Post in Polish Targeting Trans People

2023-023-FB-UA

Summary

The Oversight Board has overturned Meta's original decision to leave up a Facebook post in which a user targeted transgender people with violent speech advocating for members of this group to commit suicide. The Board finds the post violated both the Hate Speech and Suicide and Self-Injury Community Standards. However, the fundamental issue in this case is not with the policies, but their enforcement. Meta's repeated failure to take the correct enforcement action, despite multiple signals about the post's harmful content, leads the Board to conclude the company is not living up to the ideals it has articulated on LGBTQIA+ safety. The Board urges Meta to close enforcement gaps, including by improving internal guidance to reviewers.

About the Case

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, with text in Polish stating, "New technology … Curtains that hang themselves," and above that, "spring cleaning <3." The user's biography includes the description, "I am a transphobe." The post received less than 50 reactions.

Between April and May 2023, 11 different users reported the post a total of 12 times. Only two of the 12 reports were prioritized for human review by Meta's automated systems, with the remainder closed. The two reports sent for human review, for potentially violating Facebook's Suicide and Self-Injury Standard, were assessed as non-violating. None of the reports based on Hate Speech were sent for human review.

Three users then appealed Meta's decision to leave up the Facebook post, with one appeal resulting in a human reviewer upholding the original decision based on the Suicide and Self-Injury Community Standard. Again, the other appeals, made under the Hate Speech Community Standard, were not sent for human review. Finally, one of the users who originally reported the content appealed to the Board. As a result of the Board selecting this case, Meta determined the post did violate both its Hate Speech and Suicide and Self-Injury policies and removed it from Facebook. Additionally, the company disabled the account of the user who posted the content for several previous violations.

Key Findings

The Board finds the content violated Meta's Hate Speech policy because it includes "violent speech" in the form of a call for a protected-characteristic group's death by suicide. The post, which advocates for suicide among transgender people, created an atmosphere of intimidation and exclusion, and could have contributed to physical harm. Considering the nature of the text and image, the post also exacerbated the mental-health crisis being experienced by the transgender community. A recent report by the Gay and Lesbian Alliance Against Defamation (GLAAD) notes "the sheer traumatic psychological impact of being relentlessly exposed to slurs and hateful conduct" online. The Board finds additional support for its conclusion in the broader context of online and offline harms the LGBTQIA+ community is facing in Poland, including attacks and political rhetoric by influential government and public figures.

The Board is concerned that Meta's human reviewers did not pick up on contextual clues. The post's reference to the elevated risk of suicide ("curtains that hang themselves") and support for the group's death ("spring cleaning") were clear violations of the Hate Speech Community Standard, while the content creator's self-identification as a transphobe, alone, would amount to another violation. The Board urges Meta to improve the accuracy of hate speech enforcement towards LGBTQIA+ people, especially when posts include images and text that require context to interpret. In this case, the somewhat-coded references to suicide in conjunction with the visual depiction of a protected group (the transgender flag) took the form of "malign creativity." This refers to bad actors developing novel means of targeting the LGBTQIA+ community through posts and memes they defend as "humorous or satirical," but are actually hate or harassment.

Additionally, the Board is troubled by Meta's statement that the human reviewers' failures to remove the content aligns with a strict application of its internal guidelines. This would indicate that Meta's internal guidance inadequately captures how text and image can interact to represent a group defined by the gender identity of its members.

While the post also clearly violated Facebook's Suicide and Self-Injury Community Standard, the Board finds this policy should more clearly prohibit content promoting suicide aimed at an identifiable group of people, as opposed to only a person in that group.

In this case, Meta's automated review prioritization systems significantly affected enforcement, including how the company deals with multiple reports on the same piece of content. Meta monitors and deduplicates (removes) these reports "to ensure consistency in reviewer decisions and enforcement actions." Other reasons given for the automatic closing of reports included the content's low severity and low virality (amount of views the content has

accumulated) score, which meant it was not prioritized for human review. In this case, the Board believes the user's biography could have been considered as one relevant signal when determining severity scores.

The Board believes that Meta should invest more in the development of classifiers that identify potentially violating content impacting the LGBTQIA+ community and enhance training for human reviewers on gender identity-related harms.

The Oversight Board's Decision

The Oversight Board overturns Meta's original decision to leave up the content.

The Board recommends that Meta:

- Clarify on its Suicide and Self-Injury page that the policy forbids content promoting or encouraging suicide aimed at an identifiable group of people.
- Modify the internal guidance it gives to at-scale reviewers to ensure flag-based visual depictions of gender identity that do not contain a human figure are understood as representations of a group defined by the gender identity of its members.

