IG-2PJ00L4T

Reclaiming Arabic words

The Oversight Board has overturned Meta's original decision to remove an Instagram post which, according to the user, showed pictures of Arabic words which can be used in a derogatory way towards men with "effeminate mannerisms." The content was covered by an exception to Meta's Hate Speech policy and should not have been removed.

About the case

In November 2021, a public Instagram account which describes itself as a space for discussing queer narratives in Arabic culture posted a series of pictures in a carousel (a single Instagram post that can contain up to 10 images with a single caption). The caption, written in both Arabic and English, explained that each picture shows a different word that can be used in a derogatory way towards men with "effeminate mannerisms" in the Arabic-speaking world, including the terms "zamel," "foufou," and "tante/tanta." The user stated that the post intended "to reclaim [the] power of such hurtful terms."

Meta initially removed the content for violating its Hate Speech policy but restored it after the user appealed. After being reported by another user, Meta then removed the content again for violating its Hate Speech policy. According to Meta, before the Board selected this case, the content was escalated for additional internal review which determined that it did not, in fact, violate the company's Hate Speech policy. Meta then restored the content to Instagram. Meta explained that its initial decisions to remove the content were based on reviews of the pictures containing the terms "z***l" and "t***e/t***a."

Key findings

The Board finds removing this content to be a clear error which was not in line with Meta's Hate Speech policy. While the post does contain slur terms, the content is covered by an exception for speech "used self-referentially or in an empowering way," as well as an exception which allows the quoting of hate speech to "condemn it or raise awareness." The user's statements that they did not "condone or encourage the use" of the slur terms in question, and that their aim was "to reclaim [the] power of such hurtful terms," should have alerted the moderator to the possibility that an exception may apply.

For LGBTQIA+ people in countries which penalize their expression, social media is often one of the only means to express themselves freely. The over-moderation of speech by users from persecuted minority groups is a serious threat to their freedom of expression. As such, the Board is concerned that Meta is not consistently applying exemptions in the Hate Speech policy to expression from marginalized groups.

The errors in this case, which included three separate moderators determining that the content violated the Hate Speech policy, indicate that Meta's guidance to moderators assessing references to derogatory terms may be insufficient. The Board is concerned that reviewers may not have sufficient resources in terms of capacity or training to prevent the kind of mistake seen in this case.

Providing guidance to moderators in English on how to review content in non-English languages, as Meta currently does, is innately challenging. To help moderators better assess when to apply exceptions for content containing slurs, the Board recommends that Meta translate its internal guidance into dialects of Arabic used by its moderators.

The Board also believes that to formulate nuanced lists of slur terms and give moderators proper guidance on applying exceptions to its slurs policy, Meta must regularly seek input from minorities targeted with slurs on a country and culture-specific level. Meta should also be more transparent around how it creates, enforces, and audits its market-specific lists of slur terms.

The Oversight Board's decision

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

As a policy advisory statement, the Board recommends that Meta:

- Translate the Internal Implementation Standards and Known Questions into dialects of Arabic used by its content moderators. Doing so could reduce over-enforcement in Arabic-speaking regions by helping moderators better assess when exceptions for content containing slurs are warranted.
- Publish a clear explanation of how it creates its market-specific slur lists. This explanation should include the processes and criteria for designating which slurs and countries are assigned to each market-specific list.
- Publish a clear explanation of how it enforces its market-specific slur lists. This explanation should include the processes and criteria for determining precisely when and where the slurs prohibition will be enforced, whether in respect to posts originating geographically from the region in question, originating outside but relating to the region in question, and/or in relation to all users in the region in question, regardless of the geographic origin of the post.
- Publish a clear explanation of how it audits its market-specific slur lists. This explanation should include the processes and criteria for removing slurs from or keeping slurs on Meta's market-specific lists.

*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 remove an Instagram post by an account that explores "queer narratives in Arabic history and popular culture." The content falls into an exception in Meta's Hate

and popular culture." The content falls into an exception in Meta's Hate Speech policy as it reports, condemns, and discusses the negative use of homophobic slurs by others and uses them in an expressly positive context.

