Recommendations

The board issued their binding decision for this case last month overturning our initial decision in this case. Facebook had previously reinstated this content as it did not violate our policies and was removed in error. At that time the board also issued twelve non-binding recommendations, which we are responding to in the table below.

On August 6, 2021, Facebook responded to the board's recommendations for this case.

Based on feedback we've received about our commitment levels, we're adding a new category. If the board recommends something we already do, we will classify it as ***"Work Facebook already does."*** *If the board's recommendation touches on work that we already do but changes our approach or how we prioritize it, we'll use our current commitment levels (Implementing fully or Implementing in part), and we'll explain so in our response. We will include this new categorization in future updates and responses.*

Recommendation 1 (implementing fully)

Immediately restore the misplaced 2017 guidance to the Internal Implementation Standards and Known Questions (the internal guidance for content moderators), informing all content moderators that it exists and arranging immediate training on it.

Our commitment: We will restore the policy guidance and train our content reviewers on it.

Considerations: We are refreshing and restoring the policy guidance at issue in this case. We're following our standard launch protocol for distributing the guidance to content reviewers and training them in its implementation, which we describe in our response to recommendation 8, below.

Next steps: We have begun our standard training protocol for distributing policies among content reviewers, and this guidance will be fully in place within the next two weeks.

Recommendation 2 (no further action)

Evaluate automated moderation processes for enforcement of the Dangerous Individuals and Organizations policy and where necessary update classifiers to exclude training data from prior enforcement errors that resulted from failures to apply the 2017 guidance. New training data should be added that reflects the restoration of this guidance.

Our commitment: We will take no further action on this recommendation because the policy guidance in this case does not directly contribute to the performance of automated enforcement.

Considerations: The policy guidance in this case did not directly contribute to the performance of our automated detection of Dangerous Individuals and Organizations violations, nor did its absence affect our automated systems.

When content moderators review content, they record whether the content violated our Community Standards or not. They follow a series of guiding questions to arrive at the decision of whether to take down or leave up the content under review. In addition to resulting in an enforcement action, the answers that content moderators provide in response to the guiding questions help train our systems to “learn” to predict whether new content violates our Community Standards.

The policy guidance in this case was not part of the guiding questions. Instead, it was part of additional resources provided to content reviewers that contain details and examples to help minimize the role that subjectivity or interpretation may play in assessing content. Those additional resources contribute to the accuracy and precision of the enforcement actions that content reviewers take, but do not directly contribute to the performance of our automated systems.

Next steps: Because there is no training data directly resulting from the piece of policy guidance in this case, we will take no further action on this recommendation.

Recommendation 3 (implementing in part)

Publish the results of the ongoing review process to determine if any other polices were lost, including descriptions of all lost policies, the period the policies were lost for, and steps taken to restore them.

Our commitment: We are conducting a review of our internal policy documentation in order to confirm that all of our established policies are being implemented by our reviewers.

Considerations: As we described in our response to recommendation 2 above, content reviewers can access additional policy guidance containing details and examples, which helps minimize variation in content decisions. In our response to recommendation 8, below, we explain how we provide internal guidance and training to content reviewers for any new or updated policy.

To conduct our review, our policy team is examining the guidelines available to content moderators to ensure no other policy guidance is missing from the time period in which this error occurred. We are also reviewing — and updating as needed — our existing policy enforcement guidance to ensure reviewers are consistently and accurately enforcing the most recent version of our internal policy documentation.

Next steps: We are reviewing our internal policy documentation in order to confirm that all of our established policies are being implemented by our reviewers. We anticipate completing our review by the end of this year.

Recommendation 4 (implementing fully)

Reflect on the Dangerous Individuals and Organizations “policy rationale” that respect for human rights and freedom of expression can advance the value of “Safety,” and that it is important for the platform to provide a space for these discussions.

Our commitment: We will update the policy rationale of the Dangerous Individuals and Organizations section of the Community Standards with new language that makes it clear that discussion of human rights violations and abuse, as they relate to dangerous individuals and organizations, is not a violation of our policies.

