2021-012-FB-UA Indigenous Art and Kamloops School Case Meta response to recommendations

January 7, 2022

Recommendation 1 (Implementing fully)

Provide users with timely and accurate notice of action being taken on the content their appeal relates to. Where applicable, including in enforcement error cases like this one, the notice to the user should acknowledge that the action was a result of the Oversight Board's review process. Meta should share the user messaging sent when board actions impact content decisions appealed by users, to demonstrate it has complied with this recommendation.

Our commitment: We will let people know when we reverse our initial enforcement decision about a piece of content in instances where we identified an enforcement error because of a person's appeal to the board.

Considerations: We currently notify people with standard messaging when we restore their content upon appeal or when we stand by our original decision to keep the content down. This includes when we change our enforcement action because of an error we identify as a result of an appeal to the board.

To implement the board's recommendation, we will develop specific messaging making clear that the reason we identified the enforcement error was because of a person's appeal to the board. We aim to improve people's experiences by increasing transparency and providing more specific information about our enforcement and appeal decisions.

Next steps: We anticipate that we will complete this work in the first half of 2022, and will share the new messaging with the board. We will provide an update on the progress of this work in a future Quarterly Update.

Recommendation 2 (Work Meta already does)

Study the impacts of modified approaches to secondary review on reviewer accuracy and throughput. This includes an evaluation of accuracy rates when content moderators are informed that they are engaged in secondary review and an opportunity for users to provide relevant context that may help reviewers evaluate their content, in line with the

board's previous recommendations. Meta should share the results of these accuracy assessments with the board and summarize the results in its Quarterly Updates.

Our commitment: We have researched the effects of informing reviewers that they are conducting a secondary review and giving people the opportunity to provide additional context on appeal. We are exploring additional experiments to further refine and understand this research. We've provided a summary of this work below and plan to provide more detailed information about this research to the board.

Considerations: In June 2018, we ran an experiment giving people the opportunity to request a second appeal of our initial enforcement decision. In those secondary appeals, we provided people a text box to submit freeform commentary about their content, to understand if providing additional context had an impact on the outcome of the appeal. Reviewers were aware that they were reviewing content that had been previously reviewed.

The results indicated that reviewers generally did not find the additional information provided useful. Only 2% of the information people provided supported overturning the original enforcement decision. More often than not, feedback expressed either disagreement with our Community Standards or disagreement that the person had violated our Community Standards. Reviewers found that many comments didn't contain useful information, and instead contained incomplete words or phrases, and expressions of frustration or anger. Even in cases where people provided relevant information for the reviewer, it was not always possible for the reviewer to validate the information. The researchers who ran the experiment determined that people's feedback needed to be clearer to be useful to reviewers and produce changes in enforcement outcomes.

We ran another experiment in December 2020, allowing people to select reasons for their appeal from a dropdown menu (for example, "I think my post does follow the Community Standards" and "I think Facebook misunderstood the context or intent of my post"). We also provided a freeform text box allowing people to further explain why they disagreed with an enforcement decision. The results of this experiment suggested that allowing people to provide additional context on appeal could have an impact on enforcement outcomes, but we are still exploring the format that is most useful to reviewers and impactful to outcomes.

Following up on the December 2020 experiment, we are exploring additional experiments that would offer people a dropdown menu instead of a freeform text box with options to provide more granular context on appeal, grouped by violation type (for example, a specific dropdown menu with options for Hate Speech appeals). We are still exploring the most efficient way to provide reviewers additional information to (1) maximize the accuracy of their reviews while ensuring consistency and scalability, (2) minimize the review time needed to consider additional context, and (3) minimize any additional reviewer training sessions we would need to conduct on how to consider this additional context.

Next steps: We are exploring additional experiments to further refine and understand this research, and we will share more detailed results of these analyses with the board.

Recommendation 3 (Assessing feasibility)

Conduct accuracy assessments focused on Hate Speech policy allowances that cover artistic expression and expression about human rights violations (e.g., condemnation, awareness raising, self-referential use, empowering use). This includes how the location of a reviewer impacts the ability of moderators to accurately assess hate speech and counter speech from the same or different regions. Meta should share the results of this assessment with the board, including how these results will inform improvements to enforcement operations and policy development and whether it plans to run regular reviewer accuracy assessments on these allowances, and summarize the results in its Quarterly Updates.

Our commitment: We are assessing the feasibility of conducting an experiment regarding accuracy of enforcement in hate speech allowance cases.

Considerations: We are determining the feasibility of conducting an experiment measuring the accuracy of our enforcement of hate speech allowances in automated and human review. Such an experiment would pose several challenges that we are exploring.

First, we do not have specific categories in our Community Standards or Community Guidelines for allowances for artistic expression or expression about human rights violations. As the board notes in its recommendation, our hate speech policy allowances

are for condemnation, awareness raising, self-referential use, and empowering use. Any experiment we conduct to assess the accuracy of our enforcement of hate speech allowances would have to assess the accuracy of identifying these allowances generally, rather than at the granular level the board recommends.

Additionally, we do not have an easily identifiable sample of content that falls under our hate speech allowances to test our automated and human review systems against. Typically, we do not ask content reviewers to mark the reason they label content as non-violating because it requires additional time to review each piece of content, and would mean fewer pieces of content would be subject to human review. Doing so would also require additional training to ensure reviewers consistently and accurately identify categories of non-violating content. Our automated systems do not identify content that falls into a hate speech allowance, nor do they document the reason a piece of content was found not to be hate speech. To implement this recommendation, we need to identify a sample of content to test our systems against. This work requires our policy subject matter experts to agree on the correct label for each piece of content in the sample, which needs to be large enough to produce meaningful results.

Next steps: We will share an update on the progress we have made in assessing the feasibility of this recommendation in a future Quarterly Update.