2. Case description and background

In November 2021, a public Instagram account which identifies itself as a space for discussing queer narratives in Arabic culture posted a series of pictures in a carousel (a single Instagram post that can contain up to 10 images with a single caption). The caption, which the user wrote in both Arabic and English, explains that each picture shows a different word that can be used in a derogatory way towards men with "effeminate mannerisms" in the Arabic speaking world, including the terms "zamel," "foufou," and "tante"/"tanta." In the caption the user stated that they did not "condone or encourage the use of these words," but explained that they had previously been abused with one of these slurs and that the post was intended "to reclaim [the] power of such hurtful terms." The Board's external experts confirmed that the terms quoted in the content are often used as slurs.

The content was viewed approximately 9,000 times, receiving around 30 comments and approximately 2,000 reactions. Within three hours of the content being posted, a user reported it for "adult nudity or sexual activity" and another user reported it as "sexual solicitation." Each report was dealt with separately by different human moderators. No action was taken by the moderator who reviewed the first report, but the moderator who reviewed the second report removed the content for violating Meta's Hate Speech policy. The user appealed this removal and a third moderator restored the content to the platform. After the content was restored, another user reported it as "hate speech" and another moderator carried out a fourth review, again removing the content. The user appealed a second time and, after a fifth review, another moderator upheld the decision to remove the content. After Meta notified the user of that decision, the user submitted an appeal to the Oversight Board. Meta later confirmed that all of the moderators who reviewed the content were fluent Arabic speakers.

Meta explained that the initial decisions to take down the content were based on reviews of the pictures containing the terms "z***l" and "t***e/t***a". In response to a question from the Board Meta also noted that the company

considers another term used in the content, "moukhanath" to be a slur.

According to Meta, after the user appealed to the Board but before the Board selected the case, the content was independently escalated for an additional internal review, which determined that it did not violate the Hate Speech Policy. The content was subsequently restored to the platform.

3. Oversight Board authority and scope

The Board has authority to review Meta's decision following an appeal from the user whose content was removed (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 policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4).

When the Board selects cases like this one, where Meta has agreed that it made an error, the Board reviews the original decision to help increase understanding of why errors occur, and to make observations or recommendations that may contribute to reducing errors and to enhancing due process.

4.Sources of authority

The Oversight Board considered the following as sources of authority:

I.Oversight Board decisions:

The Board's most relevant decisions to this case include:

- The "Wampum Belt decision" (2021-012-FB-UA): In this decision the Board highlighted the importance of protecting the expression of marginalized groups, noting that Meta must ensure that it does not

marginalized groups, noting that Meta must ensure that it does not remove content that falls under a Hate Speech policy exception.

- The "South Africa slurs decision" (2021-011-FB-UA): In this decision the Board found that Meta must be more transparent on the procedures and criteria it uses for developing its slur lists. The Board also recommended that Meta prioritize improving procedural fairness in its enforcement of the Hate Speech policy so that users can better understand why content is removed.
- The "Myanmar bot decision" (2021-007-FB-UA): In this decision the Board emphasised the importance of context in assessing whether content falls into the exceptions to the Hate Speech policy.

The Board also refers to recommendations made in: The "Ocalan's isolation decision" (2021-006-IG-UA), the "Two buttons meme decision" (2021-005-FB-UA), and the "Breast cancer symptoms and nudity decision" (2020-004-IG-UA).

II. Meta's content policies:

This case involves Instagram's Community Guidelines and Facebook's Community Standards. Meta's Transparency Center states that "Facebook and Instagram share Content Policies. This means that if content is considered violating on Facebook, it is also considered violating on Instagram."

Instagram's Community Guidelines state:

We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech… It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent clearly.

Facebook's Community Standards define hate speech as "a direct attack on

people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability." Meta divides attacks into three tiers. The slurs section of the hate speech policy prohibits "[c]ontent that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels for the above characteristics." The rest of tier three prohibits content targeting people with segregation or exclusion.

As part of the policy rationale Meta explains that:

We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.

III. Meta's values:

Meta's values are outlined in the introduction to the Facebook Community Standards where the value of "Voice" is described as "paramount":

The goal of our Community Standards has always been to create a place for expression and give people a voice. […] We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Meta limits "Voice" in service of four values, two of which are relevant here:

"Safety": We are committed to making Facebook a safe place. Expression that threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook.

"Dignity": We believe that all people are equal in dignity and rights. We expect that people will respect the dignity of others and not harass or degrade them.