Considerations: The policy rationale of the Dangerous Individuals and Organizations Community Standard currently includes the harms this policy seeks to prevent, including violence against civilians and state actors, the dehumanization of and harm against people based on protected characteristics, and systematic criminal operations. It allows users whose intent is clear to "share content that includes references to designated dangerous organizations and individuals to report on, condemn, or neutrally discuss them." We will expand that provision to include allowances for discussing human rights violations and abuse.

Next steps: We will publish this update by the end of this year.

Recommendation 5 (implementing fully)

Add to the Dangerous Individuals and Organizations policy a clear explanation of what "support" excludes. Users should be free to discuss alleged violations and abuses of the human rights of members of designated organizations. Calls for accountability for human rights violations and abuses should also be protected.

Our commitment: We added definitions of key terms used in the Dangerous Individuals and Organizations Community Standards in June of this year in response to recommendation 2020-005-FB-UA-2 from the Nazi Quote case, including definitions and examples of "substantive support." We will provide more detailed policy guidance to our reviewers describing what "support" means, including examples of content that is allowed.

Considerations: As part of our response to recommendation 2020-005-FB-UA-2 from the Nazi Quote case, in June of this year we added definitions of key terms used in the Dangerous Individuals and Organizations Community Standards. For example, we have included definitions and examples of "praise," "substantive support," and

“representation,” as well as examples of how we apply these key terms. In addition, we created three tiers of content enforcement for different designations of severity.
In response to several recommendations from the board concerning the Dangerous Individuals and Organizations policy, we are continuing to expand the guidance we provide human reviewers for enforcing this policy. For example, in response to this recommendation, we will provide more detailed policy guidance to our reviewers describing what “support” means, outlining what type of content to leave up. And, our work in response to Recommendation 4 above will clarify for users that discussion of potential human rights abuses is allowed. Additionally, our work in response to recommendation 2020-005-FB-UA-2 from the Nazi Quote case addresses how users can make their intent clearer.
Next steps: We will update our policy guidance for our reviewers by the end of this year.

Recommendation 6 (implementing in part)
Explain in the Community Standards how users can make the intent behind their posts clear to Facebook. This would be assisted by implementing the Board’s existing recommendation to publicly disclose the company’s list of designated individuals and organizations (see: case 2020-005-FB-UA). Facebook should also provide illustrative examples to demonstrate the line between permitted and prohibited content, including in relation to the application of the rule clarifying what “support” excludes.
Our commitment: In response to 2020-005-FB-UA-2 from the Nazi Quote case, we updated the Community Standards to clarify how users can make their intent clearer, and our work in response to recommendation 4 above will expand this. We are still assessing the tradeoffs of additional transparency around our Dangerous Individuals and Organizations designations. Additionally, in response to recommendation 5 above, we are providing content reviewers with detailed definitions and examples of what “support” means, outlining what type of content to leave up.
Considerations: As part of our response to the board’s recommendation in 2020-005-FB-UA, in June of this year we added definitions of key terms used in the Dangerous Individuals and Organizations Community Standards. For example, we have included

definitions and examples of “praise,” “substantive support,” and “representation,” as well as examples of how we apply these key terms. In addition, we created three tiers of content enforcement for different designations of severity. And, we also explain that our policy is designed to allow for users who clearly indicate their intent to report on, condemn, or neutrally discuss the activities of dangerous individuals and organizations. As described in our responses to recommendations 4 and 5 above, we will expand the policy rationale of the Dangerous Individuals and Organizations section of the Community Standards to clearly state that we include allowances for discussing human rights violations and abuse. We also note that we already provide detailed policy guidance to our reviewers describing what “support” means, including examples of content that is allowed. As a result of recommendation 5, above, we will expand this.

In July 2021, we published our Q1 2021 Quarterly Update on the Oversight Board. We explained that we are still assessing the tradeoffs of additional transparency around our Dangerous Individuals and Organizations designations. Sharing this information may present safety risks to our teams and pose a tactical challenge to our ability to stay ahead of adversarial shifts. We will continue to assess how we can be more transparent about the individuals and organizations we designate, while keeping our community and employees safe.

Next steps: In response to Recommendation 4, we will update the policy rationale of the Dangerous Individuals and Organizations section of the Community Standards with new language that makes it clear that discussion of human rights violations and abuse, as they relate to dangerous individuals and organizations, are not a violation of our policies. We are updating content reviewer guidance as a result of Recommendation 5, discussed above, which we expect to complete by the end of the year. As we shared in our Q1 2021 Quarterly Update on the Oversight Board, we are still assessing the tradeoffs of additional transparency around our Dangerous Individuals and Organizations designations.

Recommendation 7 (work Facebook already does)

Ensure meaningful stakeholder engagement on the proposed policy change through Facebook’s Product Policy Forum, including through a public call for inputs. Facebook

should conduct this engagement in multiple languages across regions, ensuring the effective participation of individuals most impacted by the harms this policy seeks to prevent.

Our commitment: We'll continue conducting stakeholder engagement in the ongoing development of our Dangerous Individuals and Organizations policy.

Considerations: We continue to evaluate the most effective means of engagement on policy development, including through potentially calling for public inputs. As part of our policy development process, we consult with a broad range of stakeholders with expertise in the relevant policy area. For our ongoing development of our Dangerous Individuals and Organizations policy, we typically engage with global experts and lawyers working on the intersection of terrorism and human rights, international human rights lawyers, counter-terrorism experts, and international organizations. We also work with activists and members of civil society, including groups that sympathize with the causes for certain movements or groups that may have been designated by Facebook.

Next steps: We'll continue working with external stakeholders and evaluating the most effective means of engagement on policy development. We will have no further updates on this recommendation.

Recommendation 8 (work Facebook already does)

Ensure internal guidance and training are provided to content moderators on any new policy. Content moderators should be provided adequate resources to be able to understand the new policy, and adequate time to make decisions when enforcing the policy.

Our commitment: Content moderators receive extensive orientation and continued training on all policies, as well as updates to those policies.

Considerations: We train our content reviewers when we create or update a policy. Our global operations teams, who are subject matter experts in specific Community Standards violations, create detailed guidelines in the tool that our content reviewers

use when reviewing content. Members from our Content Policy team — who are experts in these new or updated policies — work with our internal training and operations teams to develop and deliver training sessions and resources to content reviewers on the guidelines developed by the operations team. Training materials include, among other things, the development of short videos for minor updates to policies or instructor-led training classes for more substantial policy updates. These trainings occur over a two-week period before we publish a change to our Community Standards and begin enforcing the new policy.

As we explained in our response to Recommendation 2, above, in addition to the policy guidelines, we provide content reviewers with detailed resources such as supplementary information, definitions, and examples directly in the tool they use.

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

Recommendation 9 (implementing fully)

Ensure that users are notified when their content is removed. The notification should note whether the removal is due to a government request or due to a violation of the Community Standards or due to a government claiming a national law is violated (and the jurisdictional reach of any removal).

Our commitment: In most cases, we already notify users when (1) we remove content for violations of our Community Standards and (2) we restrict access to content in particular jurisdictions on the basis of formal government reports of alleged local law violations, except where we are legally prohibited from doing so. Consistent with the board's recommendation, we are now actively working to provide more detailed notice to users when their content is restricted in response to a formal government report stating the content violates local laws. And, we are working to tell users when we remove content for a Community Standards violation on the basis of a formal government report.

Considerations: When we receive a formal government report about content that may violate local law, we first review it against our global Community Standards, just as we

would review a report from any other source. If the content violates our Community Standards, we will remove it and notify the user of the violation. Because these reports are reviewed under a standardized process in the same way and against the same policies as reports from any other source, we are not currently able to provide a different notice based on the source of the report. In addition, we may receive reports of a piece of content from multiple sources at the same time—for example, from a government and from user reports on Facebook. Such situations create additional challenges in determining whether content should be considered as removed in response to a government report.

If content reported by a government for local law violations does not violate our Community Standards, we conduct an additional legal and human rights review, and may restrict access to the content in the jurisdiction where it has been reported as unlawful. In these cases, we notify the impacted user that their content was restricted in response to a legal request, except where we are legally prohibited from doing so. We describe our process for reviewing government requests in detail in our Transparency Center.

We have been exploring ways to increase the level of transparency we provide to users and the public about formal requests we receive from governments, in line with best practices laid out by civil society efforts like the Santa Clara Principles and the Ranking Digital Rights project. We are prioritizing that work in response to this recommendation and work is now underway.

Next steps: We are beginning work to provide more detailed notice to users in two situations: (1) when we restrict their content as a result of a formal government report that it allegedly violates a local law; and (2) when we remove their content as a result of a formal government report that it violates our Community Standards. We will provide progress updates on this work in our Quarterly Updates.

Recommendation 10 (implementing fully)

Clarify to Instagram users that Facebook's Community Standards apply to Instagram in the same way they apply to Facebook, in line with the recommendation in case 2020-004-IG-UA.

Our commitment: In line with our ongoing work from recommendation 2020-004-IG-UA-2, from the Breast Cancer Symptoms and Nudity case, we are clarifying the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines.

Considerations: Our policies are applied uniformly across Facebook and Instagram, with a few exceptions — for example, people may have multiple accounts for different purposes on Instagram, while people on Facebook can only have one account using their authentic identity. As we explained in our Q1 2021 Quarterly Update on the Oversight Board, we are still working on building more comprehensive Instagram Community Guidelines to provide people with: (1) additional detail on the policies we enforce on Instagram and (2) more information about the relationship between Facebook's Community Standards and Instagram's Community Guidelines. The board's decision highlighted that we can make infrastructural improvements for handling policy updates across Instagram and Facebook.

Next steps: We are taking the time to build these systems and will provide updates on our progress by the end of the year.

Recommendation 11 (implementing fully)

Include information on the number of requests Facebook receives for content removals from governments that are based on Community Standards violations (as opposed to violations of national law), and the outcome of those requests.

Our commitment: We are actively working to provide additional transparency when we remove content under our Community Standards following a formal report by a government, including the total number of requests we receive.

Considerations: As noted in our response to Recommendation 9 above, when we receive a formal government report about content that may violate local law, we first review it against our global Community Standards, just as we would review a report from any other source. If the content violates our Community Standards, we will remove it and count it in our Community Standards Enforcement Report. These reports are reviewed under a standardized process in the same way and against the same policies

as reports from any other source. As a result, we are not currently able to report when we remove content based on a report by a government or from a Facebook user. In addition, we may receive reports of a piece of content that may violate our policies from multiple sources at the same time—for example, from a government and from user reports on Facebook. Such situations create additional challenges in determining whether content should be considered as removed in response to a government report. We have been exploring ways to increase the level of transparency we provide to users and the public about requests we receive from governments, in line with best practices laid out by civil society efforts like the Santa Clara Principles and the Ranking Digital Rights project. We are prioritizing that work in response to this recommendation.
Next steps: We are planning work that will enable us to include information on content removed for violating our Community Standards following a formal report by a government, including the number of requests we receive, as a distinct category in our Transparency Center.

Recommendation 12 (assessing feasibility)
Include more comprehensive information on error rates for enforcing rules on “praise” and “support” of dangerous individuals and organizations, broken down by region and language.

Our commitment: We are continuing to assess the feasibility of measuring and reporting enforcement and error rate data by country. In addition, we will assess whether we can capture data at the more granular violation type — such as “praise” and “support” — as a subset of the Dangerous Individuals and Organizations Community Standard.

Considerations: In line with the board’s recommendations in 2021-001-FB-FBR-18 and 2021-003-FB-UA-3, we are currently assessing the feasibility of collecting enforcement and error rates broken down in new ways, such as by country and region. There are, however, several challenges to sharing data about enforcement actions broken down by region and country. For example, reporting what location a piece of content is from is challenging because when someone in one region posts content

about another, we must determine which region takes priority. This challenge is particularly acute when it comes to groups and pages, where members, administrators, and subject matter often span countries. We are assessing how to report consistent and comprehensive data that provides meaningful transparency, while also ensuring that the information is accurate.

In addition, we are assessing whether we can create more granular categories within the Dangerous Individuals and Organizations policy, which would represent more rule-level information, such as “praise” and “support.” However, providing more granular enforcement and error data at a specific rule-level requires significant changes that may not be feasible or optimal because our current systems are not designed to capture this level of specificity.

Next steps: We are assessing how to create country- and region-level tabulations of enforcement and error data. We are beginning an assessment of whether we can capture data at a more granular violation type, such as “praise” and “support.” We will provide progress updates on this work in our Quarterly Updates.