July 2021

Facebook Q1 2021 Quarterly Update on the Oversight Board

FACEBOOK

TABLE OF CONTENTS

Introduction

We are committed to publishing regular updates to give our community visibility into our responses to the Oversight Board's independent decisions about some of the most difficult content decisions Facebook makes. These quarterly updates are designed to provide regular check-ins on the progress of this long-term work and share more about how Facebook approaches decisions and recommendations from the board. This first update, covering decisions the board issued in the first quarter of 2021, includes sections that detail (1) our content referrals to the board and (2) our progress on implementing the board's non-binding recommendations. The reports are meant to hold us accountable to the board and the public.

I. Facebook Content Referrals

In addition to providing users with direct access to appeal to the board, we regularly and proactively identify some of the most significant and difficult content decisions taken on the platform and ask the board to review them. We previously outlined how we prioritize cases we believe are significant and difficult in our Newsroom. The content at issue in these referrals generally involves real-world impact and issues that are severe, large-scale, and/or important for public discourse. Additionally, the content raises questions about current policies or their enforcement, with strong arguments on both sides for either removing or leaving up the content under review.

The process begins with an internal review of content decisions that are geographically diverse, cover questions about a wide range of policies found in our Community Standards or Community Guidelines, and represent both content removed as well as left up. Then, teams with expertise on our content policies, our enforcement processes, and specific cultural nuances from regions around the world review the candidate cases and provide feedback on both their significance and difficulty. At the end of this process, we refer the most significant and difficult content decisions to the board. The board has sole discretion to accept or decline to review the decisions referred through this process. As with user appeals, the decision the board makes about the Facebook-referred content decisions is binding on Facebook.

As of March 31, 2021, Facebook referred 26 content decision cases to the board, and the board selected three:

1. A case about supposed COVID-19 cures [link]
2. A case of a veiled threat based on religious beliefs [link]
3. A case about the decision to indefinitely suspend former US President Donald Trump's account [link]

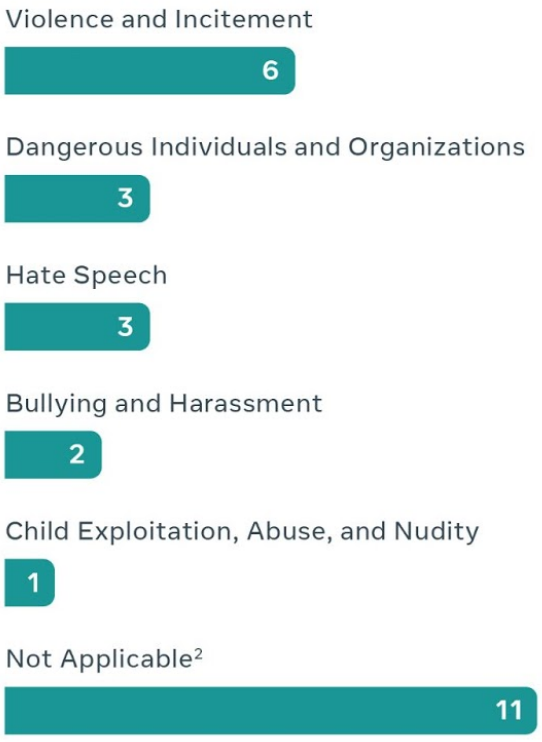
We will continue to refer content decision cases to the Oversight Board based on the process described above.

Q1 2021 Facebook referred content decision case breakdown

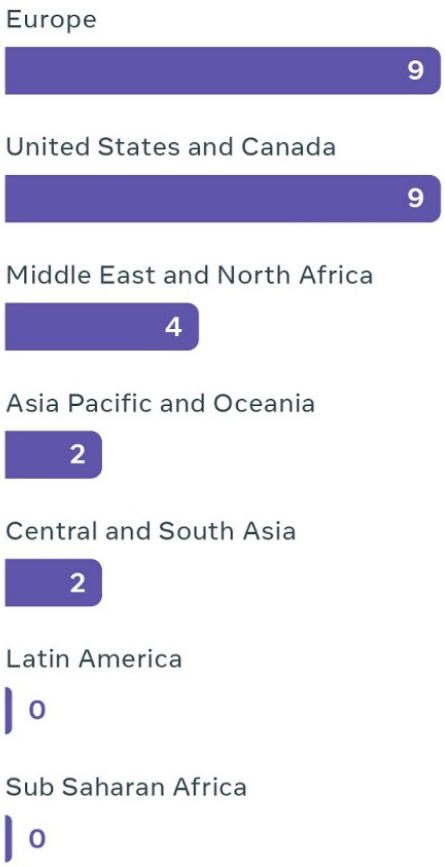
26 cases sent in total (as of March 31)[1]

3 cases selected in total (as of March 31)

POLICY VIOLATION



REGION[3]