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

Full Case Decision

1. Decision Summary

The Oversight Board overturns Meta's original decision to leave up a piece of content in Polish on Facebook in which a user targeted transgender people with violent speech that advocated for members of this group to commit suicide. After the Board identified the case for review, Meta concluded that its original decision to allow the post to remain on the platform was mistaken, removed the content and applied sanctions. The Board finds that the post violated both the Hate Speech and the Suicide and Self-Injury Community Standards. The Board takes this opportunity to urge Meta to improve its policies and guidance to its reviewers to better protect transgender people on its platforms. Specifically, in assessing whether a post is hate speech, Meta should ensure that flag-based visual depictions of gender identity that do not contain a human figure are understood as representations of a group defined by the gender identity of its members. Meta should also clarify that encouraging a whole group to commit suicide is just as violating as encouraging an individual to commit suicide. The Board finds,

however, that the fundamental issue in this case is not with the policies, but with their enforcement. Even as written, the policies clearly prohibited this post, which had multiple indications of hate speech targeting a group of people based on gender identity. Meta's repeated failure to take the correct enforcement action in this case, despite multiple user reports, leads the Board to conclude that Meta is failing to live up to the ideals it has articulated on LGBTQIA+ safety. The Board urges Meta to close these enforcement gaps.

2. Case Description and Background

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, text in Polish states: "New technology. Curtains that hang themselves." Above that line, more text in Polish states: "spring cleaning <3." A description in Polish in the user's bio reads, "I am a transphobe." The post received under 50 reactions from other users, the majority of which were supportive. The most frequently used reaction emoji was "Haha."

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." Meta generally prioritizes content for human review based on its severity, virality and likelihood of violating content policies. 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. None of the reports based on the Hate Speech policy were sent for human review. According to Meta, reviewers have both the training and tools to "assess and act" on content beyond their designated policy queue (i.e., Hate Speech or Suicide and Self-Injury). Nevertheless, both reviewers assessed the content to be non-violating and did not escalate it further.

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. This is because Meta will "monitor and deduplicate" multiple reports on the same piece of content to ensure consistency in reviewer decisions and enforcement actions.

One of the users who originally reported the content then appealed to the Board. As a result of the Board selecting this case, Meta determined that the content did violate both its Hate Speech and Suicide and Self-Injury policies and removed the post. Moreover, as part of Meta's review of the case, the company determined that the content creator's account already had several violations of the Community Standards and met the threshold to be disabled. Meta disabled the account in August 2023.

The Board noted the following context in reaching its decision in this case:

Poland is often reported to have high levels of hostility toward the LGBTQIA+ community. [Note: The Board uses "LGBTQIA+" (Lesbian, Gay, Bisexual, Transgender, Queer, Intersex and Asexual) when referring to groups based on sexual orientation, gender identity and/or gender expression. However, the Board will preserve the acronyms or usages employed by others when citing or quoting them.] The Council of Europe Commissioner for Human Rights has previously called attention to the "stigmatisation of LGBTI people" as a "long-standing problem in Poland." The International Lesbian, Gay, Bisexual, Trans and Intersex Association's (ILGA) Rainbow Europe report ranks countries on the basis of laws and policies that directly impact LGBTI people's human rights. The report ranks Poland as the lowest-performing European Union (EU) member state and 42nd out of 49 European countries assessed. National and local governments, as well as prominent public figures, have increasingly targeted the LGBTQIA+ community through both discriminatory speeches and legislative action.

Beginning in 2018, ILGA-Europe tracked what the organization called "high profile political hate-speech against LGBTI people from Polish political leaders," including statements that the "entire LGBT movement" is a "threat" to Poland. In the same year, the mayor of Lublin, Poland, attempted to ban the city's Equality March, although the Court of Appeal lifted the ban shortly before the scheduled march. In 2019, the mayor of Warsaw introduced the Warsaw LGBT+ Charter to "improve the situation of LGBT people" in the city. Poland's ruling Law and Justice party (PiS) and religious leaders criticized the charter. Poland's president and central government have also singled out the transgender community as targets. For example, the Chairman of the ruling PiS party has referred to transgender individuals as "abnormal." Poland's Minister of Justice has also asked Poland's Supreme Court to consider that "in addition to their parents, trans people should also sue their children and spouse [for permission to transition] when they want to access LGR [Legal Gender Recognition]."

