2022-005-FB-UA (Mention of the Taliban in News Case) Meta Response to Recommendations
Nov. 14, 2022

Recommendation 1 (implementing fully)

Meta should investigate why the December 2021 changes to the Dangerous Individuals and Organizations policy were not updated within the target time of six weeks, and ensure such delays or omissions are not repeated. The Board asks Meta to inform the Board within 60 days of the findings of its investigation, and the measures it has put in place to prevent translation delays in future.

Our commitment: As this recommendation relates to delayed translations of a policy update, we have reviewed our Community Standards translations processes and will conduct an internal quarterly audit to help ensure that future translation updates are published on time. We will conduct our first of these audits in the coming months.

Considerations: We have investigated the cause of these translation delays and aligned on steps to prevent similar issues in the future. While the majority of our Community Standards policy language was translated within six weeks, the "Policy Rationale" section describing the aim of our Dangerous Organizations and Individuals policy was not translated at the same time because the drafts were maintained separately. In order to prevent future translation delays, we will consolidate these drafts going forward.

In addition, to help ensure that our translation updates are comprehensive and timely, we will complete an internal audit of our Community Standards translations on a quarterly basis.

Recommendation 2 (assessing feasibility)

Meta should make its public explanation of its two-track strikes system more comprehensive and accessible, especially for "severe strikes." It should include all policy violations that result in severe strikes, which account features can be limited as a result and specify applicable durations. Policies that result in severe strikes should also be clearly identified in the Community Standards, with a link to the "Restricting Accounts" explanation of the strikes system. The Board asks Meta to inform the

Board within 60 days of the updated Transparency Center explanation of the strikes system, and the inclusion of the links to that explanation for all content policies that result in severe strikes.

Our commitment: We are reviewing our strikes system to identify opportunities to make it more comprehensive, effective and accessible.

Considerations: We agree that our explanation of our penalty systems could be more comprehensive. We are reviewing our strikes system to make it more effective and proportionate, which includes assessing the types of policy violations that result in severe strikes, which account features can be limited as a result and how long any restrictions will be in place. In the process of that review, we will also try to more clearly explain this policy and make these explanations more accessible. We will provide an update when this assessment is complete.

Recommendation 3 (assessing feasibility)

Meta should narrow the definition of "praise" in the Known Questions guidance for reviewers, by removing the example of content that "seeks to make others think more positively about" a designated entity by attributing to them positive values or endorsing their actions. The Board asks Meta to provide the Board within 60 days with the full version of the updated Known Questions document for Dangerous Individuals and Organizations.

Our commitment: We are reviewing our definition of "praise" in the Dangerous Organizations and Individuals policy through an in-depth policy development process.

Considerations: We are conducting a policy development process to see if we need to better define what "praise" means in our Dangerous Organizations and Individuals policy, as we have heard feedback from the board, experts, civil society groups and users that our current definition of "praise" is too broad. This includes looking at the definition of "praise" in the Community Standards and our internal guidance. We will provide further updates on our progress in future Quarterly Updates.

Recommendation 4 (implementing fully)

Meta should revise its internal Implementation Standards to make clear that the "reporting" allowance in the Dangerous Individuals and Organizations policy allows for positive statements about designated entities as part of the reporting, and how to distinguish this from prohibited "praise." The Known Questions document should be expanded to make clear the importance of news reporting in situations of conflict or crisis and provide relevant examples, and that this may include positive statements about designated entities like the reporting on the Taliban in this case. The Board asks Meta to share the updated Implementation Standards with the Board within 60 days.

Our commitment: We are working to clarify our internal guidance on the news reporting allowance under the Dangerous Organizations and Individuals policy and our definition of "praise."

Considerations: Under the Dangerous Organizations and Individuals policy, we allow content that includes references to designated organizations and individuals to report on, condemn or neutrally discuss them or their activities. However, as mentioned in our response to Recommendation 3, there is also an ongoing policy development process to see if we need to better define what "praise" means under this policy, and that work includes better defining and explaining how we allow for news reporting. We will update our internal policy guidance to reflect any changes as a result of this policy development process and will update the board when this process is complete.

Recommendation 5 (assessing feasibility)

Meta should assess the accuracy of reviewers enforcing the reporting allowance under the Dangerous Individuals and Organizations policy in order to identify systemic issues causing enforcement errors. The Board asks Meta to inform the Board within 60 days of the detailed results of its review of this assessment, or accuracy assessments Meta already conducts for its Dangerous Individuals and Organizations policy, including how the results will inform improvements to enforcement operations, including for HIPO.

Our commitment: Based on potential refinement to our Dangerous Organizations and Individuals policy coming out of the policy development process, we plan to develop new tools that would allow us to gather more granular details about our enforcement of policy allowances. We will also assess the feasibility of gathering and publicly sharing more detailed metrics on the accuracy of policy allowance enforcement.

Considerations: The policy development process on whether we need to better define what "praise" means in the Dangerous Organizations and Individuals policy is an important first step for our implementation of this recommendation. We expect this process to inform the development of more detailed reviewer training about how to more accurately identify news reporting as it relates to this policy.

We are also exploring decision tags that will allow human reviewers to indicate when content doesn't violate because of a policy allowance. Historically, we have not required reviewers to explain why they consider content non-violating because it requires significant additional review time and limits the amount of content that can be reviewed by people. However, we are now looking into developing new decision strings and classifiers that will help track this policy allowance enforcement going forward. Because this work is contingent on the outcome of an ongoing policy development process, we will update the board on our progress in future Quarterly Updates. We have provided more details on planned improvements to the High Impact False Positive Override (HIPO) system in our response to Recommendation 6 below.

Recommendation 6 (assessing feasibility)

Meta should conduct a review of the HIPO ranker to examine if it can more effectively prioritize potential errors in the enforcement of allowances to the Dangerous Individuals and Organizations Policy. This should include examining whether the HIPO ranker needs to be more sensitive to news reporting content, where the likelihood of false-positive removals that impacts freedom of expression appears to be high. The Board asks Meta to inform the Board within 60 days of the results of its review and the improvements it will make to avoid errors of this kind in the future.

Our commitment: We are reviewing the High Impact False Positive Override (HIPO) ranker to determine if it can more effectively prioritize potential errors across policies.

Considerations: Meta's High Impact False Positive Override (HIPO) system proactively detects content we may have incorrectly actioned. Unlike content identified through user reports and appeals, we proactively identify this content before someone notifies us of a potential mistake. Once content is flagged for the HIPO system, it is evaluated by an automated mistake prevention ranker that prioritizes the content that poses the greatest risk of harm.

We are looking into ways to improve the prioritization and effectiveness of the HIPO ranker to address false positive removals. In the first part of 2023, we plan to explore potential HIPO system improvements around the criteria it uses to select and segment content to see if this results in better prioritization of potential false positives for review. We will update the board on the status of this work in future Quarterly Updates.

Recommendation 7 (assessing feasibility)

Meta should enhance the capacity allocated to HIPO review across languages to ensure that more content decisions that may be enforcement errors receive additional human review. The Board asks Meta to inform the Board within 60 days of the planned capacity enhancements.

Our commitment: We are exploring several improvements, including new cross-system resource pooling, to increase High Impact False Positive Override (HIPO) review capacity.

Considerations: We are conducting internal experiments that appear to increase HIPO accuracy and expect that product improvements based on these findings will increase human review capacity. We are conducting these tests in a range of languages, including Vietnamese, Turkish, Italian, German, Urdu, Burmese, Arabic, Indonesian and Hindi.

Based on the results of these experiments, we plan to develop a new centralized system to allocate human review resources across our various mistake prevention tools, including HIPO. By consolidating and optimizing review resources across systems, we expect this new approach to meaningfully increase HIPO review capacity. We will share an update with the board on these improvements in future Quarterly Updates.