Oversight Board

ANNUAL REPORT 2021

Contents

Preface 2
Co-Chairs' foreword 4
Executive summary 6
Applying international human rights standards to content moderation 9
How the Board considers user appeals 10
Timeline of key events 12

Case Selection

Overview 14
Cases submitted to the Board 17
Cases considered by Case Selection Committee 21

Case Decisions

Overview 22
Table: decisions published in 2021 24
Decision spotlight: First decisions 26
Decision spotlight: Decision on former President Trump's suspension 31
Decision spotlight: Ensuring respect for freedom of expression 35
Decision spotlight: Protecting users from harmful content 40
International human rights norms in the Board's decision-making process 43
Questions the Board asked Meta as part of our decisions 48

Public Comments 51

Recommendations

Overview 54
How Meta responded to our recommendations 57
From commitments to action: getting results for users 60
Holding Meta to account on cross-check 63
Engagement and outreach 65

What's Next

2022 and beyond 67

Annex

How Meta responded to and implemented our recommendations 69

Preface

The Oversight Board's mission is to improve how Meta treats people and communities around the world by applying the Facebook and Instagram content standards in a manner that protects freedom of expression and pertinent human rights standards. We do this by providing an independent check on Meta's content moderation, making binding decisions on the most challenging content issues. We deliver policy recommendations that push Meta to improve its rules and to act in a way that is principled, transparent and treats all users fairly.

By 2018, the number of users on Facebook and Instagram had grown to over two billion people. Connecting people has created enormous benefits, bringing many closer to their loved ones, strengthening existing communities and giving rise to entirely new ones. But alongside these benefits a set of enormous challenges has emerged for society. Some content poses a threat to people's safety and freedoms. Some actors spread misinformation that threatens democracy and society. Hate speech and content promoting what Meta refers to as "dangerous individuals and organizations" have the potential to divide communities and undermine the value of social media to help bring people together.

To start a conversation about what steps might be taken to create a system of governance for conducting content moderation on a global scale, in November 2018 Meta outlined a vision for a new way for people to appeal content decisions to an independent body, whose decisions would be binding. Creating an independent entity, legally and financially distinct from the company, was the logical outcome of an increasingly strongly held belief that Meta should not make so many important decisions about free expression and safety on its own.

The Oversight Board was established based on that idea – to promote the rights and interests of users by creating transparency and bringing greater accountability, consistency, and fairness to Meta's approach to content decisions.

In September 2019, a Charter was published establishing the Board's institutional independence, founding principles and purpose. The Charter empowered the Board to make binding decisions on whether certain content should be allowed on Facebook and Instagram, which Meta had to implement within seven days of publication. It also empowered the Board to issue non-binding recommendations intended to improve how the company treated people and communities around the world. Alongside the Charter, a set of Bylaws was published setting out the Board's operational procedures.

The Oversight Board comprises three interlocking components: Board Members, the Trust and the Administration. Trustees are responsible for safeguarding the Board's independent judgment and for ensuring that it operates effectively. Board Members, including four Co-Chairs, select and decide cases, make recommendations to Meta and lead our work on implementation. The Administration consists of a team of full-time staff who assist Board Members with this work.

Our Board Members are diverse leaders experienced in working on highly challenging issues, including in relation to human rights. They include academics, civil society leaders, former judges and mandate-holders from the UN and regional human rights bodies. Our Members have lived in 27 countries, speak at least 29 languages and comprise an equal number of men and women. The Trust is also made up of an equal number of men and women. The Administration staff, while small, also speaks 28 languages between them, reinforcing the Board's global approach. Around 60% of our staff are women and 40% men.

In-keeping with our commitments to transparency and accountability, we have published several quarterly transparency reports. Now with the publication of this Annual Report, we are publicly providing: a comprehensive summary of cases submitted to the Board; a summary of Board decisions and recommendations, as well as an overview of public comments. Finally, and for the first time, this Report includes an analysis of how the Board's decisions have considered or tracked the international human rights standards implicated in several representative cases.

Co-Chairs' Foreword

Catalina Botero-Marino, Jamal Greene, Michael McConnell, Helle Thorning-Schmidt
Co-Chairs of the Oversight Board

The idea that gave birth to the Board – that social media companies should not make the defining decisions on content moderation on their own – was simple to say but complex to carry out. In our first year, we started turning this idea into reality.

We created an independent appeals system accessible to billions of users around the globe. We set up a public comments process to give people a voice in our decisions. We learned, with our different nationalities, backgrounds, and viewpoints, how to deliberate cases with no easy answers. In each case, we examined the disputed content under the relevant Facebook Community Standard or Instagram Guideline, and in light of Meta's own values. In doing so, we considered whether those principles had been correctly and consistently applied, whether users had been given adequate notice and process, and whether the standards and guidelines are compatible with the international human rights norms to which Meta has committed itself.

The Board is organized around the principle that freedom of expression is an essential component of democratic society and must be respected to protect human rights. Our commitment to freedom of expression is consistent with the International Covenant on Civil and Political Rights, which states that "Everyone shall have the right to freedom of expression." All our decisions reflect that international human rights standards are an important source of authority for the Board.

This first Annual Report, which covers the period from October 2020, when we started accepting appeals, through December 2021, describes the progress we have made in improving how Meta treats users and other affected populations around the world. The emergence of a global pandemic naturally made building an institution equipped for these tasks more challenging.

There was clearly enormous pent-up demand among Facebook and Instagram users for some way to appeal content moderation decisions Meta

made, to an organization independent from Meta. Users submitted **more than a million appeals** to the Board during this period, with the vast majority of appeals to restore content to Facebook or Instagram concerning posts which supposedly violated Meta's rules on bullying, hate speech, or violence and incitement. In April 2021, users also gained the ability to appeal content to the Oversight Board which they wanted removed from Facebook or Instagram.

We **issued decisions with full, public explanations on 20 significant cases** in 2021, on issues ranging from hate speech to COVID-19 misinformation. In doing so, we took a human rights based approach to analyzing content moderation decisions and **received nearly 10,000 public comments** that helped to shape our first judgments. In many more cases, the Board's work resulted in a voluntary decision by the company to reverse wrongful content moderation decisions.

We also made **86 recommendations to Meta** that pushed the company to be more transparent about its policies. Meta's responses to our case decisions and policy recommendations are starting to improve how it treats users. Meta now gives people using Facebook in English who break its rules on hate speech more detail on what they've done wrong. The company is rolling out new messaging in certain locations telling people whether automation or human review resulted in their content being removed, and has committed to provide new information on government requests and its newsworthiness allowance in its transparency reporting. Meta also committed to translate its Community Standards into several languages spoken in India meaning that, once completed, more than 400 million more people will be able to read Facebook's rules in their native language.

While Meta committed to implement most of the recommendations we made in 2021, our next task is to ensure that the company turns its promises into actions that will improve the experience of people using Facebook and Instagram. As such, this report applies a new, data-driven approach to track how the company is implementing each of our recommendations. We are also seeking new data from Meta to allow us to understand the precise impact of our proposals on users.

At the end of our first year of making decisions, the Board has begun to build the foundations of a project that can successfully drive change within Meta – helping the company to better serve the people and communities who are affected by the extraordinary reach of the company's social media platforms. In 2022, we will build on this strong start, adding new Board Members, expanding our work to new areas, and continuing to push Meta in the right direction. We believe that the Oversight Board's early successes in holding Meta accountable demonstrate the Board's viability and provide a self-regulatory framework for extending and improving our operations in the future. In this way, the Board will be part of a collective effort by companies, governments, academia, and civil society to shape a brighter, safer, digital future that will benefit people everywhere.

Executive Summary

1.1m+

Cases were submitted to the Board by users and Meta

More than **8 in 10**

user appeals to restore content concerned Meta's rules on bullying, hate speech or violence & incitement

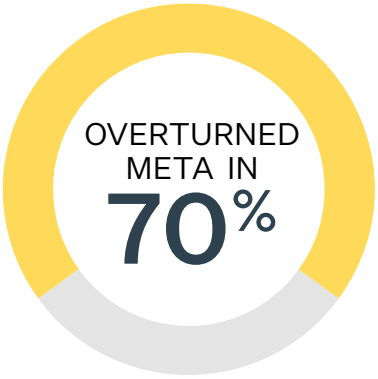
20 Decisions published

Ranging from COVID misinformation to hate speech.

OVER ½

of our decisions related to countries in the Global South

Analyzed Meta's content moderation decisions using **international human rights standards.**



of case decisions — overturning its content moderation decisions 14 times and upholding its decisions 6 times.

Nearly
10,000
public comments from people and organizations around the world shaped our decisions.

86 recommendations made to Meta

Some of the Board's recommendations that Meta committed to:

Being more specific with users when removing hate speech posts

Providing more reporting on **government requests**

Translating its rules into **languages spoken by 400+ million people**

Adopting a new **Crisis Policy Protocol** to govern its response to crisis situations

Rolling out new messaging in certain locations telling users whether **automation** or **human review** resulted in their content being removed

This report takes a **new, data-based approach** to track implementation and ensure Meta honors its commitments on recommendations.

For 2/3 of our 86 recommendations, Meta either demonstrated implementation or reported progress, with recommendations on **transparency** most likely to fall into these categories.

LOOKING TO 2022, and beyond, we are:

In dialogue with Meta about **expanding our scope,** including to review user appeals of its decisions in areas like groups and accounts.

Expanding stakeholder outreach in Asia, Latin America, the Middle East and Africa.

How the Board is Holding Meta to Account

The Board is encouraged by first-year trends in its engagement with Meta, but the company must urgently improve its transparency. We are pleased to report that the company has fully met its commitments on case decisions, has agreed to implement more than half of the Board's policy recommendations, and is increasingly answering the Board's questions.

However, the Board continues to have significant concerns, including around Meta's transparency and provision of information related to certain cases and policy recommendations.

From October '20 to December '21, Meta...

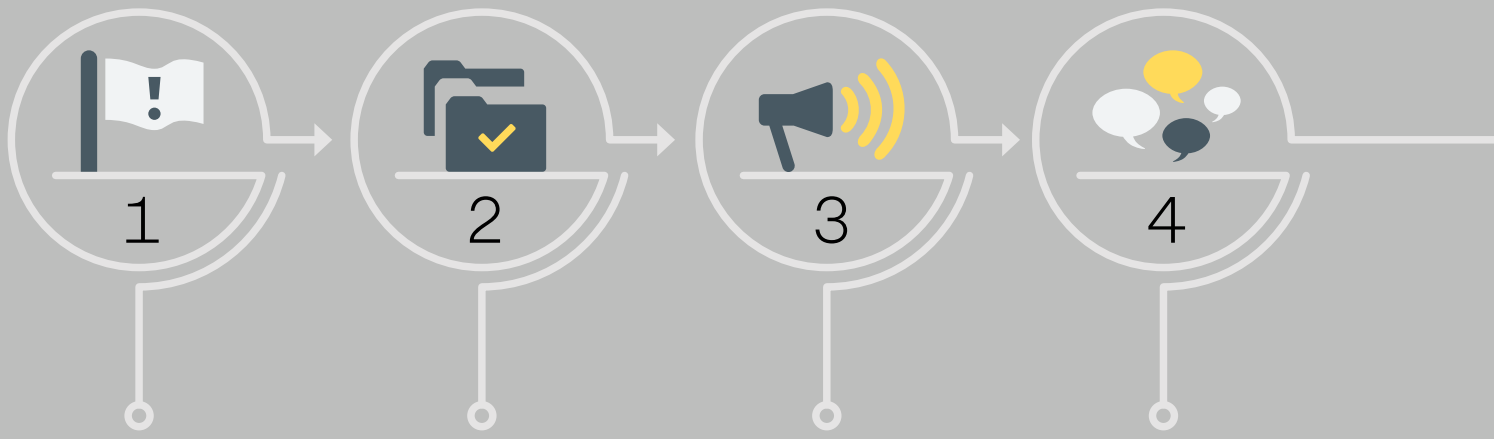
- Implemented 100% of our case decisions and committed to implement most of our recommendations.
- Answered a greater share of the questions we ask as part of our case decisions by the end of 2021, reaching 94% in Q4 2021.
- Showed improvement in how it responded to our recommendations, with the share of Meta's responses deemed either 'comprehensive' or 'somewhat comprehensive' increasing in each quarter of 2021.
- Was not fully forthcoming on information provided to the Board about its cross-check system.
- Lost and did not apply an important policy exception for three years – as described in our Öcalan's isolation decision.
- Acknowledged that in 51 of the 130 cases shortlisted for selection by the Board, its original decision was incorrect.

Applying International Human Rights Standards to Content Moderation on a Global Scale: Article 19

A defining theme of the Board's work in 2021 is a conviction that Meta will make content moderation decisions in a fairer, more principled way if it bases those decisions on the international human rights standards to which it has committed itself. To that end, the Oversight Board Charter sets out that we will "pay particular attention to the impact of removing content in light of human rights norms protecting free expression." Those norms include the International Covenant on Civil and Political Rights (ICCPR)'s Article 19, which states that while "everyone shall have the right to freedom of expression…the exercise of [that] right may…be subject to certain restrictions, but only…as provided by law and are necessary." The Article provides a three-part test for evaluating restrictions on expression:

1. Does the restriction comply with the principle of **legality**? We look at whether the rules Meta relied on in reaching its decision are accessible and sufficiently clear for users to understand and follow. It is important that rules are clear so those tasked with enforcing them can make fair and consistent decisions.
2. Does the proposed restriction have a **legitimate aim**? The Board has looked to the aims for restrictions on expression recognized in Article 19 ICCPR to assess whether the rule a decision was based on is pursuing a rights-compatible objective.
3. Was the proposed restriction **necessary and proportionate**? Was the removal of the content the least intrusive means to achieve the objective?

How the Board Considers User Appeals[1]



Appeal

Meta rejects a user's appeal on a piece of content. The user decides to appeal the case to the Board. Meta can also refer cases to the Board.

Selection

Board Members on the Case Selection Committee select the case, which is assigned to a five-member panel. Once assigned, a case should be decided by the Board and the decision should be implemented by Meta within 90 days, unless there are exceptional circumstances.

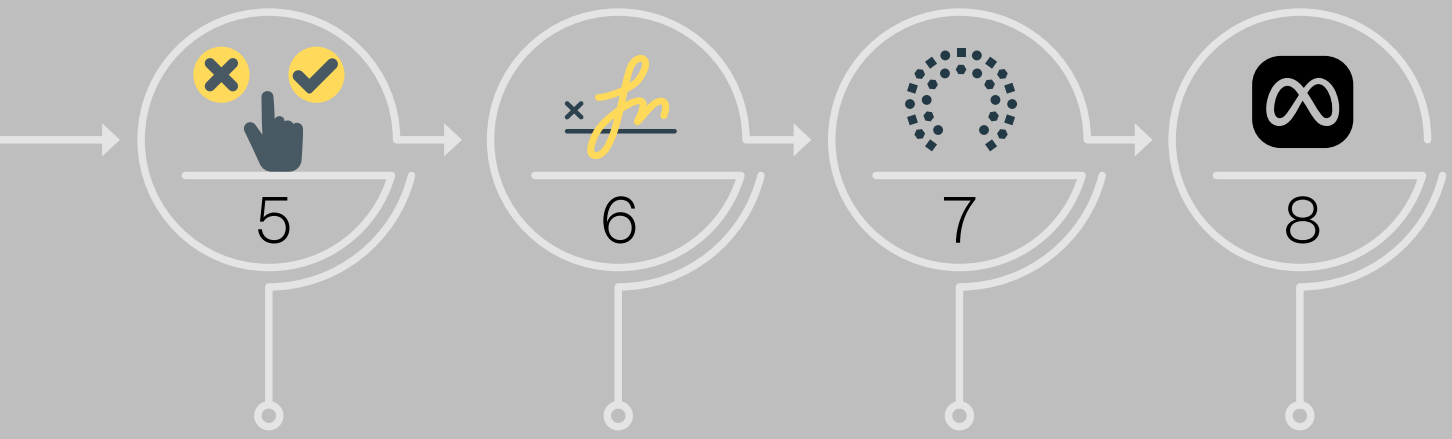
Announcement

A summary of the case is posted on the Oversight Board's website, inviting individuals and organizations to submit public comments.

Deliberation

The panel looks at whether the content violates Meta's content policies, values, and human rights standards. They consider information from the user, Meta, outside experts and public comments.

1 This graphic presents the appeals process as it applied to the 20 decisions the Board made in 2021.



Decision

The panel reaches a decision on whether to allow the content – upholding or overturning Meta.

Approval

A draft decision is circulated to all Board Members for review. A majority must sign off for a decision to be published.

Publication

Our decision is published on the Oversight Board website.

Response

In 2021, Meta had to implement our decision within seven days of publication and respond to any recommendations within 30 days. The latter deadline has since been extended to 60 days.

Timeline of Key Events

2020

OCTOBER 22

FIRST APPEALS

Board starts accepting cases from users who have exhausted the appeals process on Facebook or Instagram.

DECEMBER 1

FIRST CASES SELECTED

Select our first cases from 20,000+ submitted. These include cases from four continents covering issues including COVID misinformation, dangerous individuals and hate speech.

2021

JANUARY 28

FIRST DECISIONS

Publish our first five decisions, overturning Meta in four cases, noting the enormous complexity of the issues involved.

APRIL 13

APPEALS TO REMOVE CONTENT

Start accepting appeals from users to remove content which is live on Facebook and Instagram.

MAY 5

DECISION RELATED TO FORMER PRESIDENT TRUMP

Uphold former President Trump's suspension from Facebook and Instagram but criticize 'indefinite' penalty.

OCTOBER 21

FIRST QUARTERLY REPORTS

Publish first quarterly transparency reports, noting that the relevant team within Meta was not fully forthcoming on cross-check.

SEPTEMBER 21

RESPONSE TO CROSS-CHECK DISCLOSURES

Following press reports, we start examining whether Meta was fully forthcoming with us on its cross-check system.

JULY 8

ÖCALAN'S ISOLATION DECISION

Publish Öcalan's isolation decision, noting that Meta lost and did not apply a policy exception for three years.

JUNE 15

FIRST POLICY ADVISORY OPINION ACCEPTED

Accept policy advisory opinion on the sharing of private residential information. Policy advisory opinions allow us to suggest how Meta could improve a given policy.

MAY 26

PRO-NAVALNY PROTESTS DECISION

Publish first decision where we overturn Meta based on its human rights responsibilities, even though removing the content was in line with Facebook's rules.

Case Selection

1,152,181 Cases were submitted to the Board from October 2020 to December 2021

INCLUDING **47** from Meta

2,649

Average number of cases Board received per day

More than **8 in 10**

user appeals to restore content concerned Meta's rules on bullying, hate speech or violence & incitement.

of cases submitted were from the U.S. & Canada and Europe.

IN	OUT OF
51	130

cases shortlisted by the Board, Meta identified its original decision on the content as incorrect.

Overview

Facebook and Instagram users can challenge Meta's decisions by appealing eligible content to the Board. Between October 2020, when the Board began accepting users' appeals, through December 2021, at the close of this Report's scope, we received more than 1.1 million requests by users to independently reexamine Meta's content moderation decisions. Meta also referred 47 cases to the Board. On average, from October 2020 to December 2021, the Board received 2,649 cases a day.

The volume of cases submitted speaks to the importance of the Board's work to users. In both 2020 and 2021, we intentionally prioritized cases that had a potential to affect lots of users around the world, were critically important to public discourse or raised important questions about Meta's policies. To address issues unique to people in specific countries, we also selected cases from different regions around the world. And we selected several cases that raised major implications for applying international human rights standards to moderating content at global scale.

The Board also **listened to users** by prioritizing cases that focused on issues that were raised repeatedly in their appeals. That means that while we may only select one case that raises a particular issue for review, the Board is able to address common problems shared by a much larger number of people, including those whose cases are not selected. For example, when we saw Meta removing numerous posts referencing Nazi figures even though the content did not support or praise any "dangerous individuals or organizations," we selected the *Nazi quote* case to examine a significant issue impacting a large number of users.

To ensure that we bring a rich range of opinions to bear on our case selection process, the Board appoints a new five-member Case Selection Committee periodically to engage with a variety of widely shared user concerns. To date, and in line with our overarching criteria for selection, committees have prioritized cases where Meta's automated systems potentially moderated the content, especially in countries where users have reported a lack of human review in the original language; cases and content involving the alleged praise or promotion of "dangerous individuals and organizations"; and cases that raised important issues of press freedom, particularly for journalists facing the danger of "imminent harm" in conflict zones.

LESSONS LEARNED

Through making our selections, the Board faced several operational challenges and obstacles:

Adapting to users' free expression rights to take down their own content after case selection

One example: not long after we announced our first cases in December 2020, a user decided to delete a piece of content relating to tweets posted by the former Prime Minister of Malaysia, Dr. Mahathir Mohamad. This deletion prevented the Board from taking up the case for review. In a similar situation in late 2021, a user removed content referring to a journalist working in Afghanistan during and after U.S. troop withdrawal. In both cases, while these content removals prevented the Board from engaging with areas of common concern, we respect the right of all users to delete content for whatever reason. As we can expect such situations to occur from time to time, every time a case is withdrawn due to user action, we will announce this promptly and clarify what has occurred.

Bringing greater transparency to content removed by Meta's opaque "strikes system"

In the second quarter of 2021, two cases were assigned to a panel but not announced due to Meta taking enforcement action. A case about the origins of COVID-19 was not announced after Meta removed the page that hosted the content for receiving too many "strikes" against it. In a second case, about misgendering, Meta also removed the page that hosted the post for receiving too many "strikes."

Cases Submitted to the Board[2]

Since we started accepting cases in October 2020, the volume of cases submitted to the Board has grown each quarter until Q3 2021, with a slight dip in Q4 2021. This indicates the global prevalence of content moderation challenges for those using Facebook and Instagram.

Between Q2 and Q3 2021, the number of appeals grew by 64%. This growth is likely attributable to improvements Meta made to how users appeal to the Board through Facebook's mobile app, as well as the Board's higher profile. In one week beginning Monday, August 30, 2021, users set a record for most cases submitted in a single week, by making over 32,000 separate appeals to the Board.

Of the cases submitted, 99% related to content on Facebook,[3] while just 1% concerned content on Instagram. Starting in the second quarter of 2021, after users were empowered to appeal to the Board to remove *other* people's content, 94% of user appeals were to *restore* content to Facebook

Estimated cases submitted to the Board by quarter

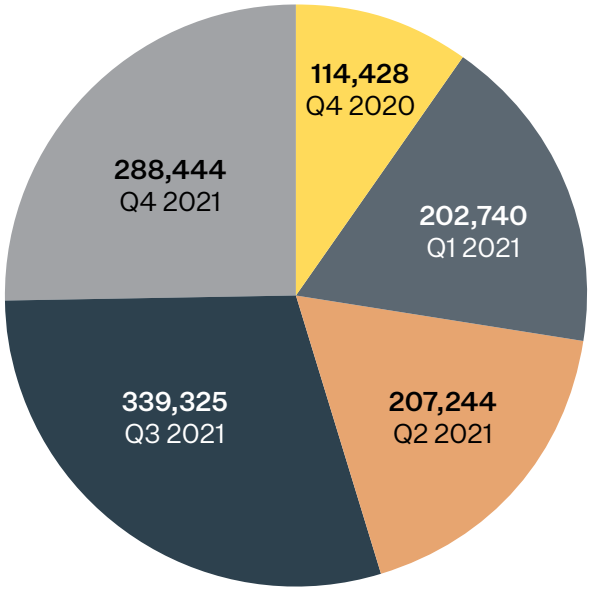


Chart: Oversight Board. Source: As seen in the Oversight Board Case Management Tool.

2 Due to limitations in the Case Management Tool, through which cases are submitted to the Board, and the data currently available to us, submitted cases were sometimes counted manually by the Board's Case Selection Team as they appeared in the CMT at the time. Breakdowns of these numbers may also be based on samples. As such, these numbers should be taken as an estimate.

3 This figure only covers Q2-Q4 2021 as data was only available for this period.

Estimated cases submitted to the Board by Community Standard, Oct '20 - Dec '21

(Only includes user appeals to restore content to Facebook and Instagram)

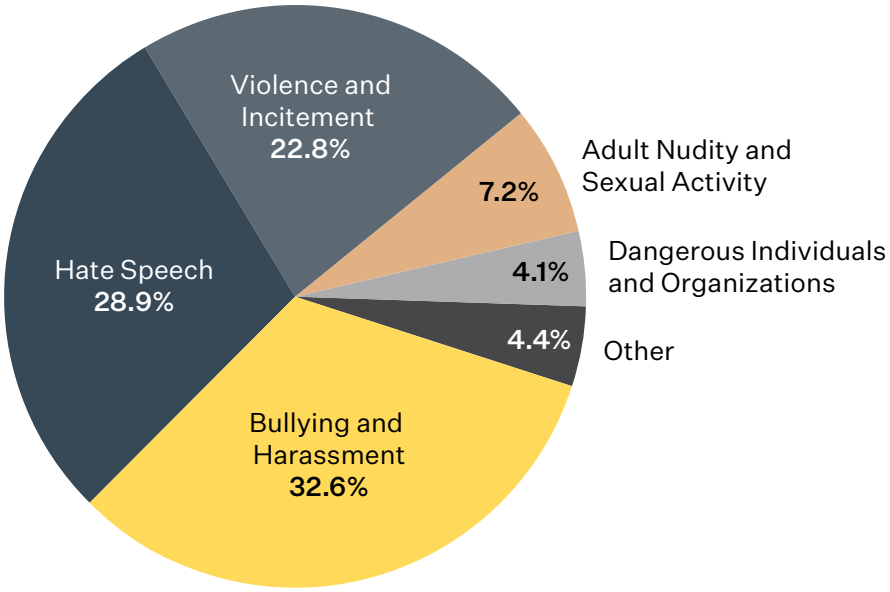


Chart: Oversight Board. Source: As seen in the Oversight Board Case Management Tool.

or Instagram, while 6% were appeals to *remove* content from the platforms.

An overwhelming majority of user appeals to restore content (84%) related to three Facebook Community Standards: Bullying and Harassment, Hate Speech, and Violence and Incitement. Cases related to Adult Nudity and Sexual Activity and Dangerous Individuals and Organizations made up most of the remainder of appeals to restore content.

To give wider context to these figures, from Q4 2020-Q4 2021 the Community Standards where Meta 'actioned' the most pieces of content (which includes removing content but also other actions such as applying warning screens) were Adult Nudity and Sexual Activity, followed by Violent and Graphic Content, and Hate Speech. The Community Standards where Meta received the most content appeals during this period, however, were Hate Speech followed by Bullying and Harassment and Adult Nudity and Sexual Activity.[4] By contrast, Bullying and Harassment followed by Hate Speech

Estimated cases submitted to the Board by Community Standard by quarter

(Only includes user appeals to restore content to Facebook and Instagram)

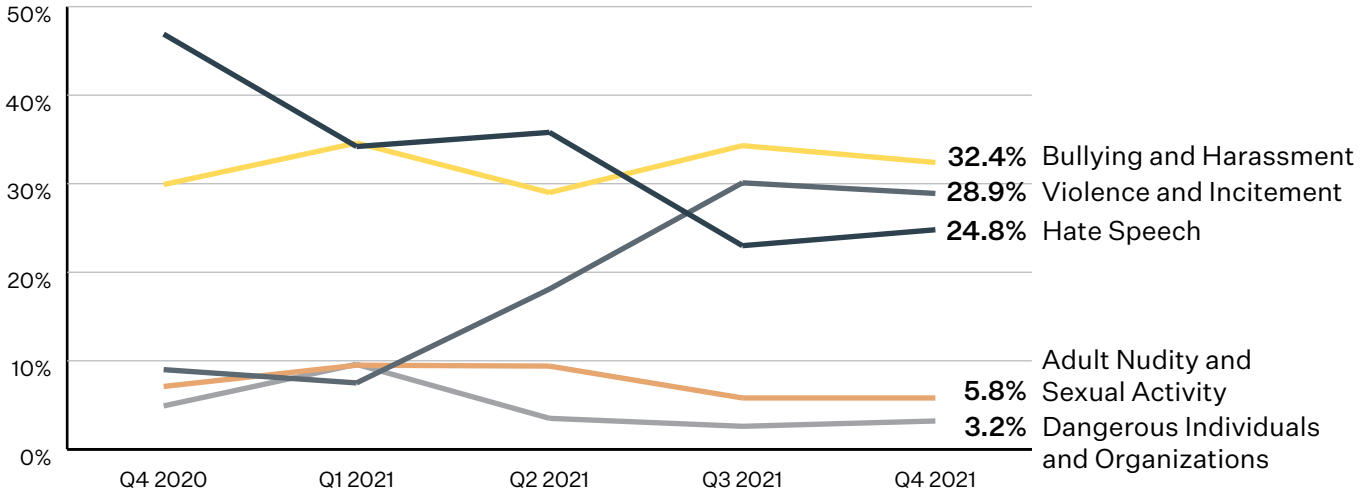
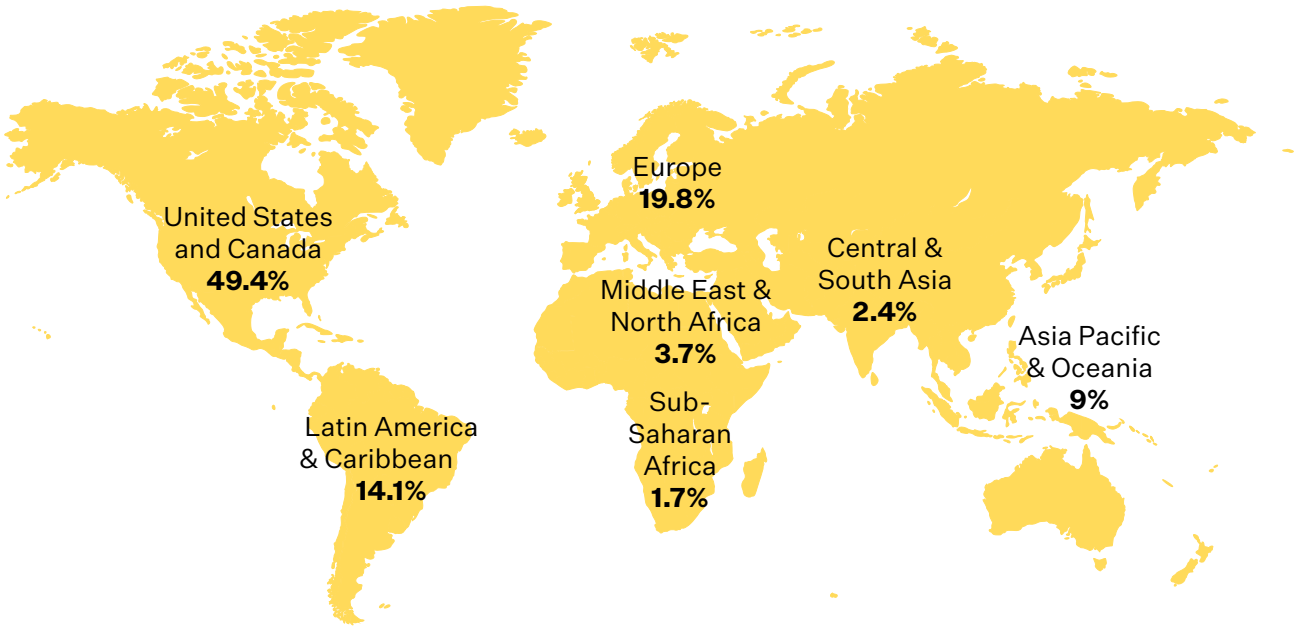


Chart: Oversight Board. Source: As seen in the Oversight Board Case Management Tool.

4 Source: Meta's Community Standards Enforcement Reports (CSER). The Community Standards mentioned here exclude policy areas outside of the Board's scope and those not mentioned in the CSER.

Estimated cases submitted to the Board by user-selected region, Oct '20 - Dec '21



Map: Oversight Board. Source: As seen in the Oversight Board Case Management Tool.

were the two Community Standards where the Board received the most user appeals to restore content during this period.

Viewed quarter by quarter, some interesting trends emerged from the data we gathered on user appeals to restore content to the two platforms. While the Bullying and Harassment Community Standard consistently represented about a third of those appeals, other Community Standards changed significantly across the year. The share of cases related to Facebook's rules on violence and incitement tripled year-on-year, rising from 9% in Q4 2020 to 29% in Q4 2021. Over the same period, the share of cases related to hate speech nearly halved, from 47% in Q4 2020 to just 25% in Q4 2021.

More than two-thirds of user appeals came from the Global North, with 49% of total appeals coming from the U.S. and Canada, and 20% coming from Europe. While 14% of appeals came from Latin America and the Caribbean and 9% from Asia Pacific and Oceania, just 4% came from the Middle East and North Africa, 2% from Central and South Asia and 2% from Sub-Saharan Africa.

We recognize that this distribution does not reflect the spread of Facebook and Instagram users worldwide. In 2019, for example, only six of the 20 countries with the most Facebook users were in Europe and North America, while India has the most Facebook and Instagram users of any country. The lower numbers of user appeals from outside Europe and the U.S. & Canada could also indicate that many of those using Facebook and Instagram in the rest of the world are not aware they can appeal Meta's content moderation decisions to the Board. We also do not believe that the distribution of appeals data on this map reflects the actual distribution of content moderation issues around the globe. If anything, we have reason to believe that users in Asia, Africa, and the Middle East experience more, not fewer, problems with Meta's platforms than other parts of the world. Our decisions so far, which covered posts from India and Ethiopia, have raised concerns about whether Meta has invested sufficient resources in moderating content in languages other than English.

In addition to expanding our outreach in regions outside Europe, the U.S. & Canada, we are also

considering and selecting many cases from outside the Global North to address the concerns of users globally. More than half the 278 cases considered for selection by the Case Selection Committee were from outside the U.S. and Canada, and Europe, while over half of the 20 decisions we published in 2021 concerned countries in the Global South.

Cases referred by Meta

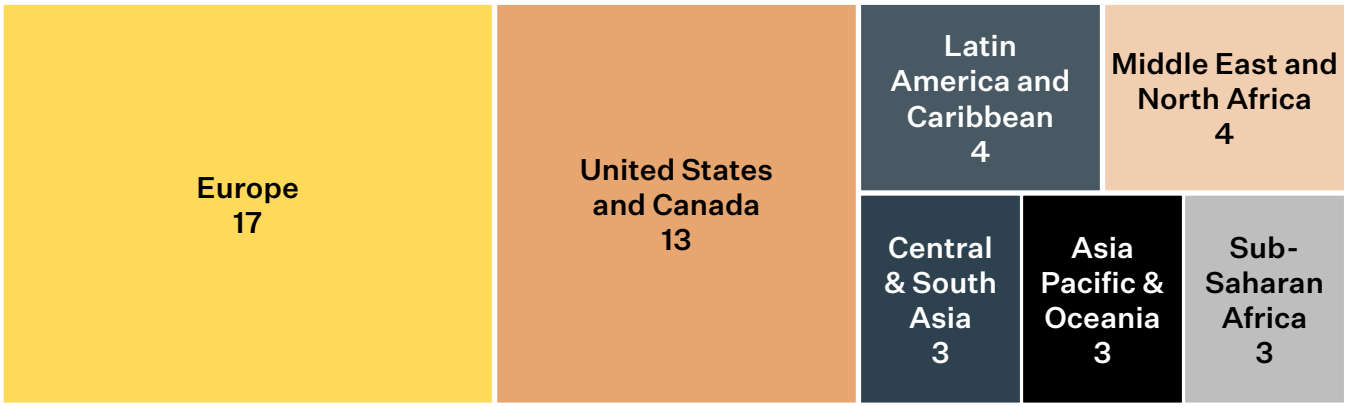
In addition to appeals from Facebook and Instagram users, Meta can refer significant and difficult cases to us.

Of the 47 cases Meta referred to the Board through December 2021, 41 reference content shared on Facebook while six involved content on Instagram. 29 of the 47 cases referred by Meta concerned content that was still live on Facebook or Instagram and had not been deemed to violate the company's rules.

The remaining 18 cases involved content Meta removed from Facebook or Instagram for allegedly violating the following Community Standards: Violence and Incitement (six cases), Hate Speech (three cases), Dangerous Individuals and Organizations (three cases), Child Sexual Exploitation, Abuse and Nudity (three cases), Bullying and Harassment (two cases) and Adult Nudity and Sexual Activity (one case).

Nearly two-thirds of the cases Meta referred to the Board related to content from Europe and the United States and Canada, with 30 of the 47 cases from these regions. Meta referred 17 cases covering the rest of the world: four from Latin America and the Caribbean, four from the Middle East and North Africa, three from Central and South Asia, three from Asia Pacific and Oceania and three from Sub-Saharan Africa.

Meta referred cases by region



Cases Considered by Case Selection Committee

To address issues facing Facebook and Instagram users across the globe, the Board selects a diverse range of cases in many different languages. The 278 cases considered by the Case Selection Committee through December 2021 covered more than 70 countries, ranging from Fiji to Chad, and Trinidad & Tobago. These cases featured content in 38 different languages.

After the Board's Case Selection Committee shortlists cases for Board review, Meta sometimes determines that its original decision on a piece of content was incorrect. This is known as an 'enforcement error.'

The chart on this page shows cases which Meta identified as enforcement errors broken down by quarter. In both Q4 2020 and Q2 2021, Meta found that in one third of shortlisted cases its original decision was incorrect. In Q1 2021, this rate was 44%, while in Q3 2021 it was 46%. The Board did not shortlist any cases in Q4 2021.

Of the 130 cases shortlisted by the Board up until December 2021, Meta identified 51 occasions where its original decision on the content was incorrect. While this is only a small sample, and the Board intentionally seeks out challenging cases, it is concerning that in just under 4 out of 10 shortlisted cases Meta found its decision to have been incorrect. This high error rate raises wider questions both about the accuracy of Meta's content moderation and the appeals process Meta applies before cases reach the Board.

Enforcement errors per number of shortlisted cases

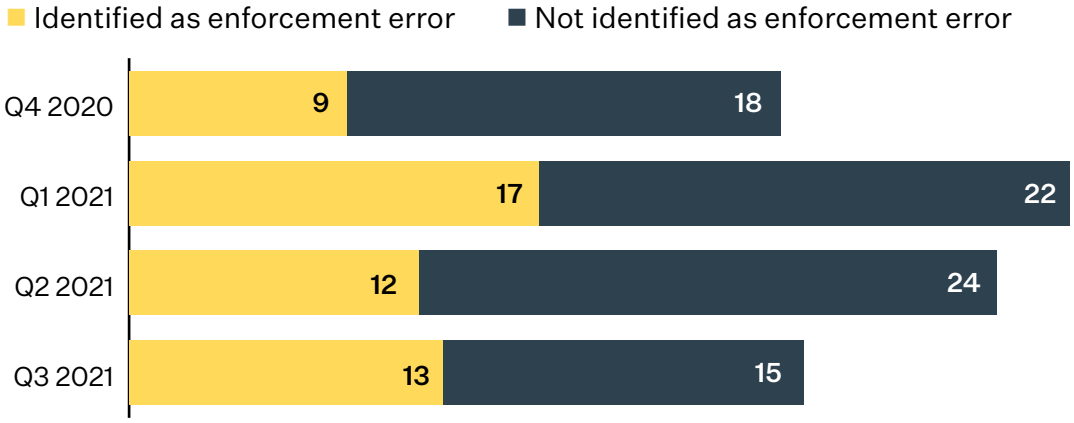


Chart: Oversight Board. Source: As seen in the Oversight Board Case Management Tool.

Case Decisions

20
Decisions published in 2021

14 Decisions Overturned Meta

6 Decisions Upheld Meta

16 Cases from Users

4 Cases from Meta

17 Cases from Facebook

3 Cases from Instagram

83 Average days to decide and implement case

OVER ½ of our decisions related to Global South countries.

313 Questions asked to Meta as part of our case review

Analyzed Meta's content moderation decisions using **international human rights standards**

Overview

The Oversight Board exists to make binding and independent decisions on the most challenging content issues facing Facebook and Instagram. Our first 20 full decisions and related policy recommendations repeatedly raised key issues about how the company treats users worldwide. The Board not only overturns Meta's decisions where necessary, but also recommends policy changes that will improve the treatment of users and provide greater transparency. Collectively, these actions advocate for users in cases where company policies are unclear, inconsistent, or not applied. Major themes covered by our decisions included:

- **Protecting freedom of expression and other human rights.** Although Meta, a company that serves as a primary network of communication for billions of people, deems the value of voice to be "paramount," freedom of expression must be limited when necessary to protect against real harms to users and others. Reconciling those concerns is the most important, and the most difficult, of the Board's responsibilities.
- **Considering meaning in context.** Our decisions often rested on questions of context. These included the meaning of a particular word or the post itself, as well as the circumstances in which the content was posted, such as violent conflict or political protests.
- **Being clear with users.** We decided a case where Meta didn't tell a user what Community Standard they'd broken, a case where the company didn't review the user's appeal and a case where Meta lost a piece of policy guidance for three years. Three decisions looking at content from Instagram also highlighted that the relationship between Facebook and Instagram's rules are not clear to users.
- **Treating users fairly no matter where they are.** Our decisions highlighted that Meta's rules for Facebook and Instagram are not available in all user languages. We also raised concerns about whether Meta was investing sufficient resources in moderating content in languages other than English.
- **Ensuring that Meta's rules are enforced correctly.** Our decisions made clear that Meta needs to improve the accuracy of the automated and human aspects of content moderation and be more transparent with the Board and public on why mistakes happen.

LESSONS LEARNED

Revising timeframes to reflect operational realities

As we drafted our first decisions, it quickly became clear that the original timeline of 90 days starting from Meta's last decision on a case would not be sufficient. To ensure that all cases had the same amount of time for deliberation, in March we changed the starting point for the 90-day timeline for a case to be decided and implemented to when a case was assigned to a panel. As the Board developed its decision-making processes and started asking Meta more questions, the 90-day timeline was changed again in November to start when a case is announced.

Refining our processes and standards

As we deliberated on our first cases, we grappled with several complex considerations. For example, when looking at a piece of content, should we apply all the information provided to us through our case review process, or instead assess only whether Meta's decision was a reasonable response given the circumstances of the review and information available to it at the time? Also, how should the Board apply international human rights standards largely developed for states, to decisions made by social media platforms? The human rights section of this chapter discusses this issue in greater detail.

DECISIONS PUBLISHED IN 2021

	PLATFORM	SOURCE	COMMUNITY STANDARD	COUNTRIES	BOARD'S DECISION
Myanmar post about Muslims Case No.: 2020-002-FB-UA		User		Myanmar, France, China	**Overturned** Meta's decision to remove
Armenians in Azerbaijan Case No.: 2020-003-FB-UA		User		Armenia, Azerbaijan	**Upheld** Meta's decision to remove
Breast Cancer Symptoms and Nudity Case No.: 2020-004-IG-UA		User		Brazil	**Overturned** Meta's decision to remove
Nazi Quote Case No.: 2020-005-FB-UA		User		United States	**Overturned** Meta's decision to remove
Claimed COVID Cure Case No.: 2020-006-FB-FBR		Meta		France	**Overturned** Meta's decision to remove
Protest in India Against France Case No.: 2020-007-FB-FBR		Meta		India, France	**Overturned** Meta's decision to remove
Former President Trump's Suspension Case No.: 2021-001-FB-FBR		Meta		United States	**Upheld** Meta's decision to remove
Depiction of Zwarte Piet Case No.: 2021-002-FB-UA		User		Netherlands	**Upheld** Meta's decision to remove
Punjabi Concern Over the RSS in India Case No.: 2021-003-FB-UA		User		India	**Overturned** Meta's decision to remove
Pro-Navalny Protests in Russia Case No.: 2021-004-FB-UA		User		Russia	**Overturned** Meta's decision to remove

Dangerous Individuals and Organizations

Violence and Incitement

Hate Speech

Adult Nudity and Sexual Activity

Bullying and Harassment

	PLATFORM	SOURCE	COMMUNITY STANDARD	COUNTRIES	BOARD'S DECISION
"Two buttons" meme Case No.: 2021-005-FB-UA		User		Armenia, Turkey, United States	**Overturned** Meta's decision to remove
Öcalan's isolation Case No.: 2021-006-IG-UA		User		Turkey, United States	**Overturned** Meta's decision to remove
Myanmar Bot Case No.: 2021-007-FB-UA		User		Myanmar	**Overturned** Meta's decision to remove
COVID lockdowns in Brazil Case No.: 2021-008-FB-FBR		Meta	N/A	Brazil	**Upheld** Meta's decision to leave up
Shared Al Jazeera post Case No.: 2021-009-FB-UA		User		Israel, Egypt	**Overturned** Meta's decision to remove
Colombia Protests Case No.: 2021-010-FB-UA		User		Colombia	**Overturned** Meta's decision to remove
South Africa Slurs Case No.: 2021-011-FB-UA		User		South Africa	**Upheld** Meta's decision to remove
Wampum Belt Case No.: 2021-012-FB-UA		User		Canada	**Overturned** Meta's decision to remove
Ayahuasca Brew Case No.: 2021-013-IG-UA		User		Brazil	**Overturned** Meta's decision to remove
Alleged Crimes in Raya Kobo Case No.: 2021-014-FB-UA		User		Ethiopia	**Upheld** Meta's decision to remove

Dangerous Individuals and Organizations

Hate Speech

Regulated Goods

DECISION SPOTLIGHT

First Decisions

> Up until now it's been up to [Mark] Zuckerberg to make content decisions... It's a historic day. It's the first time that content moderation has been done outside of Facebook and Facebook has to follow our decisions."
>
> Helle Thorning-Schmidt
> *Oversight Board Co-Chair*

On January 28, 2021, the Oversight Board published five decisions, based on deliberations conducted since we selected our first cases from 20,000 appeals. Board Members conducted deliberations according to a process outlined in our Bylaws: each case was assigned to a five-Member panel including at least one Member from the region implicated in the content, with mixed-gender representation. Once the panel had reached a decision, a draft was circulated to all Board Members for comments, and then put to a vote by the full Board. The cases covered content across four continents: Asia, Europe, North America, and South America.

Of the first five cases, four overturned Meta's decisions and one upheld Meta's decision to remove a post. The cases explored a wide range of issues. In one case, Board Members examined whether, in the context of an armed conflict, Meta was right to remove an otherwise-permissible post because it contained a hateful slur. In another, they examined whether a post accused of spreading COVID-19 misinformation contributed to imminent harm. Another decision set an important precedent by highlighting Meta's automated enforcement systems' inability to distinguish permitted images of uncovered female nipples in a breast cancer awareness campaign.

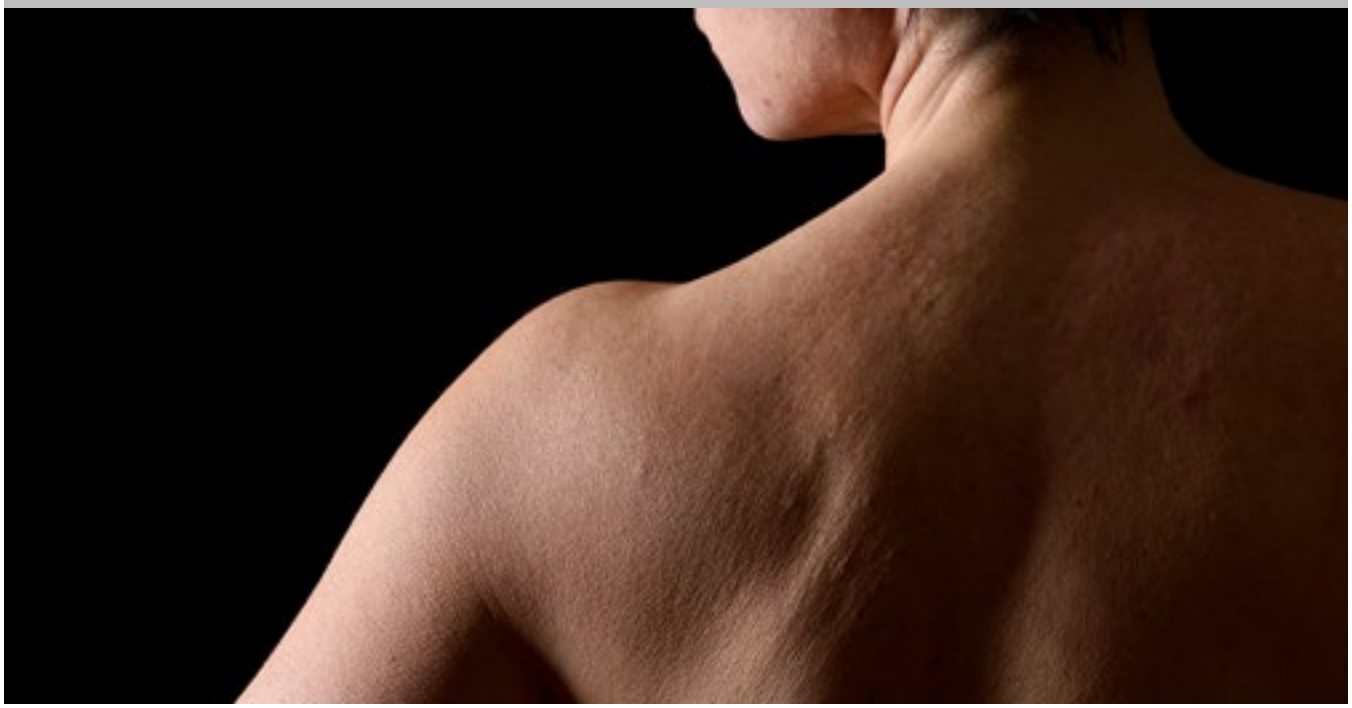
Breast Cancer Symptoms and Nudity

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
January 28, 2021	Brazil	Adult Nudity and Sexual Activity

In October 2020, a user in Brazil posted a picture to Instagram with a title, in Portuguese, intended to raise awareness of breast cancer. The post was removed by an automated system enforcing Facebook's Community Standard on Adult Nudity and Sexual Activity, which the company explained also applies to Instagram. We overturned Meta's decision because the Community Standard includes a clear exception that allows nudity when the user seeks to "raise awareness about a cause for educational and medical reasons," and specifically permits the posting of images of uncovered female nipples to "advance breast cancer awareness."

After the content was removed, the user appealed this decision to Meta. In public statements, Meta had previously said that it could not always offer users the option to appeal due to a temporary reduction in its review capacity as a result of COVID-19. Moreover, Meta had also stated that not all appeals will receive human review.

We rejected Meta's argument that the case was "moot" because, having restored the post, the user and Meta no longer disagreed that the content should stay up. We made clear that the Board has the authority to review cases from users even when Meta chooses to later rectify its mistake and restore the content. Finally, the Board recommended that users be able to appeal decisions made by automated systems to human review when their content is found to have violated Facebook's Community Standard on Adult Nudity and Sexual Activity.



Nazi Quote

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
January 28, 2021	United States	Dangerous Individuals and Organizations

In October 2020, a user in the United States posted a quote, in English, misattributed to Joseph Goebbels, Reich Minister of Propaganda in Nazi Germany. The quote asserted that there is no point in appealing to intellectuals, as they will not be converted and in any case, yield to the stronger man in the street. The quote further stated that arguments should appeal to emotions and instincts. Meta removed the post for violating Facebook's Community Standard on Dangerous Individuals and Organizations. The user, for their part, maintained to the Board that their intent was to draw a comparison between the sentiment expressed in the quote and the presidency of U.S. President Donald Trump.

Reviewing the case, the Board found that since the quote did not support the Nazi party's ideology or the regime's acts of hate and violence, it did not violate Facebook's Community Standard. We noted that, in this case, Meta's approach requiring content moderators to review content without considering the full context led to the user's expression being restricted unnecessarily. The Board further recommended that Meta should explain and provide examples of the applications of key terms in the Dangerous Individuals and Organizations policy, including its definitions of "praise," "support" and "representation," which should align with the definitions in Facebook's Internal Implementation Standards.

Claimed COVID Cure

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
January 28, 2021	France	Violence and Incitement

Another decision considered when health-related misinformation rises to the threshold of contributing to a risk of imminent physical harm. In October 2020, a user posted a statement in French about a decision by a French public health agency, Agence Nationale de Sécurité du Médicament, not to authorize hydroxychloroquine combined with azithromycin for the treatment of COVID-19. The user referred to this combination as a "harmless drug" and claimed it was a "cure." Meta removed the post under its policy on Violence and Incitement, which included a particular rule on misinformation and imminent harm. The Board reversed Meta's decision on the basis that Meta had not demonstrated that this content contributed to imminent harm - the standard in the company's own policies. The fact that the post was addressed to public policy rather than recommending treatments to individuals, and that the drugs in question were not available in France without a prescription, were relevant considerations. We recommended that Meta adopt less intrusive measures to address posts that criticize the positions of health authorities and pose potential but not imminent harms, including links to public health agency websites.

Myanmar Post About Muslims

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
January 28, 2021	Myanmar, France, China	Hate Speech

One decision concerned a post by a user in Myanmar, who shared two widely shared photos of a Syrian toddler of Kurdish ethnicity who had drowned attempting to reach Europe in September 2015. The post questioned the lack of response by Muslims generally to the treatment of Uyghur Muslims in China, compared to killings in response to cartoon depictions of the Prophet Muhammad in France. The accompanying text, in Burmese, Meta translated as: "[I]t's indeed something's wrong with Muslims psychologically," while the Board's translators suggested, "[t]hose male Muslims have something wrong in their mindset." As the Board noted in overturning Meta's decision to remove it, when viewed in context the post was "commentary pointing to the apparent inconsistency between Muslims' reactions to events in France and in China."

