisions. As such, the majority finds the athlete could reasonably expect to receive critical commentary about their biological sex. Waiving protections under Tier 3 of the Bullying and Harassment policy recognizes that agency, as well as the public interest in the speech at issue, and does not violate the principle of upholding the best interests of the child.

Some Board Members who support the majority position note that Meta's human rights responsibilities provide the company with a degree of discretion to take a stance on social issues. For these members, the Board's prior relevant cases around hateful content (see Depiction of Zwarte Piet and South Africa Slurs) mean it would be within Meta's discretion to take a more restrictive stance against the misgendering of transgender people or other use of gender- or sex-exclusive language. In doing so, they should provide clear and accessible policies to this effect, provided they are enforced consistently and fairly. However, Meta's human rights responsibilities do not require it to adopt this position. Here, Meta has chosen to provide limited protections for individuals against misgendering in the Bullying and Harassment policy. It has taken steps to prevent overreach by requiring self-reporting, and by creating the public figure criteria to allow discussion of individuals in the news. For this reason, these Board Members also uphold Meta's decisions not to remove either post.

For the minority, Meta's decisions to leave up both posts contradicts its human rights responsibilities.

The minority notes that rules to address the harms of hate speech and bullying and harassment are consistent with freedom of expression because they are essential to ensure that vulnerable minorities can express themselves, including

their gender identities. Meta seeks to provide an expressive space to LGBTQIA+ people to maximize diversity and pluralism (see UN Independent Expert on Sexual Orientation and Gender Identity, report A/HRC/56/49, July 2024, at para. 7 and 66).

Meta has a specific and additional responsibility to remove from its platforms any advocacy of hatred against LGBTQIA+ people that constitutes incitement to discrimination, hostility or violence (Article 20, para 2, ICCPR; report A/74/486, at para. 9). However, for the minority Meta's Hateful Conduct policy exists to limit the use of language that contributes to an environment that makes discrimination and violence more acceptable and therefore sets a different threshold in terms of intent and causation. In this way, this policy is distinct from Meta's Violence and Incitement policy. Even so, in these two cases, a minority find that the incitement to discrimination threshold was met, as demonstrated under the Rabat Plan of Action:

- **Context**: Crucially, violence and discrimination against LGBTQIA+ people, especially transgender people, is worsening, globally and in the United States. Transgender people are four times more likely to suffer violent crime than others. In the US, in 2023, more than 20% of hate crimes were motivated by anti-LGBTQIA+ bias (2023 Hate Crime Statistics of the US Federal Bureau of Investigations). In 2024, over 30 violent killings (Human Rights Campaign) and at least 447 incidents directly targeting trans and non-binary people (GLAAD) were documented. Transgender people are frequently subject to harassment, abuse and threats online and offline, with consequences including bullying, isolation, increased suicide rates and violence (see PC-30338, PC- 30409; also Report A/74/181 at para. 32; Report A/HRC/56/49, at para. 7 and 66). In the US, and globally, governments are legislating to ban access to healthcare and to remove the rights of transgender people to be legally recognized and participate in society. In this highly inflammatory context, Meta must take additional care to ensure that its services are not used in ways that contribute to an environment of hostility that makes further harm more likely.

- **Speaker's status and extent of dissemination**: The sharing account has a large following on Facebook and Instagram that wields substantial public influence and is known for its anti-LGBTQIA+ stances. It is committed to spreading inflammatory material that challenges what it pejoratively

describes as transgender 'ideology.' These posts received more than 180,000 views and reactions, and hundreds of hateful comments, increasing significantly the risks.

- **Content, form and intent**: Both posts display animus, deliberately attacking specific transgender individuals and transgender people as a group. They intentionally misgender identifiable individuals, referring to a transgender woman and girl as "a boy who thinks he's a girl," denying the validity of transgender identities and harassing them. The first post uses a frequently invoked stereotype that transgender women access women's bathrooms to sexually assault cisgender women. This claim has been weaponized against LGBTQIA+ people, to intimidate, exclude and incite violence. It is particularly prominent in debates over school bathrooms, where the threat is falsely presented as accusations that transgender people are "pedophiles" intent on "grooming" young people in bathrooms. The second post targets a minor competing in a sports event, fueling controversy and hatred against transgender people.

- **Likelihood and imminence of acts of violence, discrimination or hostility**: According to the Board's research, the account creator is linked to multiple instances of harassment and threats of violence against LGBTQIA+ people. The promotion of anti-trans rhetoric contributes to a climate where violence against LGBTQIA+ people, including mass shootings in Buffalo, Colorado Springs, Orlando and Bratislava, becomes more likely. While direct causation is not necessary – actual "imminence" would inevitably be too late for Meta to make effective interventions to prevent violence – the link between the rhetoric in these posts and real-world violence is undeniable.

For the minority, taking all of these factors into consideration, both posts clearly contribute to an imminent risk of further "discrimination, hostility, or violence," and no measure short of removal would adequately prevent harm on this basis in either case.