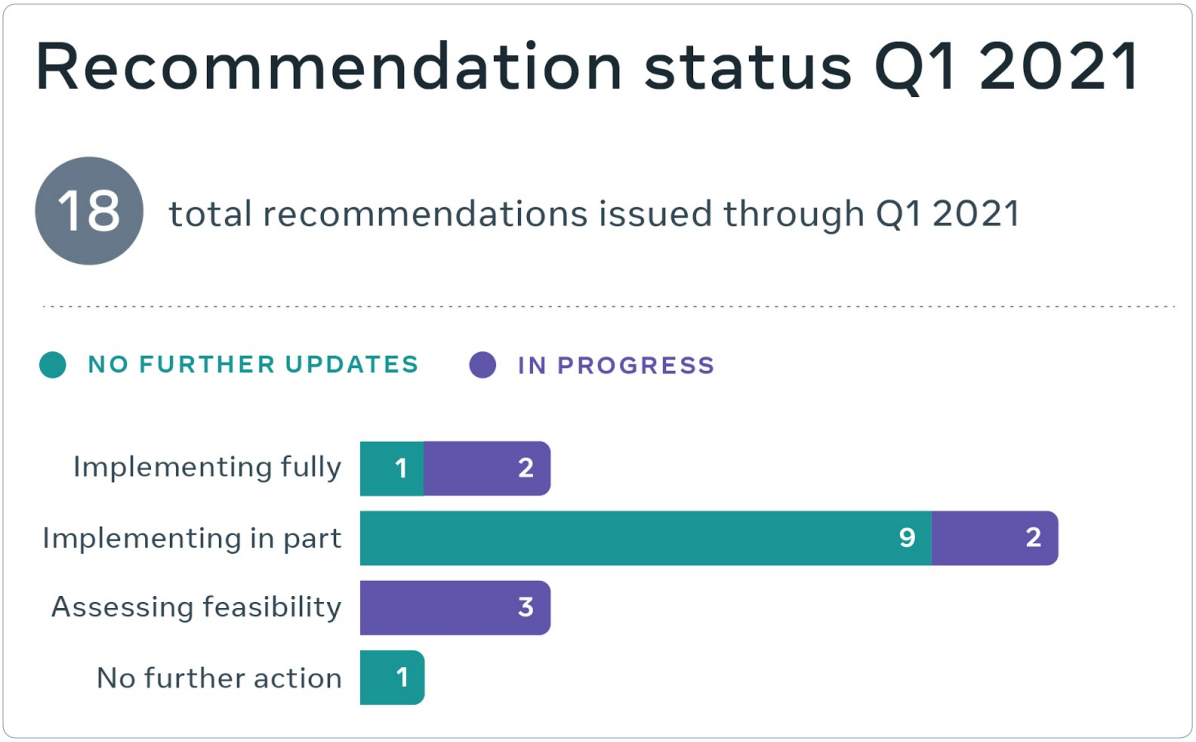
[1] As this is our first quarterly report, this number captures all referrals from November 2020 - March 2021.

[2] When we decide to leave content up, there is, by definition, no policy violation. As a result, we categorize the policy violation as "not applicable" for referrals of content we left up on Facebook and Instagram.

[3] Facebook defines "region" according to an analysis of several factors, including the location of the posting user, the language(s) the content includes, and countries/regions referenced in the content.

II. Progress on Non-Binding Recommendations

This section provides a detailed update on how we continue to address the non-binding recommendations provided by the board. In the first quarter of 2021, the board issued 18 recommendations in six cases. We are implementing fully or in part 14 recommendations, still assessing the feasibility of implementing three, and are taking no action on one.[1]



The size and scope of the board's recommendations go beyond the policy guidance that we first anticipated when we set up the board, and several require multi-month or multi-year investments. The board's recommendations touch on how we enforce our policies, how we inform users of actions we've taken and what they can do about it, and additional transparency reporting. We welcome these recommendations — the changes they have sparked make

[1] We will not include recommendations where, in a previous response or report, we shared that we would have no further updates. For this report, there is only one recommendation (2020-006-FB-FBR-7) from the hydroxychloroquine, azithromycin, and COVID-19 case where we said we would have no further updates, which is why we are providing updates on 17 of the 18 recommendations from Q1 2021.

Facebook more transparent with users and the public, more consistent with our policy applications, and more proportional in our enforcement.

For example, in the last quarter, in response to the board's recommendations, we've launched and continue to test new user experiences that are more specific about why we are removing content. We've made progress on the specificity of 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 more specific messaging when they violate our hate speech policy. We will continue to roll out more specific notifications for hate speech violations to other languages in the future. And, as a result of the board's recommendations, we're running tests to assess the impact of telling people more about whether automation was involved in enforcement. Additionally, we've updated our Dangerous Organizations and Individual policy: We have created three tiers of content enforcement for different designations of severity and added definitions of the key terms used in the policy.

We hope our responses also add to the dialogue around the challenges of content moderation at scale, by providing more insight into tradeoffs. Where we disagree in part or whole with a board recommendation — or where implementation will take a long time — we explain why.

This is our first quarterly update, which we recognize is still a work in progress. We welcome the board's feedback and review — along with the feedback from the public — of our implementation of the recommendations, as well as how we can continue to improve.

1. How to Read This Update

We designed this update in partnership with Business for Social Responsibility (BSR), based on best practices in human rights reporting principles, corporate disclosures, and goal-tracking reports. These include the Sustainability Accounting Standards Board (SASB) Conceptual Framework, International Integrated Reporting Council Framework, GRI Reporting Principles, and UN Guiding Principles for Business and Human Rights, among others.