9. Poland has also enacted anti-LGBTQIA+ legislation. In the words of Human Rights Watch, cities began calling for "the exclusion of LGBT people from Polish society" by implementing, among other measures, "LGBT-free zones" in 2019. Human Rights Watch has reported that these zones are places "where local authorities have adopted discriminatory 'family charters' pledging to 'protect children from moral corruption' or declared themselves free from 'LGBT ideology.'" More than 100 cities have created such zones. ILGA-Europe reports that due to local, EU and international pressure, some of these municipalities have withdrawn "anti-LGBT resolutions or Family Rights Charters." On June 28, 2022, Poland's Supreme Administrative Court ordered four municipalities to withdraw their anti-LGBTQI+ resolutions. Nevertheless, as Rainbow Europe report's ranking of Poland suggests, the climate in the country is notably hostile to the LGBTQIA+ community.

A 2019 survey of LGBTI people in the EU, conducted by the European Union Agency for Fundamental Rights, compared LGBTI peoples' experiences of assault and harassment in Poland and other parts of the European Union. According to the survey, 51% of LGBTI people in Poland often or always avoid certain locations for fear of being assaulted. This compares to 33% for the rest of the European Union. The survey also found that one in five transgender people were physically or sexually attacked in the five years before the survey, more than double that of other LGBTI groups.

The Board commissioned external experts to analyze social-media responses to derogatory statements by Polish government officials. Those experts noted "a concerning uptick in online hate speech targeting minority communities in Poland, including LGBTQIA+ communities since 2015." In its analysis of anti-LGBTQIA+ content in Polish on Facebook, these experts noted that spikes occurred during "court rulings relating to anti-LGBTQIA+ legislation." These include the Supreme Administrative Court decision discussed above and determinations relating to legal challenges to the adoption of several anti-LGBT declarations brought before local administrative courts by the Polish ombudsmen for the Commissioner for Human Rights, which have been ongoing since 2019.

The Board also asked linguistic experts about the meaning of the two Polish phrases in the post. With regard to the phrase "curtains that hang themselves," the experts observed that in the context of a "trans flag hanging in the window," the phrase was "a play on words" that juxtaposed "to hang curtains" with "to commit suicide by hanging." The experts concluded that the phrase was "a veiled transphobic slur." On the "spring cleaning" phrase, experts said

that the phrase “normally refers to thorough cleaning when spring comes” but, in certain contexts, “it also means ‘throwing out all trash’ and ‘getting rid of all unwanted items (and/or people).’” Several public comments, including the submission from the Human Rights Campaign Foundation (PC-16029) argued that the post’s reference to “spring cleaning” was a form of “praising the exclusion and isolation of trans people out of Polish society (through their deaths).”

The issues of online and offline harms at play in this case extend beyond the LGBTQIA+ community in Poland to affect that community around the globe. According to the World Health Organization, suicide is the fourth-leading cause of death among 15–29-year-olds worldwide. WHO notes that “suicide rates are also high amongst vulnerable groups who experience discrimination, such as refugees and migrants, indigenous peoples; and lesbian, gay, bisexual, transgender, intersex (LGBTI) persons.” Other research studies have found a “positive association” between cyber-victimization and self-injurious thoughts and behaviors.

Suicide risk is a particular concern for the transgender and nonbinary community. The Trevor Project’s 2023 National Survey on LGBTQ Mental Health found that half of transgender and nonbinary youth in the United States considered attempting suicide in 2022. The same study estimates that 14% of LGBTQ young people have attempted suicide in the past year, including nearly one in five transgender and nonbinary young people. According to the CDC’s Youth Risk Behavior Survey, 10% of high school students in the United States attempted suicide in 2021. Numerous studies from around the world have found that transgender or nonbinary people are at a higher risk of both suicidal thoughts and attempts compared to cisgender people.

In a public comment to the Board, the Gay and Lesbian Alliance Against Defamation (GLAAD) (PC-16027) underscored findings from their annual survey, the Social Media Safety Index, on LGBTQ user safety on five major social-media platforms. The 2023 report assigned Facebook a score of 61% based on 12 LGBTQ-specific indicators. This score represented a 15-point increase from 2022, with Facebook ranked second to Instagram and above the three other major platforms. However, GLAAD wrote, “safety and the quality of safeguarding of LGBTQ users remain unsatisfactory.” The report found that there are “very real resulting harms to LGBTQ people online, including a chilling effect on LGBTQ freedom of expression for fear of being targeted, and the sheer traumatic psychological impact of being relentlessly exposed to slurs and hateful conduct.”

3. Oversight Board Authority and Scope

The Board has authority to review Meta's decision following an appeal from the person who previously reported content that was left up (Charter Article 2, Section 1; Bylaws Article 3, Section 1).

The Board may uphold or overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). When Meta commits to act on recommendations, the Board monitors their implementation.

When the Board selects cases like this one, in which Meta subsequently acknowledges that it made an error, the Board reviews the original decision to increase understanding of the content moderation process and to make recommendations to reduce errors and increase fairness for people who use Facebook and Instagram.

4. Sources of Authority and Guidance

The following standards and precedents informed the Board's analysis in this case:

I. Oversight Board Decisions

The most relevant previous decisions of the Oversight Board include:

- Reclaiming Arabic Words
- Knin Cartoon
- Colombia Protests
- Armenians in Azerbaijan

II. Meta's Content Policies

The Hate Speech policy rationale defines hate speech "as a direct attack against people – rather than concepts or institutions – on the basis of . . . protected characteristics," including sex and gender identity. Meta defines "attacks" as "violent or dehumanizing speech, harmful

stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation." In the policy rationale, Meta further states: "We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence."

Meta's Hate Speech Community Standard separates attacks into "tiers." Tier 1 attacks include content targeting a person or group of people on the basis of their protected characteristic(s) with "violent speech or support in written or visual form." Meta ultimately found that the post in this case violated this policy line. On December 6, 2023, Meta updated the Community Standards to reflect that the prohibition on violent speech against protected-characteristic groups was moved to the Violence and Incitement policy.

Tier 2 attacks include content targeting a person or group of people on the basis of their protected characteristic(s) with "expressions of contempt (in written or visual form)." In the Hate Speech Community Standard, Meta defines expressions of contempt to include "[s]elf-admission to intolerance on the basis of a protected characteristic" and "[e]xpressions that a protected characteristic shouldn't exist."

The Suicide and Self-Injury Community Standard prohibits "any content that encourages suicide or self-injury, including fictional content such as memes or illustrations." Under this policy, Meta removes "content that promotes, encourages, coordinates, or provides instructions for suicide and self-injury."

The Board's analysis was informed by Meta's commitment to voice, which the company describes as "paramount," and its values of safety and dignity.

III. Meta's Human-Rights Responsibilities

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human-rights responsibilities of private businesses. In 2021, Meta announced its Corporate Human Rights Policy, in which it reaffirmed its commitment to respecting human rights in accordance with the UNGPs. The

Board's analysis of Meta's human-rights responsibilities in this case was informed by the following international standards:

- The rights to freedom of opinion and expression: Articles 19, International Covenant on Civil and Political Rights (ICCPR); General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur (UNSR) on freedom of opinion and expression, reports: A/HRC/38/35 (2018), A/74/486 (2019); and Rabat Plan of Action, UN High Commissioner for Human Rights report: A/HRC/22/17/Add.4 (2013).
- The right to life: Article 6, ICCPR.
- The right to the enjoyment of the highest attainable standard of physical and mental health: Article 12, International Covenant on Economic, Social and Cultural Rights (ICESCR).
- The right to equality and non-discrimination: Article 2, para. 1 and Article 26, ICCPR.

5. User Submissions

In their appeal to the Board, the user who reported the content noted that the person who posted the image had previously harassed transgender people online and had created a new account after being suspended from Facebook. They also said that applauding the high rate of suicide in the transgender community "shouldn't be allowed."

6. Meta's Submissions

Meta eventually removed the post under Tier 1 of its Hate Speech Community Standard because the content violated the policy line prohibiting content targeting a person or group of people on the basis of their protected characteristics with "violent speech or support in written or visual form." In its internal guidelines about how to apply this policy, Meta says that content should be removed if it is "violent speech in the form of calls for action or statements of intent to inflict, aspirational or conditional statements about, or statements advocating or supporting death, disease or harm (in written or visual form)."

In those internal guidelines, the company also describes what it considers to be visual representation of protected-characteristic groups in an image or video. Meta did not allow the Board to publish more detailed information related to this guidance. The company instead said that, "under the Hate Speech policy, Meta may take visual elements in the content into consideration when establishing whether the content targets a person or group of people based on their protected characteristics."

Meta said that the multiple assessments of the content as a non-violation of the Hate Speech Community Standard by its reviewers align with "a strict application of our internal guidelines." It elaborated: "Although the curtains resemble the Trans Pride flag, we would interpret an attack on a flag, standing alone, as an attack on a concept or institution, which does not violate our policies, and not on a person or groups of people." However, Meta subsequently determined that the "reference to hanging indicates this post is attacking a group of people." This assessment was based on the determination that the phrase "'curtains which hang themselves' implicitly refers to the suicide rate in the transgender community because the curtains resemble the Trans Pride flag, and the curtains hanging in the photo (as well the text overlay) is a metaphor for suicide by hanging oneself." Meta also noted that "concepts or institutions cannot 'hang themselves,' at least not literally." For this reason, Meta found that the user was referring to "Transgender people, not just the concept." Therefore, according to Meta, "this content violates the Hate Speech policy because it is intended to be interpreted as a statement in favor of a P[rotected] C[haracteristic] group's death by suicide."

Following an update to the Hate Speech policy, in which the policy line prohibiting violent speech against protected-characteristic groups was moved to the Violence and Incitement policy, Meta told the Board that the content remains violating.

Meta also reported that the statement in the biography of the user's account reading, "I am a transphobe," violated Tier 2 of the Hate Speech policy as a "self-admission to intolerance on the basis of protected characteristics." This statement was, according to Meta, assessed as violating as part of Meta's review of both the case and the user's account following the Board's selection of the case. Meta said that this statement helped inform its understanding of the user's intent in the case content.

In response to the Board asking whether the content violates the Suicide and Self-Injury policy, Meta confirmed that the "content violates the Suicide and Self-Injury policy by encouraging suicide, consistent with our determination that the content constitutes a statement in favor of a protected characteristic group's death by suicide." Meta also reported that the Suicide and Self-Injury policy "does not differentiate between content that promotes or encourages suicide aimed at a specific person versus a group of people."

The Board asked Meta 13 questions in writing. Questions related to Meta's content-moderation approach to transgender and LGBTQIA+ issues; the relationship between the Hate Speech and Suicide and Self-Injury Community Standards; how "humor" and "satire" are assessed by moderators when reviewing content for hate speech violations; the role of "virality" and "severity" scores in prioritizing content for human review; and how Meta's content-moderation practices deal with prioritization of content for human review that has multiple user reports. Meta answered all 13 questions.

7. Public Comments

The Oversight Board received 35 public comments relevant to this case, including 25 from the United States and Canada, seven from Europe and three from Asia Pacific and Oceania. This total includes public comments that were either duplicates or were submitted with consent to publish, but did not meet the Board's conditions for publication. Such exclusion can be based on the comment's abusive nature, concerns about user privacy and/or other legal reasons. Public comments can be submitted to the Board with or without consent to publish, and with or without consent to attribute.

The submissions covered the following themes: the human-rights situation in Poland, particularly as it is experienced by transgender people; LGBTQIA+ safety on social-media platforms; the relationship between online and offline harms in Poland; the relationship between humor, satire, memes and hate/harassment against transgender people on social-media platforms; and the challenges of moderating content that requires context to interpret.

To read public comments submitted for this case, please click here.

8. Oversight Board Analysis

The Board examined whether this content should be removed by analyzing Meta's content policies, human rights-responsibilities and values. The Board also assessed the implications of this case for Meta's broader approach to content governance.

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 involves both hate speech and the promotion of suicide or self-injury.

8.1 Compliance With Meta's Content Policies

I. Content Rules

Hate Speech

The Board finds that the content in this case violates the Hate Speech Community Standard. The post included "violent speech or support" (Tier 1) in the form of a call for a protected-characteristic group's death by suicide, which clearly violates the Hate Speech policy.

The Board agrees with Meta's eventual conclusion that the reference to hanging in the post is an attack on a group of people rather than a concept because "concepts or institutions cannot 'hang themselves.'" The Board also finds support for its conclusion in the broader context around online and offline harms that members of the LGBTQIA+ community, and specifically transgender people, face in Poland. A post that uses violent speech to advocate for and support the death of transgender people by suicide creates an atmosphere of intimidation and exclusion, and could contribute to physical harm. This context in which the post's language was used makes clear it was meant to dehumanize its target. In light of the nature of the text and image, the post also exacerbates the ongoing mental health crisis that is currently being experienced by the transgender community. According to multiple studies, transgender or nonbinary people are at a higher risk of both suicidal thoughts and attempts than cisgender individuals. Moreover, the experience of online attacks and victimization has been positively correlated with suicidal thoughts. In this context, the Board finds that the presence of the transgender flag, coupled with the reference to the elevated risk of suicide within the transgender community ("curtains that hang themselves"), is a clear indication that transgender people are the post's target. The Board finds that the phrase "spring cleaning," followed by a "<3" (heart) emoticon, also constitutes support for the group's death. As such, it also violates the Hate Speech Community Standard's prohibition (Tier 2) on "expressions that a protected characteristic shouldn't exist."

With respect to this Community Standard, the Board believes that the policy and internal guidelines for its enforcement could be more responsive to "malign creativity" in content trends that target historically marginalized groups. The Wilson Center coined this phrase with research on gendered and sexualized abuse, and GLAAD also underscored its relevance in their public comment (PC-16027). "Malign creativity" refers to "the use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection on social-media platforms." In applying the concept to the post in this case, GLAAD said "malign creativity" involves "bad actors develop[ing] novel means of targeting the LGBTQ community" and vulnerable groups more generally through posts and memes that they defend as "humorous or satirical," but are "actually anti-LGBTQ hate or harassment." Specifically, "malign creativity" took the form of a post that uses two somewhat-coded references to suicide ("curtains that hang themselves" and "spring cleaning") in conjunction with a visual depiction of a protected group (the transgender flag) to encourage death by suicide. In the Armenians in Azerbaijan decision, the Board noted the importance of context in determining that the term under consideration in that case was meant to target a group based on a protected characteristic. While the context of war prevailed in that case, the threats faced by transgender people in Poland show that situations can be dire for a community, short of war. As noted above, one in five transgender people in Poland reported having been physically or sexually attacked in the five years before 2019, more than double the number of individuals from other LGBTI groups reporting such attacks.

The Board is concerned that Meta's initial human reviewers did not pick up on these contextual clues within the content and, as a result, concluded the content was non-violating. While the Board is recommending some revisions to the guidance on enforcing the Hate Speech Community Standard, it underscores that the post violated the policies even as they were written at the time. Both statements made in the post support the death of transgender people by suicide. An additional signal in the user's bio supports this conclusion. The user's self-identification as a transphobe would – in and of itself – constitute a Tier 2 violation of the Hate Speech Standard's prohibition on "self-admission to intolerance on the basis of protected characteristics." Meta must improve the accuracy of its enforcement on hate speech towards the LGBTQIA+ community, either through automation or human review, especially when posts include images and text that require context to interpret. As GLAAD observed (PC-16027), Meta "consistently fails to enforce its policies when reviewing reports on content that employs 'malign creativity.'"

The Board is also troubled by Meta's statement that the reviewers' failure to remove the content "aligns with a strict application of our internal guidelines." Meta's statement indicates that the internal guidance to reviewers inadequately captures how text and image can interact in a social-media post to represent a group defined by the gender identity of its members. The Board finds that the guidance may not suffice for at-scale content reviewers to be able to reach the correct enforcement outcome on content that targets protected-characteristic groups that are represented visually, but are not named or depicted in human figures. Meta did not allow the Board to publish additional details that would have enabled a more robust discussion of how enforcement of this type of content could be improved. However, the Board believes Meta should modify its guidance to ensure that visual depictions of gender identity are adequately understood when assessing content for attacks. The Board underscores that in suggesting this course, it does not seek to diminish Meta's protection of challenges to concepts, institutions, ideas, practices or beliefs. Rather, the Board wants Meta to clarify that posts need not depict human figures to constitute an attack on people.

Suicide and Self-Injury

The Board finds that the content in this case also violates the Suicide and Self-Injury Community Standard. This policy prohibits "content that promotes, encourages, coordinates, or provides instructions for suicide and self-injury." According to internal guidelines that Meta provides to reviewers, "promotion" is defined as "speaking positively of." The Board agrees with Meta's eventual conclusion that the content constitutes a statement in favor of a protected-characteristic group's death by suicide, and therefore encourages suicide.

The Board also finds that the Suicide and Self-Injury Community Standard should more expressly prohibit content that promotes or encourages suicide aimed at an identifiable group of people, as opposed to a person in that group. Meta disclosed to the Board that the policy does not differentiate between these two forms of content. Given the challenges that reviewers faced in identifying a statement encouraging a group's suicide in this case, however, the Board urges Meta to clarify that the policy forbids content that promotes or encourages suicide aimed at an identifiable group of people. Meta should clarify this point on its Suicide and Self-Injury policy page as well as in its associated internal guidelines to reviewers.

II. Enforcement Action

The Board finds that Meta's automated review prioritization systems significantly affected the enforcement actions in this case. Of the 12 user reports of the post, 10 were automatically closed by Meta's automated systems. Of the three user appeals against Meta's decisions, two were automatically closed by Meta's automated systems. The Board is concerned that the case history shared with the Board contains numerous indications of a violation and thus suggests that Meta's policies are not being adequately enforced.

The Board notes that many user reports were closed as a result of Meta's content moderation practices to deal with multiple reports on the same piece of content. The first user report for hate speech was not prioritized for human review because of a "low severity and a low virality score." Subsequent reports for hate speech were not prioritized for human review because when multiple reports are given on the same piece of content, Meta will "deduplicate those reports to ensure consistency in reviewer decisions and enforcement actions." The Board acknowledges that deduplication is a reasonable practice for content moderation at scale. However, the Board notes that the practice puts more pressure on the initial determination made on a report, as that will also determine the fate of all reports that are grouped with it.

The Board believes it would be important for Meta to prioritize improving the accuracy of automated systems that both enforce content policies and prioritize content for review, particularly when dealing with content that potentially impacts LGBTQIA+ people. Such improvements to the ability of automated systems to recognize the kind of coded language and context-based images considered in this case would undoubtedly improve enforcement on content that targets other protected-characteristic groups as well. The Board believes that the user's biography, for example, which included a self-admission of transphobia, could have been considered as one relevant signal when determining severity scores for the purpose of deciding whether to prioritize content for review and/or to take an enforcement action. This signal could supplement existing behavioral and social-network analyses that Meta might use to surface potentially violating content.

Additionally, the Board emphasizes that it would be important for Meta to ensure automated systems are well calibrated and content reviewers are trained to effectively assess LGBTQIA+ community-related posts at scale. The Board is concerned about Meta's current approach, under which reviewers tasked with assessing appeals often seem to have the same level of expertise as those performing the first content assessment. The Board believes that Meta should invest more in the development and training of classifiers that surface potentially

violating content impacting the LGBTQIA+ community and prioritize that content for human review. Hate speech, especially the highest severity content that falls under Tier 1 of Meta's policy, should always be prioritized for review. The Board also suggests bolstering these process improvements with: i) enhanced training on harms relating to gender identity for reviewers; ii) a task force on transgender and non-binary people's experiences on Meta's platforms; and iii) the creation of a specialized group of subject-matter experts to review content related to issues impacting the LGBTQIA+ community. While the facts of this case pertain specifically to the harms faced by transgender people on Facebook, the Board also encourages Meta to explore how to improve enforcement against hateful content impacting other protected-characteristic groups.

While the Board is only issuing two formal recommendations below, the Board underscores that this is because the challenges highlighted in this case have less to do with the policies as written than with their enforcement. The Board counts at least five indicia of harmful content in this case: (1) the post's references to "self-hanging curtains"; (2) the post's reference to "spring cleaning <3"; (3) the user's self-description as a "transphobe" in a country context where high levels of hostility toward the LGBTQIA+ community are reported; (4) the number of user reports and appeals on the content; and (5) the number of reports and appeals relative to the virality of the content. The Board is concerned Meta missed these signals and believes this suggests that its policies are underenforced. The Board is adamant that Meta should think rigorously and creatively about how to close the gap between its ideals of safeguarding LGBTQIA+ individuals on its platforms and its enforcement of those ideals.

8.2 Compliance With Meta's Human-Rights Responsibilities

Freedom of Expression (Article 19 ICCPR)

Article 19, para. 2 of the International Covenant on Civil and Political Rights (ICCPR) provides that "everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media." General Comment No. 34 (2011) further specifies that protected expression includes expression that may be considered "deeply offensive" (para. 11).

Where 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 voluntary human-rights commitments, 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 under international human-rights law requires rules that limit expression to be clear and publicly accessible (General Comment No. 34, para. 25). Rules restricting expression "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and "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.*). Applied to rules that govern online speech, the UN Special Rapporteur on freedom of expression has said they 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 on their enforcement.

The Board finds that Meta's prohibitions of "violent speech or support" in written or visual form targeting groups with protected characteristics, of expressions that a protected characteristic shouldn't exist and of speech that promotes or encourages suicide and self-injury are sufficiently clear.

The Board notes, however, that Meta could improve enforcement accuracy in relation to the policies engaged in this case by providing clearer guidance for human reviewers, as addressed under section 8.1 above. Meta should clarify that visual depictions of gender identity, such as through a flag, need not depict human figures to constitute an attack under the Hate Speech policy. Meta also should clarify that a call for a group (as opposed to an individual) to commit suicide violates the Suicide and Self-Injury Policy.

II. Legitimate Aim

Any restriction on expression should pursue one of the legitimate aims of the ICCPR, which include the "rights of others." In several decisions, the Board has found that Meta's Hate Speech policy, which aims to protect people from the harm caused by hate speech, has a legitimate aim that is recognized by international human-rights law standards (see, for example, Knin Cartoon decision). Additionally, the Board finds that, in this case, the Suicide and Self-Injury policy lines on content that encourages suicide or self-injury serve the legitimate aims of protecting people's right to the enjoyment of the highest attainable standard of physical and mental health (ICESCR 12) and the right to life (Article 5, ICCPR). In cases such as this one, where a protected-characteristic group is encouraged to commit suicide, the Suicide and Self-Injury policy also protects people's rights to equality and non-discrimination (Article 2, para. 1, ICCPR).

III. Necessity and Proportionality

The principle of necessity and proportionality provides that any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; [and] they must be proportionate to the interest to be protected" (General Comment No. 34, para. 34).

When analyzing the risks posed by violent content, the Board is typically guided by the six-factor test described in the Rabat Plan of Action, which addresses advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence. Based on an assessment of the relevant factors, especially the content and form of expression, the intent of the speaker and the context further described below, the Board finds that removing the content is in compliance with Meta's human-rights responsibilities as it poses imminent and likely harm. Removing the content is a necessary and proportionate limitation on expression in order to protect the right to life, as well as the right to enjoyment of the highest attainable standard of physical and mental health of the broader LGBTQIA+ community and, in particular, transgender people in Poland.

While the Board has previously noted the importance of reclaiming derogatory terms for LGBTQIA+ people in countering disinformation (see Reclaiming Arabic words decision), that is not the case here. The post also does not contain political nor newsworthy speech (see

Colombia Protests decision). The post in this case features the image of the transgender flag hanging as curtains, with the description that curtains hang themselves. According to experts consulted by the Board, the use of curtains – in both visual and textual form – does not appear to be recurring coded language aimed at the transgender community. Nonetheless, as discussed above, the phenomenon of "malign creativity," or the use of novel language and strategies of representation to express hate and harassment, has come to characterize content trends that target transgender people. The Board finds that the content in this case fits squarely within that trend. Although the post used imagery that some found "humorous" (as evidenced by the "Haha" emoji reactions), the post can still be interpreted as a violent and provocative statement targeting the transgender community. Humor and satire can, of course, be used to push the boundaries of legitimate criticism, but it cannot be a cover for hate speech. The post only engages with the topic of high suicide rates among the transgender community to celebrate this fact.

When considering the intent of the content creator, the Board notes that their biography openly stated they are a "transphobe." While Meta only later considered the implications of this statement for the case content itself, the Board finds it to be highly relevant to determining the intent of the user. It would also be an independent ground for removing the content as a Tier 2 violation of the Hate Speech policy. The post also described the act of transgender individuals dying by suicide to be "spring cleaning," including a heart emoticon alongside the description. In light of this statement of support for a group's death by suicide, the Board finds intent to encourage discrimination and violence based on the content of the post, image used and accompanying text and caption. The content in this case not only encourages transgender people to take violent action against themselves but also incites others to discriminate and act with hostility towards transgender people. This understanding is confirmed by the fact that the reaction emoji most frequently employed by other users engaging with the content was "Haha."

Finally, the Board notes the significant offline risks that the Polish LGBTQIA+ community faces in the form of increasing attacks through legislative and administrative action, as well as political rhetoric by central government figures and influential public voices. Since 2020, Poland has consistently ranked as the lowest-performing EU member country for LGBTQIA+ rights, according to ILGA-Europe. It is also important to note that Poland does not have LGBTQIA+ protections in its hate speech and hate crime laws, an issue that ILGA-Europe and Amnesty International, among others, have called upon Poland to address. Furthermore, the

rise in anti-LGBTQIA+ rhetoric in Polish on Facebook, flagged by external experts and numerous public comments, is not happening in isolation. Many organizations and institutions have expressed alarm at the prevalence of anti-LGBTQIA+ speech on social media. UN Independent Expert on Sexual Orientation and Gender Identity (IE SOGI) Victor Madrigal-Borloz has said that levels of violence and discrimination against gender-diverse and transgender people "offend the human conscience." GLAAD's research and reporting has found that there are "very real resulting harms to LGBTQ people online, including … the sheer psychological trauma of being relentlessly exposed to slurs and hateful conduct." Content like the post in this case, especially when considered at scale, may contribute to an environment in which the already pervasive harm of dying by suicide within the transgender community is exacerbated. Moreover, content that normalizes violent anti-transgender speech, as is the case with this post, risks contributing to both the ongoing mental-health crisis that impacts the transgender community, as well as an increase in violence targeting the community offline.

9. Oversight Board Decision

The Oversight Board overturns Meta's original decision to leave up the content.

10. Recommendations

Content Policy

1. Meta's Suicide and Self-Injury policy page should clarify that the policy forbids content that promotes or encourages suicide aimed at an identifiable group of people.

The Board will consider this implemented when the public-facing language of the Suicide and Self-Injury Community Standard reflects the proposed change.

Enforcement

2. Meta's internal guidance for at-scale reviewers should be modified to ensure that flag-based visual depictions of gender identity that do not contain a human figure are understood as representations of a group defined by the gender identity of its members. This modification

would clarify instructions for enforcement of this form of content at-scale whenever it contains a violating attack.

The Board will consider this implemented when Meta provides the Board with the changes to its internal guidance.

Procedural Note:

The Oversight Board's decisions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg, which draws on a team of more than 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. Memetica, an organization that engages in open-source research on social-media trends, also provided analysis. Linguistic expertise was provided by Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world.