IV. International human rights standards:

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, where 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 human rights standards which are applied in Section 8 of this decision:

- The rights to freedom of opinion and expression: Article 19, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human Rights Committee, 2011; Communication 488/1992, Toonen v. Australia, Human Rights Committee, 1992; Resolution 32/2, Human Rights Council, 2016; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019); UN High Commissioner for Human Rights, report: A/HRC/19/41 (2011).
- The right to non-discrimination: Article 2, para. 1 and Article 26, ICCPR.

5. User submissions

In their statement to the Board, the user described their account as a place "to celebrate queer Arab culture." They explain that while it is a "safe space," as its following has grown it has increasingly been targeted by homophobic trolls who write abusive comments and mass-report content.

The user explained that their intent in posting the content was to celebrate "effeminate men and boys" in Arab society who are often belittled with the derogatory language highlighted in the post. They further explained that they were attempting to reclaim these derogatory words used against them as a form of resistance and empowerment, and argued that they made clear in the post's content that they do not condone or encourage the use of the words in the pictures as slurs. The user also stated that they believed their content complied with Meta's content policies which specifically permit the use of otherwise banned terms when used self-referentially or in an empowering way.

6. Meta's submissions

Meta explained in its rationale that the content was originally removed under its Hate Speech Policy as the content contains a prohibited word on Meta's slur list which is "a derogatory term for gay people." Meta ultimately reversed its original decision and restored the content as the use of the word concerned fell within Meta's exceptions for "content that condemns a slur or hate speech, discusses the use of slurs including reports of instances when they have been used, or debates about whether they are acceptable to use." Meta accepted that the context indicated that the user was drawing attention to the hurtful nature of the word and was therefore non-violating.

In response to questions from the Board about how context is relevant in Meta's application of Hate Speech policy exceptions, Meta stated that "hate speech and slurs are allowed" when they are mocked, condemned, discussed, reported, or used self-referentially and that the responsibility is on the user to make their intent clear when mentioning a slur.

In response to another question from the Board, Meta stated that they "did not speculate" as to why the content was erroneously removed because its content reviewers do not document the reasons for their decisions.

The Board asked Meta a total of 17 questions, 16 of which were answered fully and 1 of which was answered partially.

7. Public comments

The Board received three public comments related to this case. One of the comments was submitted from the United States and Canada, one from the Middle East and North Africa, and one from Latin America and the Caribbean.

The submissions covered the following themes: LGBT safety on major social media platforms, the consideration of local context in the enforcement of the hate speech policy, and the changing meanings of Arabic words.

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

Additionally, as part of ongoing stakeholder engagement efforts, members of the Board held informative and enriching discussions with organizations that work on freedom of expression and the rights of LGBTQIA+ people, including Arabic speakers. This discussion highlighted concerns including: the difficulty in proclaiming a slur to be categorically reclaimed and universally inoffensive when the term in question may continue to be heard as a slur by some audiences, regardless of the intent of the speaker, the problems caused by a lack of input on content policy from LGBTQIA+ advocacy groups and non-English speaking communities, and the risks of content moderation which is not sufficiently sensitive to context.

8.Oversight Board analysis

The Board looked at the question of whether this content should be restored through three lenses: Meta's content policies, the company's values, and its human rights responsibilities.

This case was selected by the Board as the over-moderation of speech by users from persecuted minority groups is a serious and widespread threat to their freedom of expression. Online spaces for expression are particularly important to groups that face persecution and their rights require heightened attention for protection from social media companies. This case also demonstrates the tension for Meta in seeking to protect minorities from hate speech, while also seeking to create a space where minorities can fully express themselves, including by reclaiming hateful slurs.

8.1 Compliance with Meta’s content policies

I.Content rules

The Board finds that, while slur terms are used, the content is not hate speech because it falls into an exception in the Hate Speech policy for slur words that are “used self-referentially or in an empowering way,” as well as the exception for quoting hate speech to “condemn it or raise awareness.”

In the "Wampum Belt" and "Two buttons meme" decisions the Board noted

that it is not necessary for a user to explicitly state their intention in a post in order for it to meet the requirements of an exception to the Hate Speech

policy. It is enough for a user to be clear in the context of the post that they are using hate speech terminology in a way for which the policy allows.

However, the content in this case included the user's statements that they did not "condone or encourage" the offensive use of the slur terms in question but that the post was instead "an attempt to resist and challenge the dominant narrative" and "to reclaim the power of such hurtful terms." While clear statements of intent will not always be necessary or sufficient to legitimize the use or quotation of hate speech, they should alert a moderator to the possibility that an exception may apply. In this case, the Board finds that the statement of intent, coupled with the context, make clear that the content unambiguously falls within the exception.

Despite this, Meta initially removed the content, with three separate moderators determining that the content violated the Hate Speech policy. While there are a range of possible reasons as to how multiple moderators failed to properly classify the content, Meta was unable to provide specific explanations for the error since the company does not require moderators to record the reasoning for their decisions. As noted in the "Wampum Belt" decision, the types of mistakes and the people or communities who bear the burden of them reflect design choices for enforcement systems on the platform that risk impairing the free speech rights of members of persecuted groups. When Meta observes a pattern of persistent over-enforcement of content in relation to a persecuted or marginalized group, such as in this case, it would be appropriate to investigate the reasoning behind the enforcement decisions and consider what modifications to moderation rules, or increased training or supervision with respect to existing rules, are necessary to avoid overzealous enforcement that burdens members of groups whose expressive rights are at particular risk.

II.Enforcement action

In response to questions from the Board, Meta explained that the content was only restored to the platform because it happened to be flagged by a Meta

employee for an escalated level of review. “Escalated for review” means that, instead of the decision being revisited by at-scale review, which is often outsourced, it goes to an internal team at Meta. This appears to have required

a Meta employee to notice the removal of the content, then fill out and submit an internal webform highlighting the issue. In addition to the element of chance, systems such as these can only identify errors in content to which Meta staff are personally exposed. Accordingly, content that is not in English, content not posted by accounts with many followers in the US, or content created for and by groups not well represented within Meta is far less likely to be noticed, flagged and given the additional attention.

As part of its outreach, the Board was made aware of concerns from stakeholders that accurate enforcement of the exceptions to the Hate Speech policy requires a degree of subject-matter expertise and local knowledge that Meta may either lack or not always be able to apply. The Board shares concerns that, unless Meta regularly seeks input from minority groups targeted with slurs on a country-specific level, it will be unable to formulate nuanced lists of designated slur terms and give its moderators proper guidance on how exceptions to the slurs policy should be applied.

8.2 Compliance with Meta’s values

The Board finds that the original decision to remove this content was inconsistent with Meta's values of "Voice" and "Dignity" and did not serve the value of "Safety." While it is consistent with Meta's values to prevent the use of slurs to abuse people on its platforms, the Board is concerned that Meta is not consistently applying exceptions in the policy to expression from marginalized groups.

In the context of this case, “Voice” that seeks to promote free expression from members of a marginalized group is of the utmost importance. Meta is right to attempt to limit the use of slurs to denigrate and intimidate their targets, and also to allow good faith attempts to deprive those words of their negative impact through reclamation.

The Board recognizes that the circulation of slurs impacts “Dignity.” Particularly when used with the intent to offend or absent contextual clues

signifying that they are not being used to offend, encounters with slur words can intimidate, upset or offend users in ways that inhibit online expression. Where there are clear contextual clues that the slur is mentioned to condemn it, raise awareness for it, or mentioned self-referentially or in an empowering way, the value of "Dignity" does not dictate that the word must be removed from the platform. On the contrary, over-enforcement that ignores the exceptions particularly affects minority and marginalized groups. As recommended by the Board in the "Two buttons meme" decision, Meta must ensure that its moderators are sufficiently resourced and supported such that relevant context could be assessed properly. It is important that moderators are able to distinguish between permitted references to slurs and impermissible uses of slurs to protect the "Voice" and "Dignity" of its users, especially those from marginalized communities.

As the “Dignity” and “Safety” of marginalized communities are at a heightened level of risk on social media platforms, those platforms have heightened responsibilities to protect them. The Board has already recommended in the "Wampum Belt" decision that Meta should conduct accuracy assessments on the application of Hate Speech policy allowances. Accuracy can be improved through the training of moderators so that they are able to identify content involving discriminated communities and receive instructions to carefully assess whether exceptions to the Hate Speech policy apply. An assessment of the content, along with supporting contextual cues, should be the triggering factor for the application of these exceptions.

With regards to “Safety,” the Board also notes the particular importance of both safe online spaces and careful moderation to marginalized and threatened communities. LGBTQIA+ Arabic speakers, especially in the MENA region, face a degree of danger when openly expressing themselves online. Meta must balance the need to provide supportive arenas for this expression with ensuring that it does not over-moderate and silence people who already face censorship and oppression. While the Board acknowledges the complexity of moderation in this area, especially at scale, it is vital that platforms invest the resources required to do it properly.

8.3 Compliance with Meta’s human rights responsibilities

The Board concludes that Meta's initial decision to remove the content was inconsistent with its human rights responsibilities as a business. Meta has committed itself to respect human rights under the UN Guiding Principles on Business and Human Rights (UNGPs). Facebook's Corporate Human Rights Policy states that this includes the International Covenant on Civil and Political Rights (ICCPR).

1. Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including discussion of human rights and expression which may be regarded as "deeply offensive" (General Comment 34, para. 11). The right to freedom of expression is guaranteed to all people without discrimination as to "sex" or "other status" (Article 2, para. 1, ICCPR). This includes sexual orientation and gender identity (Toonen v. Australia (1992); A/HRC/19/41, para. 7).

This post relates to important social issues of discrimination against LGBTQIA+ people. The UN High Commissioner for Human Rights has noted concerns regarding restrictions on the freedom of expression arising from discriminatory limitations on advocacy for LGBTQIA+ rights (A/HRC/19/41, para. 65).

Article 19 requires that 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). Relying on the UNGPs framework, the UN Special Rapporteur on freedom of opinion and expression has called on social media companies to ensure that their content rules are guided by the requirements of Article 19, para. 3, ICCPR (A/HRC/38/35, paras. 45 and 70).

I. Legality (clarity and accessibility of the rules)

The requirement of legality provides that any restriction on freedom of expression is accessible and clear enough to provide guidance as to what is permitted and what is not.

The Board recommended in the "Breast cancer symptoms and nudity" case

(2020-004-IG-UA, Recommendation no. 9), the "Ocalan's isolation" case (2021-006-IG-UA, Recommendation no. 10) and the Policy Advisory Opinion on sharing private residential information (Recommendation no. 9) that Meta should clarify to Instagram users that Facebook's Community Standards apply to Instagram in the same way they apply to Facebook, with some exceptions. In the Policy Advisory Opinion, the Board recommended that Meta complete this within 90 days. The Board notes Meta's response to the Policy Advisory Opinion that, while this recommendation will be implemented fully, Meta is still working on building more comprehensive Instagram Community Guidelines clarifying their relationship with the Facebook Community Standards and cannot commit to the 90-day deadline. The Board, having reiterated this recommendation on multiple occasions, believes Meta has had sufficient time to prepare for these changes. The unclear relationship between the Instagram Community Guidelines and Facebook Community Standards is a source of continual confusion for users of Meta's platforms. Currently, while the Instagram Community Guidelines contain a link to the Facebook Community Standard on Hate Speech, it is not clear to the user that the entire Facebook Community Standard on Hate Speech, including the slurs prohibition and exceptions, applies to Instagram. Timely and comprehensive updates to the Instagram Community Guidelines remain a top priority for the Board.

With regards to the development of the slurs list, the Board reiterates the point made in the "South Africa Slurs" case (2021-011-FB-UA) that Meta should be more transparent on the procedures and criteria for developing the list. In this case, Meta explained that it defines slur lists for each established market based on "analysis and vetting from relevant internal partners such as process, markets, and content policy teams." Meta also stated that its market experts audit the slur list annually, with each term being assessed qualitatively and quantitatively, differentiating "words which are inherently offensive, even if written on their own, and words which are not inherently offensive." It is unclear to the Board when that annual review takes place, but after the Board selected this case, Meta audited the use of the word "z***l." Following this audit, the word was removed from the "Arabic" slur list while remaining on the slur list for the "Maghreb market." The Board does not know whether this audit was part of regular procedures or an ad hoc review in response to the Board's selection of this case. More generally, it is not apparent to the Board what the qualitative and quantitative assessments in annual reviews entail.

Information on the processes and criteria for development of the slur list and market designation, especially regarding how linguistic and geographic markets are distinguished, is not available to users. Without this information, the users may have difficulty assessing what words might be considered slurs, based solely on the definition of slurs in the Hate Speech policy that relies on subjective concepts such as inherent offensiveness and insulting nature (A/74/486, para. 46; see also A/HRC/38/35, para. 26).