We updated the categorization of our responses to the board's recommendations as follows:

- **Implementing fully:** Facebook agrees with the recommendation and has or will implement it in full.
- **Implementing in part:** Facebook agrees with the overall aim of the recommendation and has or will implement work related to the board's guidance.
- **Assessing feasibility:** Facebook is assessing the feasibility and impact of the recommendation.
- **No further action:** Facebook will not implement the recommendation, for example, due to a lack of feasibility or disagreement about how to reach the desired outcome.

Based on feedback from BSR and from stakeholders, we have updated our previous label of "committed to action" to one of two new labels, "implementing fully" or "implementing in part," to be clearer about what actions we are taking. We use the label "implementing in part" for one of two reasons:

- We implemented a more specific portion of a recommendation while we continue to assess the feasibility of the more general recommendation.
- We are addressing the board's guidance, but not in the method specified by the board.

Below we provide:

- The text of the recommendation
- The previous and new categorization (*e.g.*, implementing in part)
- Whether our work is in progress or we will have no further updates ("Current status")
- The text of our initial 30-day response
- New information on progress ("July 2021 Update")

2. Update on Non-Binding Recommendations

A. Case 1: Breast Cancer Symptoms and Nudity (2020-004-IG-UA)

2020-004-IG-UA-1[2]: Improve the automated detection of images with text-overlay to ensure that posts raising awareness of breast cancer symptoms are not wrongly flagged for review.

- **New category:** Implementing fully
- **Previous category:** Committed to Action
- **Current status**: In progress

Initial Response:

COMMITMENT

We agree we can do more to ensure our machine learning models don't remove the kinds of nudity we allow (e.g., female nipples in the context of breast cancer awareness). We commit to refining these systems by continuing to invest in improving our computer vision signals, sampling more training data for our machine learning, and leveraging manual review when we're not as confident about the accuracy of our automation.

CONSIDERATIONS

Facebook uses both: 1) automated detection systems to flag potentially violating content and "enqueue" it for a content reviewer and 2) automated enforcement systems to review content and decide if it violates our policies. We want to avoid wrongfully flagging posts both for review and removal, but our priority will be to ensure our models don't remove this kind of content (content wrongfully flagged for review is still assessed against our policies before any action is taken).

In this case, our automated systems got it wrong by removing this post, but not because they didn't recognize the words "breast cancer." Our machine learning works by predicting whether a piece of content violates our policies or not, including text overlays. We have observed patterns of abuse where people mention "breast cancer" or "cervix cancer" to try to confuse and/or

[2] For ease of tracking the board's recommendations, we have labeled them with the board's alphanumeric case reference (2020-004-IG-UA) and the recommendation number we assigned in our Transparency Center for our initial responses (-1).

evade our systems, meaning we cannot train our system to, say, ignore everything that says "breast cancer."

So, our models make predictions about posts like breast cancer awareness after "learning" from a large set of examples that content reviewers have confirmed either do or do not violate our policies. This case was difficult for our systems because the number of breast cancer-related posts on Instagram is very small compared to the overall number of violating nudity-related posts. This means the machine learning system has fewer examples to learn from and may be less accurate.

NEXT STEPS

We will continue to invest in making our machine learning models better at detecting the kinds of nudity we do allow. We will continue to improve computer vision signals, sampling more training data for our machine learning, and increase our use of manual review when we're less sure about the accuracy of our automation.

July 2021 Update:

We're making progress toward improving our automatic detection models by developing a new signal specific to nudity in health contexts, including breast cancer awareness. In the first part of this year, we launched keyword-based improvements to our automated systems. We are now developing an additional predictive model that will contribute more detail to the original system by identifying whether a piece of content is not only nudity, but also related to a health context. This additional layer of granularity should result in better precision when detecting non-violating, health-related nudity that does not violate our policies. We plan to launch the new model by the end of the third quarter this year.

2020-004-IG-UA-2: Revise the Instagram Community Guidelines around adult nudity. Clarify that the Instagram Community Guidelines are interpreted in line with the Facebook Community Standards, and where there are inconsistencies, the latter take precedence.

- **New category:** Implementing fully
- **Previous category:** Committed to Action
- **Current status**: In progress

Initial Response:

COMMITMENT

In response to the board's recommendations, we updated the Instagram Community Guidelines on nudity to read: "...photos in the context of breastfeeding, birth-giving and after-birth moments, health-related situations (for example, post-mastectomy, breast cancer awareness, or gender confirmation surgery), or an act of protest are allowed." We'll also clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines, including in the Transparency Center we'll be launching in the coming months (see hydroxychloroquine, azithromycin, and COVID-19 recommendation 2 for more detail).

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. We'll update Instagram's Community Guidelines to provide additional transparency about the policies we enforce on the platform. Our teams will need some time to do this holistically (for example, ensuring the changes are reflected in the notifications we send to people and in our Help Center), but we'll provide updates on our progress.