The minority stresses that the purpose of the Bullying and Harassment policy is to ensure the safety of individuals, including children, from violence and physical harm, and to safeguard their psychological health, to prevent isolation, self-harm and suicide, so they can be free to express themselves free of that intimidation.

Meta's human rights responsibilities are heightened in respect of children. One in three internet users globally are under 18. The Committee on the Rights of the Child has recognized bullying as a form of violence against children (CRC, General Comment No. 25 on children's rights in relation to the digital environment, at para. 81). For a minority, Meta's threshold for classifying children as "voluntary public figures" is too low, with implications beyond LGBTQIA+ youth. When influential and popular accounts engage in anti-LGBTQIA+ bullying and harassment, they knowingly signal to their hundreds of thousands of followers to engage in online abuse. A minority is concerned that Meta does not consider the power imbalance between the accounts leading to harassment and targeted individuals. This can cause severe near-term harms that are especially acute for LGBTQIA+ youth, and, as discussed in the analysis above, makes the removal of both posts necessary and proportionate.

According to Meta, in situations where a child's gender identity is weaponized in public debates for political purposes, and this is reported on by the media, they become *by virtue of that attention* a voluntary public figure who can be subject to Tier 3 attacks in the same way as an elected official. This circular cruelty is not in the best interests of the child (CRC Article 3), and in the view of a minority, Meta should have a higher threshold to apply public figure status to minors and require more robust evidence to demonstrate that they have engaged with their fame. Otherwise, a child in this situation has only two options: to stop pursuing their passions or face harassment by their bullies.

Non-Discrimination

The Board observes that gender identity is a protected characteristic recognized under international human rights law, and this is reflected in Meta's listing of protected characteristics in the Hateful Conduct policy. The Board is concerned Meta has incorporated the term "transgenderism" into this policy. This term suggests that being transgender is a question of ideology, rather than an identity. For its rules to have legitimacy, Meta must seek to frame its content policies neutrally, in ways that respect human rights principles of equality and non-discrimination. This could be achieved, for example, by stating "discourse about gender identity and sexual orientation" in place of "discourse about transgenderism and homosexuality."

Human Rights Due Diligence

The UN Guiding Principles on Business and Human Rights, Guiding Principles 13, 17 (c) and 18, require Meta to engage in ongoing human rights due diligence for significant policy and enforcement changes, which the company would ordinarily do through its Policy Product Forum, including engagement with impacted stakeholders. The Board is concerned that Meta's January 7, 2025, policy and enforcement changes were announced hastily, in a departure from regular procedure, with no public information shared as to what, if any prior human rights due diligence it performed.

Now these changes are being rolled out globally, it is important that Meta ensures adverse impacts of these changes on human rights are identified, mitigated and prevented, and publicly reported. This should include a focus on how groups may be differently impacted, including women and LGBTQIA+ people. In relation to enforcement changes, due diligence should be mindful of the possibilities of both overenforcement (Call for Women's Protest in Cuba, Reclaiming Arabic Words) as well as underenforcement (Holocaust Denial, Homophobic Violence in West Africa, Post in Polish Targeting Trans People).

6. The Oversight Board's Decision

The Oversight Board upholds Meta's decision to leave up the content in both cases.

7. Recommendations

Content Policy

1. As part of its ongoing human rights due diligence, Meta should take all of the following steps in respect of the January 7, 2025, updates to the Hateful Conduct Community Standard. First, it should identify how the policy and enforcement updates may adversely impact the rights of LGBTQIA+ people, including minors, especially where these populations are at heightened risk. Second, Meta should adopt measures to prevent and/or mitigate these risks and monitor their effectiveness. Third, Meta should update the Board on its progress and learnings every six months, and report on this publicly at the earliest opportunity.

The Board will consider this recommendation implemented when Meta provides the Board with robust data and analysis on the effectiveness of its prevention or mitigation measures on the cadence outlined above and when Meta reports on this publicly.

2. To ensure Meta's content policies are framed neutrally and in line with international human rights standards, Meta should remove the term "transgenderism" from the Hateful Conduct policy and corresponding implementation guidance.

   The Board will consider this recommendation implemented when the term no longer appears in Meta's content policies or implementation guidance.

Enforcement

3. To reduce the reporting burden on targets of Bullying and Harassment, Meta should allow users to designate connected accounts, which are able to flag potential Bullying and Harassment violations requiring self-reporting on their behalf.

   The Board will consider this recommendation implemented when Meta makes these features available and easily accessible to all users via their account settings.

4. To ensure there are fewer enforcement errors on Bullying and Harassment violations requiring self-reporting, Meta should ensure the one report representing multiple reports on the same content is chosen based on the highest likelihood of a match between the reporter and the content's target. In doing this Meta should guarantee that any technological solutions account for potential adverse impacts on at-risk groups.

   The Board will consider this recommendation implemented when Meta provides sufficient data to validate the efficacy of improvements in the enforcement of self-reports of Bullying and Harassment violations as a result of this change.

***Procedural Note:**

- The Oversight Board's decisions are made by panels of five Members and approved by a majority vote of the full Board. Board decisions do not necessarily represent the views of all Members.
- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.
- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology.