2022-003-IG-UA (Reclaiming Arabic Words Case): Meta Response to Recommendations

Recommendation 1 (no further action)

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.

Our commitment: We believe that maintaining one set of internal policy guidelines in the language in which all of our content reviewers are fluent, reinforced with supplementary lists of context-specific terms and phrases, is the best way to ensure standardized global enforcement of our rapidly evolving policies. Therefore, we will take no further action on this recommendation.

Considerations: As we shared in our response to the board's recommendation in the Post Discussing the Situation in Myanmar While Using Profanity case (2021-007-FB-UA), our Community Standardsapply to everyone, all around the world, and to all types of content. These Community Standards are currently available in 68 translations spoken by people who use Facebook and Instagram around the world.

In addition to our Community Standards, we also provide our content reviewers with detailed implementation guidance in English, in which all reviewers are fluent. Maintaining our internal review guidance in English is important for maintaining global enforcement consistency. Because this guidance rapidly evolves (it is constantly being updated with new clarifications, definitions, and language including market-specific slurs) relying on translations could lead to irregular lags and unreliable interpretations.

While we aim to be as consistent as possible in our enforcement of slurs, we also recognize that some words are offensive only in a specific language and cultural context, and we account for this in our guidance to reviewers as much as possible. This is why content reviewers also need specific language knowledge, such as any terms or phrases that could be considered an attack on a protected group in their region. Content reviewers are also supported by teams with regional and linguistic expertise.

Like the people who use our technologies, our content reviewers speak languages from regions across the globe. With this in mind, we will continue to iterate on our policies through rigorous policy development, but will not pursue translating our internal guidance at this time.

Next steps: We will have no further updates on this recommendation.

Recommendation 2 (implementing fully)

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.

Our commitment: We will share a clear overview of how we develop our market-specific slur lists here and in the "Enforcement" section of our Transparency Center.

Considerations: Our Community Standards define a slur as a word that is inherently offensive and used as an insult for a protected characteristic. Multiple teams, including policy, markets, and stakeholder engagement teams are involved in designating a slur. To create these lists, our regional teams conduct ongoing qualitative and quantitative analysis on the language and culture of their region or community (which we call a market). This includes reviewing how a word is locally and colloquially used, the prevalence of the word on our platforms, and the meaning associated with it when it is used. They may use cultural knowledge based on news articles, academic studies, and other linguistic research. Our regional teams are assisted by other experts on our

policies and operational processes. Cultural context is provided (news articles, academic articles, etc.), and at least 50 pieces of content containing that term must be collected and assessed. Once that analysis is complete, policy teams review everything provided by local markets, so that the content is assessed based on the relevant Meta policy. Once content policy teams are comfortable with a new designation, they will then loop in public policy and communications teams to review from a government regulation and local culture perspective. Market teams are responsible for ensuring that their relevant slur lists are as exhaustive and up to date as possible.

We also analyze the ways certain words are used on our platforms to determine the extent to which they meet our slur definition. For example, the use of words on our platforms may indicate some slurs that include previously unidentified variations or related terms that should be considered. We'll analyze the use of slurs across our platforms to identify those instances. Additionally, slur lists and policies include guidance on circumstances in which a given slur might be used in a permissible way, as when they are used in a clear self-referential way, when used in its alternative meaning, when discussing the use of the slurs, when reporting on the slur, when condemning the use of the slur, or when the slur is used in a explicitly positive way.

Next steps: We have provided an overview of how we develop market slur lists here and now consider this recommendation complete.

Recommendation 3 (implementing fully)

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.

Our commitment: We will share a clear overview of how we develop our market-specific slur lists here and in the "Enforcement" section of our Transparency Center.

Considerations: A language can be shared across nations and cultures, but slurs are often specific to a region or community (which we call a market). This is why we use slur lists that are specific to markets, not just languages. Across all violation areas, we have reviewers that are covering multiple regions across multiple languages (to cover all dialects as much as possible). These reviewers are assigned to queues based on language expertise and violation type skill set, so they have an informed sense of which slur lists will be most relevant for their respective content queues. Our content moderation routing incorporates both language and region to determine the appropriate reviewer(s) for content, but generally, language plays a larger role in that complex routing. For example, Southern Cone market queues will encompass content that comes from Chile, Uruguay, Argentina, and Paraguay and for which the primary language is Spanish. Each market queuing algorithm also has a condition called a “catch all.” This condition allows for all jobs in a language that is not assigned to the countries that the market covers to automatically fall in the markets most relevant for that language. For example, French language jobs that geographically originate in the South Cone Market (Argentina for instance) would fall in the French markets review queues and vice versa. When a slur in a language that is different from the rest of a piece of content appears, our scaled review technology highlights it as appearing on a different market’s slur list so that it is flagged at scale and in all market queues as potentially violating language.

Queuing algorithms account for both language and country because slurs can have caveats (or benign alternative use cases) related to what is going on in the world at the time and market context. Market context is important for reviewers in determining if a word is appearing in a permissible use case or not. If context is lacking, and in the absence of any other permissible use case, we err towards considering the content violating.

Next steps: We have shared a public overview of how we enforce upon market slur lists here and consider this recommendation complete.

Recommendation 4 (implementing fully)

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.

Our commitment: We will share a clear overview of how we develop our market-specific slur lists here and in the "Enforcement" section of our Transparency Center.

Considerations: We conduct an annual audit of our slur lists. This is done by our operational teams in collaboration with our regional market teams, who together review the slurs and reach a conclusion as to whether the word retains the offensive character that initially qualified it for the list. We also encourage our regional teams, including at-scale review partners, to continually monitor the linguistic development of their market and, based on this, propose new slurs that should be added to their market list or suggest that existing words on the list be revised. Finally, we ask the civil society and non-governmental organizations with whom we engage to provide input on what words should be considered slurs.

Next steps: We have shared a public overview of how we audit our market slur lists here and consider this recommendation complete.