NEXT STEPS

We'll build more comprehensive Instagram Community Guidelines that provide additional detail on the policies we enforce on Instagram today. We'll also provide people with more information on the relationship between Facebook's Community Standards and Instagram's Community Guidelines.

July 2021 Update:

We have updated the Instagram Community Guidelines on nudity in line with the board's guidance. 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. We are taking the time to build these systems.

2020-004-IG-UA-3: When communicating to users about how they violated policies, be clear about the relationship between the Instagram Community Guidelines and Facebook Community Standards.

- **New category:** Implementing in part
- **Previous category:** Committed to Action
- **Current status**: No further updates

Initial Response:

COMMITMENT

We'll continue to explore how best to provide transparency to people about enforcement actions, within the limits of what is technologically feasible. We'll start with ensuring consistent communication across Facebook and Instagram to build on our commitment above to clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines.

CONSIDERATIONS

Over the past years, we've invested in improving the way we communicate with people when we remove content, and we have teams dedicated to continuing to research and refine these user experiences. As part of this work, we've updated our notifications to inform people under which of Instagram's Community Guidelines a post was taken down (for example, was it taken down for Hate Speech or Adult Nudity & Sexual Activity), but we agree with the board that we'd like to

provide more detail. As part of our response to the recommendation in the case about Armenians in Azerbaijan, we are working through multiple considerations to explore how we can provide additional transparency. In addition to confirming the need to provide more specificity about our decisions, the board's decision also highlighted the need for consistency in how we communicate across Facebook and Instagram. In this case, we did not tell the user that we allow female nipples in health contexts, but the same notification on Facebook would have included this detail. As we clarify the overall relationship between Facebook's Community Standards and Instagram's Community Guidelines, we commit to ensuring our notification systems keep up with those changes.

NEXT STEPS

We will continue to work toward consistency between Facebook and Instagram and provide updates within the next few months.

July 2021 Update:

The board has recommended, in four decisions (2020-003-FB-UA-1, 2020-004-IG-UA-3, 2020-005-FB-UA-1, and 2021-002-FB-UA-2), that Facebook communicate the specific rule within the Community Standard it is enforcing to the user. We are consolidating these four recommendations into one workstream to easily track progress. For additional information about how we are addressing these recommendations, see our response to the recommendation about user notifications in the Armenians in Azerbaijan case (2020-003-FB-UA-1) below. As discussed in our update to 2020-004-IG-UA-2, we will continue to report progress on providing people with more information on the relationship between Facebook's Community Standards and Instagram's Community Guidelines in our response to that recommendation.

2020-004-IG-UA-4: Ensure users can appeal decisions taken by automated systems to human review when their content is found to have violated Facebook's Community Standard on Adult Nudity and Sexual Activity.

- **New category:** Implemented fully
- **Previous category:** Committed to Action
- **Current status:** No further updates

Initial Response:

COMMITMENT

Our teams are always working to refine the appropriate balance between manual and automated review. We will continue this assessment for appeals, evaluating whether using manual review would improve accuracy in certain areas, and if so how best to accomplish it.

CONSIDERATIONS

Typically, the majority of appeals are reviewed by content reviewers. Anyone can appeal any decision we make to remove nudity, and that appeal will be reviewed by a content reviewer except in cases where we have capacity constraints related to COVID-19. That said, automation can also be an important tool in re-reviewing content decisions since we typically launch automated removals only when they are at least as accurate as content reviewers.

NEXT STEPS

We'll continue to monitor our enforcement and appeals systems to ensure that there's an appropriate level of manual review and will make adjustments where needed.

July 2021 Update:

We want to clarify our initial response about the role of automation in the appeal of content decisions. We use automation to help our teams prioritize content for review as part of our appeals process. In cases where we have capacity constraints, like during COVID-19, in addition to this prioritization, automation can help determine which reviews should be sent to a content reviewer and which to close without further review.

We will not have further updates related to this recommendation because typically, the majority of appeals are reviewed by content reviewers. If users appeal a decision we make to remove nudity, that appeal will be reviewed by a content reviewer, except in cases where we have capacity constraints, such as those related to COVID-19.

2020-004-IG-UA-5: Inform users when automation is used to take enforcement action against their content, including accessible descriptions of what this means.

- **New category:** Assessing Feasibility
- **Previous category:** Assessing Feasibility
- **Current status:** In progress

Initial Response:

COMMITMENT

Our teams will test the impact of telling people whether their content was actioned by automation or manual review.

CONSIDERATIONS

Over the past several years we've invested in improving the experience that we provide people when we remove content. We have teams who think about how to best explain our actions and conduct research to help inform how we can do this in a way that's accessible and supportive to people. We also need to ensure that this experience is consistent across billions of people all over the world, with differing levels of comprehension. From prior research and experimentation, we've identified that people have different perceptions and expectations about both manual and automated reviews. While we agree with the board that automated technologies are limited in their ability to understand some context and nuance, we want to ensure that any additional transparency we provide is helping all people more accurately understand our systems, and not instead creating confusion as a result of pre-existing perceptions. For example, we typically launch automated removal technology when it is at least as accurate as content reviewers. We also don't want to overrepresent the ability of content reviewers to always get it right. Additionally, many decisions made are a combination of both manual and automated input. For example, a content reviewer may take action on a piece of content based on our Community Standards, and we may then use automation to detect and enforce on identical copies. We would need to research to identify the best way of explaining these and other permutations to people.