With regards to how the slur list is enforced, Meta stated in the "South Africa Slurs" case (2021-011-FB-UA) that its "prohibition against slurs is global, but the designation of slurs is market-specific." It explained that "[i]f a term appears on a market slur list, the hate speech policy prohibits its use in that market." Meta's explanation is confusing as to whether its enforcement practices, which may be global in scope, mean that market-designated slurs are also prohibited globally. Meta explained that it defined a market as "a combination of country(ies) and language(s)/dialect(s)" and that "the division between…market[s] is primarily based on a combination of language /dialect and country of the content." Meta's content reviewers are "designated to their market based on their linguistic aptitude and cultural and market knowledge." According to Meta, this content involved the Arabic and Maghreb markets on the slur list. It was routed to these markets "based on a combination of multiple signals such as location, language, and dialect detected in the content, the type of the content and the report type." It is not sufficiently clear to the Board how the multiple signals work together to determine which markets a piece of content would engage, and whether content containing a word which is a slur in a given market would only be removed if the content relates to that market, or whether it would be removed globally. The Community Standard itself does not explain this process.

Meta should issue a comprehensive explanation of how slurs are enforced on the platform. There are multiple areas of opacity in the current policy, including whether slurs designated for particular geographies are removed from the platform only when posted in those geographies or when viewed in those geographies, or regardless of where they are posted or viewed. Meta should also explain how it handles words that are considered a slur in some settings but have an entirely different meaning, one that does not violate any

of Meta's policies, elsewhere.

The structure of the Community Standard on Hate Speech may also cause confusion. Although the prohibition on slurs appears below the heading for tier three hate speech, the Board finds it unclear whether the prohibition does belong to tier three as slurs do not necessarily target people with segregation or exclusion, which are the focus in the rest of that tier.

II.Legitimate aim

Any restriction on expression should pursue one of the legitimate aims listed in the ICCPR, which include the "rights of others." The policy at issue in this case pursued the legitimate aim of protecting the rights of others (General Comment No. 34, para. 28) to equality, protection against violence, and discrimination based on sexual orientation and gender identity (Article 2, para. 1, Article 26 ICCPR; UN Human Rights Committee, Toonen v. Australia (1992); UN Human Rights Council Resolution 32/2 on the protection against violence and discrimination based on sexual orientation and gender identity).

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 34, para. 34).

It was not necessary to remove the content in this case as the removal was a clear error which was not line with the exception in Meta's Hate Speech policies. Removal was also not the least intrusive instrument to achieve the legitimate aim because, in each review resulting in removal, the entire carousel containing 10 photos was taken down for alleged policy violations in only one of the photos. Even if the carousel had included one image with impermissible slurs not covered by an exception, removal of the entire carousel would not be a proportionate response.

Meta explained to the Board that "the post is considered violating when any photo contains a violation of the Community Standards" and "[u]nlike Facebook, it is not possible for Meta to remove a single image from an Instagram multi-photo post." Meta stated that an update to the content review tool had been proposed with the aim that reviewers could remove just the violating photo in a carousel, but the update had not been prioritized. The Board does not find this explanation clear, and believes that deprioritizing the update could lead to systemic overenforcement where entire carousels are taken down even though only parts of them are deemed violating. The Board also notes that, where a user posts the same series of photos on Facebook and Instagram, the different treatments of this kind of content on the two platforms would lead to inconsistent results which are not justified by any meaningful policy difference: if one of the photos is violating, this will cause removal of the whole carousel on Instagram, but not on Facebook.

2.Non-discrimination

Given the importance of reclaiming derogatory terms for LGBTQIA+ people in countering discrimination, the Board expects Meta to be particularly sensitive to the possibility of wrongful removal of the content in this case and similar content on Facebook and Instagram. As the Board noted in the "Wampum Belt" decision (2021-012-FB-UA) regarding artistic expression from Indigenous persons, it is not sufficient to evaluate the performance of Meta's enforcement of Facebook's Hate Speech policy as a whole – effects on particular marginalized groups must be taken into account. Under the UNGPs, "business enterprises should pay special attention to any particular human rights impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization" (UNGPs, Principles 18 and 20). For LGBTQIA+ people in countries which penalize their expression, social media is often one of the only means through which they can still express themselves freely. This is especially the case for Instagram, where the Community Guidelines permit users to not use their real name. The Board notes the same freedoms are not provided to Facebook users in the Community Standards. It would be important for Meta to demonstrate that it has undertaken human rights due diligence to ensure its systems are operating fairly and are not contributing to discrimination (UNGPs, Principle 17). The Board notes that Meta routinely evaluates the accuracy of its

enforcement systems in dealing with hate speech ("Wampum Belt" decision). However these assessments are not broken down into evaluations of accuracy

that specifically measure Meta's ability to distinguish impermissible hate speech from permitted content that attempts to reclaim derogatory terms.

