Case on the depiction of Zwarte Piet (2021-002-FB-UA)

Facebook Response
May 19, 2021

Recommendation 1 (committed to action)

Facebook should link the rule in the Hate Speech Community Standard prohibiting blackface to the company's reasoning for the rule, including harms it seeks to prevent.

Our commitment: We will add language to the Policy Rationale section of the Hate Speech Community Standard explaining why we remove harmful stereotypes like blackface.

Considerations: The Policy Rationale sections in the Community Standards contain the overall reasoning for our policies as well as specific guidance for each policy. We also provide additional details about our policy development process and the rationale behind our policies in the minutes of our Policy Forum. We do this so people can easily understand the reasons behind our policies as well as the specifics of what is and is not allowed. We want our policies to be consistent, and we do not often publish rationales for each specific policy line in our Community Standards. However, that is under consideration. At the board's request, we'll add the reasons supporting our prohibition of blackface to the Policy Rationale section of the Hate Speech Community Standard.

When a content reviewer reviews a post and determines it violates a policy, they often provide some additional data to our systems about the type of violation, but not always to the granularity of each line in the policy. Additionally, when we build technology to take automated action, it is often at the level of a policy area (e.g., Hate Speech) as it is not technologically feasible to create separate AI systems for each individual line in the policy. We understand the benefit in additional detail and will continue to explore how best to provide additional transparency.

Next steps: We will add language described above to the Policy Rationale section of the Hate Speech Community Standard.

Recommendation 2 (committed to action)

In line with the board's recommendation in the case about Armenians in Azerbaijan, the board said that Facebook should "ensure that users are always notified of the reasons for any enforcement of the Community Standards against them, including the specific rule Facebook is enforcing." In this case any notice to users should specify the rule on blackface, and also link to the above-mentioned resources that explain the harm this rule seeks to prevent. The board asked Facebook to provide a detailed update on its "feasibility assessment" of the prior recommendations on this topic, including the specific nature of any technical limitations and how these can be overcome.

Our commitment: In four decisions (Armenians in Azerbaijan, Breast Cancer Symptoms and Nudity, Nazi Quote, and this case), the board has recommended that Facebook communicate the specific rule within the Community Standard it is enforcing against. We are consolidating these recommendations into one workstream.

We've made some progress on our hate speech notifications by using an additional classifier that is able to predict what kind of hate speech is contained in the content — violence, dehumanization, mocking hate crimes, visual comparison, inferiority, contempt, cursing, exclusion and/or slurs. People using Facebook in English now receive messaging specific to the type of hate speech contained in their content, so if a person's content is removed because it contains blackface, they will receive a notification that the content was removed for being dehumanizing. We'll continue to explore more granularity and expand these notifications for hate speech to other languages in the future. We will also roll this out to Instagram users in the coming months.

Considerations: Over the past several years, we've worked to increase transparency in our messaging to people when we've removed their content. For example, we updated our notifications to tell people which Community Standard prompted us to take the post down, such as Hate Speech, Adult Nudity and Sexual Activity, etc. When a content reviewer determines a post violates a policy, they often provide some additional information about the type of violation, but not always to the granularity of each line in the policy. Additionally, when we build technology to take action against violating

content automatically, it is often at the level of a policy area (e.g., Hate Speech) as it is difficult to build sufficiently accurate systems for each individual line in a policy.

Next steps: We'll work to expand this notification for hate speech to other languages in the future, and we'll expand it to Instagram in the coming months. We'll also continue to explore how best to provide transparency to people about the action we take against content that violates our policies.