NEXT STEPS

We will continue experimentation to understand how we can more clearly explain our systems to people, including specifically testing the impact of telling people more about how an enforcement action decision was made.

July 2021 Update:

We are continuing to assess the feasibility of this recommendation to ensure that this experience is consistent across billions of people all over the world, with differing levels of comprehension. We've launched a test on Facebook to assess the impact of telling people more about whether automation was involved in enforcement. People in the test now see whether technology or a Facebook content reviewer made the enforcement decision about their content. We will analyze the results to see if people had a clearer understanding of who removed their content, while also watching for a potential rise in recidivism and appeals rates. We expect to be able to complete our analysis by the end of the third quarter of 2021 and will provide an update on our progress.

2020-004-IG-UA-6: Expand transparency reporting to disclose data on the number of automated removal decisions, and the proportion of those decisions subsequently reversed following human review.

- **New category:** Assessing feasibility
- **Previous category:** Assessing feasibility
- **Current status:** In progress

Initial Response:

COMMITMENT

We need more time to evaluate the right approach to share more about our automated enforcement. Our Community Standards Enforcement Report currently includes our "proactive rate" (the amount of violating content we find before people report it), but we agree that we can add more information to show the accuracy of our automated review systems.

CONSIDERATIONS

The board uses the term "automation" broadly, however many decisions are made with a combination of both manual and automated input. For example, a content reviewer may take action on a piece of content based on our Community Standards, and we may then use automation to detect and enforce on identical copies. We need to align on the best way to study and report this information.

NEXT STEPS

We will continue working on this recommendation and the most appropriate and meaningful metrics reported in our Community Standards Enforcement Report that take into account the complexities of scale, technology, and manual review.

July 2021 Update:

We are continuing to identify appropriate accuracy metrics to include in the Community Standards Enforcement Report (CSER). We are still assessing how to report a consistent, comprehensible measurement about automated enforcement. We will also need additional time to assess the quality of this measurement and ensure its accuracy before adding it to CSER.

B. Case 2: Armenians in Azerbaijan (2020-003-FB-UA)

2020-003-FB-UA-1: Go beyond the Community Standard that Facebook is enforcing, and add more specifics about what part of the policy they violated.

- **New category:** Implementing in part
- **Previous category:** Assessing feasibility
- **Current status:** In progress

Initial Response:

COMMITMENT

We will continue to explore how best to provide transparency to people about enforcement actions, within the limits of what is technologically feasible.

CONSIDERATIONS

Over the past several years, we've invested in improving the experiences for people when we remove their content, and we have teams dedicated to continuing to improve these. As part of this work, we updated our notifications to inform people under which Community Standard a post was taken down (for example, Hate Speech, Adult Nudity & Sexual Activity, etc.), but we agree with the board that we'd like to provide more. 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

Our teams will continue to explore potential ways to address this challenge. We will provide updates with any future developments.

July 2021 Update:

The board has recommended, in four decisions (2020-003-FB-UA-1, 2020-004-IG-UA-3, 2020-005-FB-UA-1, and 2021-002-FB-UA-2), that Facebook communicate the specific rule within the Community Standard it is enforcing to the user. We are consolidating these four recommendations into one workstream to easily track progress.

We've made progress on the specificity of 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 more specific messaging when they violate our hate speech policy. We will continue to roll out more specific notifications for hate speech violations to other languages in the future.

As a result of the board's recommendations, we have reviewed our framework for notifying users — our integrity transparency framework — and made it more robust. We've adopted a new strategy for creating user-level transparency, which addresses questions like whether to share information, what to share, and at what level of detail. We are working on sharing this framework

with teams across the company to unify a principled approach. We will continue to update on our progress.

Additionally, earlier this year, we launched “Account Status” on Facebook, an in-product experience to help users understand the penalties Facebook applied to their accounts. It provides information about the penalties on a person’s account (currently active penalties as well as past penalties), including why we applied the penalty. In general, if people have a restriction on their account, they can see their history of certain violations, warnings, and restrictions their account might have, as well as how long this information will stay in Account Status on Facebook.

C. Case 3: Nazi Quote (2020-005-FB-UA)

2020-005-FB-UA-1: Ensure that users are always notified of the Community Standards Facebook is enforcing.

- **New category:** Implemented in part
- **Previous category:** Committed to Action
- **Current status:** No further updates

Initial Response:

COMMITMENT

We’ve fixed the mistake that led to the user not being notified about the Community Standard used for our enforcement action.

CONSIDERATIONS

People should be able to understand our decisions when we take action on their content. This is why we’ve worked to ensure a consistent level of detail is provided when content is removed from our platforms, specifically by referencing at least the Community Standard or Community Guideline in question.

NEXT STEPS

After the board surfaced this issue, we fixed the mistake.

July 2021 Update:

The board has recommended, in four decisions (2020-003-FB-UA-1, 2020-004-IG-UA-3, 2020-005-FB-UA-1, and 2021-002-FB-UA-2), that Facebook communicate the specific rule within the Community Standard it is enforcing to the user. We are consolidating these four recommendations into one workstream to easily track progress, under the Armenians in Azerbaijan case, 2020-003-FB-UA-1. The mistake specific to this recommendation has been fixed.

2020-005-FB-UA-2: Explain and provide examples of the application of key terms used in the Dangerous Individuals and Organizations policy. These should align with the definitions used in Facebook's Internal Implementation Standards.

- **New category:** Implemented in part
- **Previous category:** Committed to Action
- **Current status:** No further updates

Initial Response:

COMMITMENT

We commit to adding language to the Dangerous Individuals and Organizations Community Standard clearly explaining our intent requirements for this policy. We also commit to increasing transparency around definitions of "praise," "support," and "representation."

CONSIDERATIONS

Facebook agrees with the board that we can be clearer about how we define concepts like "praise," "support" and "representation," and we're committed to increasing transparency here. Ahead of sharing more details about these terms, we need to ensure that this information doesn't inadvertently allow bad actors to circumvent our enforcement mechanisms. Over the

next few months, our teams will determine the best way to explain these terms and how they are used in our policy.

NEXT STEPS

We will add language to our Dangerous Individuals and Organizations Community Standard within a few weeks explaining that we may remove content if the intent is not made clear. We will also add definitions of "praise," "support" and "representation" within a few months.

July 2021 Update:

We added definitions of the key terms used in the Dangerous Individuals and Organizations policy to the Community Standards. For example, we have included definitions of "praise," "substantive support," and "representation" and examples of how we apply these key terms. In addition, we created three tiers of content enforcement for different designations of severity. Tier 1, which includes terrorist, hate, and criminal organizations, results in the most extensive enforcement because we believe these entities have the most direct ties to offline harm. 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 organizations and individuals.

2020-005-FB-UA-3: Provide a public list of the organizations and individuals designated "dangerous" under the Dangerous Individuals and Organizations Community Standard.

- **New category:** Assessing feasibility
- **Previous category**: Assessing Feasibility
- **Current status:** In Progress

Initial Response:

COMMITMENT

We commit to increasing transparency around our Dangerous Individuals and Organizations Policy. In the short term, we will update the Community Standard and link to all of our Newsroom content related to Dangerous Individuals and Organizations so that people can access it with one click.

CONSIDERATIONS

Ahead of sharing more details about these terms, we need to ensure that this information will not allow bad actors to circumvent our enforcement mechanisms.

Our teams need more time to fully evaluate whether sharing examples of designations will help people better understand our policy, or if we should publish a wider list. Before publishing, we also have to be confident it will not jeopardize the safety of our employees.

NEXT STEPS

We will update the link in the Community Standards within a few weeks. We will continue to work toward more clarity on our Dangerous Individuals and Organizations policies while protecting the safety of our employees and platform.

July 2021 Update:

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.

D. Case 4: Hydroxychloroquine, Azithromycin, and COVID-19 (2020-006-FB-FBR)

2020-006-FB-FBR-1: Clarify the Community Standards with respect to health misinformation, particularly with regard to COVID-19. Facebook should set out a clear and accessible Community Standard on health misinformation, consolidating and clarifying existing rules in one place.

- **New category:** Implemented in part
- **Previous category**: Committed to action
- **Current status:** No further updates

Initial Response:

COMMITMENT

In response to the board's recommendation, we have consolidated information about health misinformation in a Help Center article, which we now link to in the Community Standards. This article includes details about all of our Community Standards related to COVID-19 and vaccines, including how we treat misinformation that is likely to contribute to imminent physical harm. We also added a "Commonly Asked Questions" section to address more nuanced situations (e.g. how humor and satire relate to these policies, how we handle personal experiences or anecdotes). We have also clarified our health misinformation policy as part of a larger COVID-19 update earlier this month. As part of that update, we added more specificity to our rules, including giving examples of the type of false claims that we will remove.

CONSIDERATIONS

Our policies and principles for enforcement of health misinformation are continuously updated to reflect the feedback we get from our global conversations with health experts.

NEXT STEPS

We'll continue to update the Help Center as necessary as our policies evolve with the pandemic.

July 2021 Update:

As we said in our initial response, in response to the board's recommendation, we consolidated information about health misinformation in a Help Center article, which we now link to in the Community Standards. We will continue to update the Help Center in line with this recommendation and will have no further updates on this recommendation.

2020-006-FB-FBR-2: Facebook should 1) publish its range of enforcement options within the Community Standards, ranking these options from most to least intrusive based on how they infringe freedom of expression, 2) explain what factors, including evidence-based criteria, the platform will use in selecting the least intrusive option when enforcing its Community Standards to protect public health, and 3) make clear within the Community Standards what enforcement option applies to each rule.

- **New category:** Implemented in part
- **Previous category:** Committed to action
- **Current status:** No further updates

Initial Response:

COMMITMENT

In the coming months, we will launch the Transparency Center. The website will be a destination for people to get more information about our Community Standards and how we enforce them on our platform, including when and why we remove violating content, and when we choose to provide additional context and labeling.