The errors in this case indicate that Meta's guidance to moderators assessing references to derogatory terms may be insufficient. The Board is concerned that reviewers may not have sufficient resources in terms of capacity or training to prevent the kind of mistake seen in this case, especially in respect of content permitted under policy exceptions. In this case, Meta informed the Board that the Known Questions and Internal Implementation Standards are available in English only to "ensure standardized global enforcement" of its policies, and that "all of its content moderators are fluent in English." In the "Myanmar bot" decision (2021-007-FB-UA), the Board recommended that Meta should ensure its Internal Implementation Standards are available in the language in which content moderators review content. Meta took no further action on this recommendation, giving a similar response that its content moderators were fluent in English. The Board observes that providing reviewers with guidance in English on how to moderate content in non-English languages is innately challenging. The Internal Implementation Standards and Known Questions are often based in US-English language structures that may not apply in other languages, such as Arabic.

In the "Wampum Belt" decision (2021-012-FB-UA, Recommendation no. 3), the Board recommended that Meta conduct accuracy assessments focused on Hate Speech policy exceptions that cover expression about human rights violations (e.g. condemnation, awareness-raising, self-referential use, empowering use), and that Meta should share results of the assessment, including how these results will inform improvements to enforcement operations and policy development. The Board issued this recommendation based on its understanding that the costs of over-removal of expression about human rights violations are particularly great. The Board notes Meta's concerns with the recommendation in assessing feasibility, including (a) lack of specific categories in its policies on exceptions for areas such as human rights violations, and (b) lack of an easily identifiable sample of content that falls under Hate Speech exceptions. The Board believes these challenges can

be overcome, as Meta could focus analysis on existing Hate Speech exceptions and prioritize identifying samples of content. The Board encourages Meta to commit to implement the recommendation in the "Wampum Belt" case (2021-012-FB-UA) and welcomes updates from Meta in its next quarterly report.

9. Oversight Board decision

The Oversight Board overturns Meta's original decision to take down the content.

10. Policy advisory statement

Enforcement

1. Meta should translate the Internal Implementation Standards and Known Questions to Modern Standard Arabic. Doing so could reduce over-enforcement in Arabic-speaking regions by helping moderators better assess when exceptions for content containing slurs are warranted. The Board notes that Meta has taken no further action in response to the recommendation in the "Myanmar Bot" case (2021-007-FB-UA) that Meta should ensure that its Internal Implementation Standards are available in the language in which content moderators review content. The Board will consider this recommendation implemented when Meta informs the Board that translation to Modern Standard Arabic is complete.

Transparency

2. Meta should publish a clear explanation on how it creates its market-specific slur lists. This explanation should include the processes and criteria for designating which slurs and countries are assigned to each market-specific list. The Board will consider this implemented when the information is published in the Transparency Center.

3. Meta should publish a clear explanation of how it enforces its market-specific slur lists. This explanation should include the processes and criteria for determining precisely when and where the slurs prohibition will be

enforced, whether in respect to posts originating geographically from the region in question, originating outside but relating to the region in question, and/or in relation to all users in the region in question, regardless of the geographic origin of the post. The Board will consider this recommendation implemented when the information is published in Meta's Transparency Center.

4. Meta should publish a clear explanation on how it audits its market-specific slur lists. This explanation should include the processes and criteria for removing slurs from or keeping slurs on Meta's market-specific lists. The Board will consider this recommendation implemented when the information is published in Meta's Transparency Center.

***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. An independent research institute headquartered at the University of Gothenburg and drawing on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. The company Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world, provided linguistic expertise.

Date published
Jun 13, 2022

Platform
Instagram

Relevant Community Standard
Hate speech

Related topics
LGBT, Marginalized communities, Sex and gender equality

Region
Middle East and North Africa

Country(s) affected
Morocco, Egypt, Lebanon

Attachments
Reclaiming Arabic words public comments

Policies cited

Hate Speech