CONSIDERATIONS

As our content moderation practices have grown in sophistication and complexity, our efforts to provide people with comprehensive but clear information about our systems have to catch up. The Transparency Center is a step in this effort, building on our Community Standards to help people understand our integrity efforts overall. The Transparency Center will add more detail about what isn't allowed, as well as how we use interventions like downranking and labels for content that we think may benefit from more context.

NEXT STEPS

Launch the Transparency Center in the coming months.

July 2021 Update:

We launched the Transparency Center, which builds on our Community Standards to help people understand our integrity efforts overall. We explain how we enforce our policies, including detecting violations and taking action. We also recently published detailed information about

our strikes and penalties in response to a board recommendation in a different case. Our goal is to provide people with more information about our process for restricting profiles, pages, groups, and accounts on Facebook and Instagram. We will continue to add content to the Transparency Center to explain more about our approach to enforcement. We will have no further updates to this recommendation.

2020-006-FB-FBR-3: To ensure enforcement measures on health misinformation represent the least intrusive means of protecting public health, Facebook should clarify the particular harms it is seeking to prevent and provide transparency about how it will assess the potential harm of particular content.

- **New category:** Implemented in part
- **Previous category:** Committed to action
- **Current status:** No further updates

Initial Response:

COMMITMENT

In response to the board's guidance, we updated our Help Center to provide greater detail on the specific harms that our COVID-19 and vaccine policies are intended to address. The Help Center explains that we will "remove misinformation when public health authorities conclude that the information is false and likely to contribute to imminent violence or physical harm." As noted in the Help Center, some of these examples of imminent physical harm include "increasing the likelihood of exposure to or transmission of the virus, or having adverse effects on the public health system's ability to cope with the pandemic."

CONSIDERATIONS

For COVID-19, we assessed harm by working closely with public health authorities, who are better equipped to answer the complex question of causality between online speech and offline harm. We also consulted with experts from around the world with backgrounds in public health, vaccinology, sociology, freedom of expression, and human rights on updates we made to our policies on vaccine misinformation. These experts came from academia, civil society, public health organizations, and elsewhere. We rely on these experts to help us understand whether

claims are false and likely to contribute to the risk of increased exposure and transmission or to adverse effects on the public health system. We then remove content that includes these claims.

NEXT STEPS

We won't take any additional actions since based on the board's recommendation we've already updated our Help Center.

July 2021 Update:

In response to the board's recommendation, we have already updated the Help Center with more detail about the COVID-19 harms we seek to combat. For example, we explain that we're working to remove COVID-19 content that contributes to the risk of real-world harm, including through our policies prohibiting coordination of harm, hate speech, bullying and harassment, and misinformation that contributes to the risk of imminent violence or physical harm. Additionally, based on input from experts in health communication and related fields, we are also taking additional steps amid the pandemic to reduce the distribution of content that does not violate our policies but may present misleading or sensationalized information about vaccines in a way that would be likely to discourage vaccinations. As the situation evolves, we continue to look at content on the platform, assess speech trends, and engage with experts like the World Health Organization (WHO), government health authorities, and stakeholders from across the spectrum of people who use our service, and we will provide additional policy guidance when appropriate to keep the members of our community safe during this crisis. We will have no further updates on this recommendation.

2020-006-FB-FBR-4: To ensure enforcement measures on health misinformation represent the least intrusive means of protecting public health, Facebook should conduct an assessment of its existing range of tools to deal with health misinformation and consider the potential for development of further tools that are less intrusive than content removals.

- **New category:** Implemented in part
- **Previous category:** Committed to action
- **Current status:** No further updates

Initial Response:

COMMITMENT

We will continue to develop a range of tools to connect people to authoritative information as they encounter health content on our platforms, starting with information about COVID-19 vaccines.

CONSIDERATIONS

We continually assess and develop a range of tools, in consultation with public health experts, to address potential health misinformation in the least intrusive way depending on the risk of imminent physical harm. Our current range of enforcement tools include:

- Working with independent third-party fact-checking partners to debunk claims that are found to be false, but do not violate our Community Standards. Once third-party fact-checkers rate something as false, we reduce its distribution and inform people about factual information from authoritative sources.
- Sending notifications to people who shared false content to let them know it's since been rated false. We add a notice and an overlay to the post and show a fact-checker's articles when someone tries to share the content.
- Connecting people to authoritative information based on their behavior. For example, if someone searches for "COVID-19" or "vaccines," we will redirect them to our COVID-19 Info Center on Facebook. And, we may show educational modules to people who we know have interacted with misinformation we removed for violating our Community Standards.

These tools are part of our larger effort to respond proportionally to content, as the board recommends, while keeping people safe on the platform.

NEXT STEPS

Our immediate focus for this recommendation is to work on tools to connect people with authoritative information about COVID-19 vaccines.

July 2021 Update:

In our Transparency Center, we describe our three-part approach to content enforcement on Facebook and Instagram: remove, reduce, inform. We remove content that violates our Community Standards and Community Guidelines. We reduce distribution of certain content that creates a negative experience for users even when it doesn't quite meet the standard for removal under our policies. We may inform users through warnings and additional information from independent fact-checkers when content is potentially sensitive or misleading. As described in our response to 2020-006-FB-FBR-3, we've engaged and continue to engage with health experts like the WHO and government health authorities to inform our approach and to keep the members of our community safe during this crisis.

2020-006-FB-FBR-5: Publish a transparency report on how the Community Standards have been enforced during the COVID-19 global health crisis.

- **New category:** Implementing in part
- **Previous category:** Committed to Action
- **Current status**: In progress

Initial Response:

COMMITMENT

We will continue to look for ways to communicate the efficacy of our efforts to combat COVID-19 misinformation.

CONSIDERATIONS

We regularly publish information on the efforts we are taking to combat COVID-19 misinformation. For example, we have previously shared detailed data points on our response to COVID-19 misinformation, including the number of pieces of content on Facebook and Instagram we removed for violating our COVID-19 misinformation policies, the number of warning labels applied to content about COVID-19 that was rated by independent third-party fact-checkers, the number of visits to the COVID-19 Information Hub, and the number of people who clicked through these notifications to go directly to the authoritative health

sources. We have also shared information with the EU Commission's COVID-19 monitoring programme reports.

NEXT STEPS

We began consistently sharing COVID-19 metrics in the Spring of 2020, and we will continue to do so for the duration of the pandemic. Given the temporary and unique circumstances of COVID-19, we are not planning to add it into the Community Standards Enforcement Report as an additional policy area.

July 2021 Update:

We will continue to share COVID-19 enforcement metrics throughout the duration of the pandemic. We understand the board's recommendation also sought to address increased transparency of enforcement overall and the effectiveness of our systems during the COVID-19 health crisis. As indicated in our response to 2020-004-IG-UA-6, above, we are continuing to work on identifying appropriate accuracy metrics to include in CSER.

2020-006-FB-FBR-6: Conduct a human rights impact assessment with relevant stakeholders as part of its process of rule modification.

- **New category:** Implemented in part
- **Previous category:** Committed to action
- **Current status:** No further updates

Initial Response:

COMMITMENT

We will ask the board to clarify if its recommendation relates to all rule modifications or those related to COVID-19 misinformation. We will explore approaches to strengthen the incorporation of human rights principles into our policy development process.

CONSIDERATIONS

Facebook has a dedicated Human Rights Policy Team that consults on policy development and rule changes. Given the frequency with which we update our policies conducting a full human rights impact assessment for every rule change is not feasible. The Human Rights Policy Team, informed by authoritative guidance and an independent literature review, advised on access to authoritative health information as part of the right to health and on permissible restrictions to freedom of expression related to public health. It also participated in structuring an extensive global rights holder consultation. These elements were directly incorporated into Facebook's overall strategy for combating misinformation that contributes to the risk of imminent physical harm.

NEXT STEPS

We will ask the board to clarify if its recommendation relates to all rule modifications or those related to COVID-19 misinformation. Based on this, we will assess whether there are opportunities to strengthen the inclusion of human rights principles in our policy development process, including the possibility of additional formal human rights impact assessments.

July 2021 Update:

We now understand from the board that the recommendation was to conduct a human rights impact assessment as part of establishing a new Community Standard on health misinformation or to clarify our Community Standards with respect to health misinformation. As we described in our initial response to this recommendation, our Human Rights Policy Team was involved in the development of our policies combating health misinformation that contributes to the risk of imminent physical harm. We will have no further updates on this response.

E. Case 5: India Incitement (2020-007-FB-FBR)

2020-007-FB-FBR-1: Provide users with additional information regarding the scope and enforcement of this Community Standard. Enforcement criteria should be public and align with Facebook's internal Implementation Standards. Specifically, Facebook's criteria should address intent, the identity of the user and audience, and context.

- **New category:** Implemented in part
- **Previous category:** Committed to action
- **Current status**: No further updates

Initial Response:

COMMITMENT

We commit to adding language to the Violence and Incitement Community Standard to make it clearer when we remove content for containing veiled threats.

CONSIDERATIONS

Facebook removes explicit statements that incite violence under our Violence and Incitement Community Standard. Facebook also removes statements that are not explicit when they act as veiled or implicit threats. The language we will add to our Community Standards will elaborate on the criteria we use in this policy to evaluate whether a statement is a coded attempt to incite violence.

In its enforcement of this policy, Facebook currently does not directly use the identity of the person who shared the content or the content's full audience as criteria for assessing whether speech constitutes a veiled threat, so the added language will not include such criteria. As the board notes, we are informed by our trusted partner network to tell us when content is potentially threatening or likely to contribute to imminent violence or physical harm, so it is possible that these partners use such signals in their assessments.

NEXT STEPS

We will add language described above to the Violence and Incitement Community Standard within a few weeks.

July 2021 Update:

In April, we added language to the Violence and Incitement Community Standard to make it clearer when we remove content for containing veiled threats. We explain that we look at certain signals to determine whether there is a threat of harm in the content. For example, among other things, we look to see if the content was shared in a retaliatory context, references to historical or fictional incidents of violence, indicates knowledge of or shares sensitive information that could expose others to harm, or acts as a threatening call to action. We will have no further updates on this response.