While we acknowledged that online hate speech in Myanmar has been linked to accusations of potential crimes against humanity and genocide targeting the Rohingya Muslim minority, the Board found the post "did not advocate hatred or intentionally incite any form of imminent harm." Meta's decision that the post violated Facebook's Hate Speech Community Standard, in short, was wrong.

Armenians in Azerbaijan

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
January 28, 2021	Armenia, Azerbaijan	Hate Speech

Of those first five decisions, the only one that upheld Meta's decision concerned the November 2020 posting of historical photos described as churches in the Azerbaijani capital of Baku. Accompanying text in Russian claimed that Armenians had built Baku, and that this heritage, including these churches, had been destroyed. The user included hashtags in the post calling for an end to Azerbaijani aggression and vandalism. The post used the term, "taziks" to describe Azerbaijanis, whom they claimed have "no history compared to Armenians."

The term "taziks," means "wash bowl," and appeared to have been used as wordplay on the Russian word "aziks," a derogatory term for Azerbaijanis which is on Meta's internal list of slur terms. In this decision, we noted that "the context in which the term was used makes clear it was meant to dehumanize its target." As such, we found that the removal of this post in this context was consistent with Meta's human rights responsibilities. We further found that, if left up, "an accumulation of such content may create an environment in which acts of discrimination and violence are more likely." We recommended that Meta promptly inform the user of the reason for the violation – in this case, a single term – so that they can repost the remainder of the message if they wish.

When announcing our first round of decisions, the Board observed:

> *We believe the first case decisions by the Oversight Board demonstrate our commitment to holding [Meta] to account, by standing up for the interests of users and communities around the world, and by beginning to reshape Facebook's approach to content moderation. This is the start of a process that will take time, and we look forward to sharing our progress through the Board's many subsequent case decisions.*

DECISION SPOTLIGHT

Decision on Former President Trump's Suspension

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
May 5, 2021	United States	Dangerous Individuals and Organizations

At 4:21 PM Eastern Standard Time on January 6, 2021, hours after a mob forcibly entered the United States Capitol Building in Washington, D.C. and was rampaging through the hallways, President Donald J. Trump posted a one-minute video on Facebook, which was also shared to his Instagram account, in which he said:

> *I know your pain. I know you're hurt. We had an election that was stolen from us. It was a landslide election, and everyone knows it, especially the other side, but you have to go home now. We have to have peace. We have to have law and order. We have to respect our great people in law and order. We don't want anybody hurt. It's a very tough period of time.*

He went on to say:

> *There's never been a time like this where such a thing happened, where they could take it away from all of us, from me, from you, from our country. This was a fraudulent election, but we can't play into the hands of these people. We have to have peace. So go home. We love you. You're very special. You've seen what happens. You see the way others are treated that are so bad and so evil. I know how you feel. But go home and go home in peace.*

One hour and 20 minutes later, Meta took down the video for violating its Community Standard on Dangerous Individuals and Organizations. At 6:07 PM, as police were securing the Capitol, Mr. Trump posted a statement on Facebook:

Credit: Tyler Merbler

These are the things and events that happen when a sacred landslide election victory is so unceremoniously and viciously stripped away from great patriots who have been badly and unfairly treated for so long. Go home with love in peace. Remember this day forever!

Eight minutes later, Meta removed that post from Facebook for violating its Community Standard on Dangerous Individuals and Organizations and imposed a 24-hour block on Mr. Trump's ability to post on Facebook or Instagram.

The following day, January 7, after reviewing Mr. Trump's posts, his recent communications off Facebook, and additional information about the severity of the violence in Washington, Meta extended its block on Mr. Trump's accounts "indefinitely and for at least the next two weeks until the peaceful transition of power is complete."

Meta's referral to the Board

Two weeks later, on January 21, the day after President Joe Biden's inauguration, Meta referred its decision to indefinitely suspend Mr. Trump's access to his Facebook and Instagram accounts to the Board. It also asked the Board for observations or recommendations about account suspensions when the user is a political leader.

Meta's indefinite suspension of Mr. Trump's Instagram account and Facebook page had, not surprisingly, drawn intense public, political and press attention, and prompted heated debate as to the fairness, complexity, and limitations of policies governing content moderation on social media platforms. It also drew renewed attention, the Board noted, "to the value of independent oversight of the most consequential decisions by companies such as Facebook." Our pivotal role in attempting to resolve this tension prompted a comment from The Washington Post:

Officials and policymakers all over the world, seeking to design new frameworks for regulating the social media industry are watching the board closely. If it's judged a success, it may lessen calls for regulation. If it fails, the failure may hasten demands to create more-stringent legal guardrails for content moderation in many countries."[5]

The New York Times succinctly summed the stakes up.[6]

This decision has major consequences not just for American politics, but also for the way in which social media is regulated, and for the possible emergence of a new kind of transnational corporate power at a moment when almost no power seems legitimate.

In reaching a decision, the Board considered whether Meta's suspension of Mr. Trump's access to posting content on Facebook and Instagram was in line with Meta's own policies, its values, and international human rights standards.

The Board considered multiple sources of information: a decision rationale provided by Meta that laid out the reasoning behind its decision; Meta's responses to specific questions asked by the Board; a Content Creator Statement submitted to the Board on Mr. Trump's behalf through the American Center for Law and Justice and a page administrator; and independent research.

Meta's explanation

In cases where Meta must make an emergency decision that has widespread interest, it often shares the reasoning behind its decision through a post in its Newsroom. With regard to this case,

5 "Facebook's New 'Supreme Court' Overturns Firm in First Rulings," Elizabeth Dwoskin, Craig Timberg, *The Washington Post,* January 28, 2021

6 "Trump Wants Back on Facebook. This Star-Studded Jury Might Let Him." Ben Smith, *The New York Times,* Jan. 24, 2021

Meta stated that it removed the two pieces of content posted on January 6, 2021, for violating the Community Standard on Dangerous Individuals and Organizations. Specifically, the content was removed for violating its policy prohibiting "praise, support and representation of designated violent events." The content, Meta maintained, also violated the part of that policy prohibiting "praise of individuals who have engaged in acts of organized violence."

Meta noted that, while Mr. Trump did ask people in his video to "go home in peace," he also reiterated allegations that the election was fraudulent and suggested a common purpose in saying, "I know how you feel." Given the ongoing instability at the time of his comments and the overall tenor of his words, Meta concluded that "We love you. You're very special" was intended as praise of people breaking the law by storming the Capitol. The company also believed that the second post contained praise of the event, as Mr. Trump referred to those who stormed the Capitol as "great patriots" and urged people to "[r]emember this day forever."

As for its subsequent decision to extend its 24-hour block on Mr. Trump's ability to post content for an indefinite period, Meta stated that after further assessing the evolving situation and emerging details of the violence at the Capitol, it concluded that the 24-hour ban was not sufficient to address "the risk that Trump would use his Facebook and Instagram presence to contribute to a risk of further violence." The company cited a National Terrorism Advisory System Bulletin issued by the Department of Homeland Security (DHS), which described a "heightened threat environment across the United States, which DHS believes will persist in the weeks following the successful Presidential Inauguration."

After Mr. Biden's inauguration on January 20, Meta left the block in place "indefinitely," stating neither a determinate time period nor specific criteria for restoration, leaving the duration of the block to its future discretion.

> The Trump judgment cannot possibly satisfy everyone. But this 38-page text is a serious contribution to thinking about how to handle free speech in an age of information chaos."
>
> Alan Rusbridger
> *Board Member*

Trump's indefinite suspension, the company claimed, was informed by international human rights standards, specifically Article 19 of the International Covenant on Civil and Political Rights (ICCPR) and the UN Human Rights Committee's General Comment No. 34, which permits necessary and proportionate restrictions of freedom of expression in situations of public emergency that threaten the life of the nation.

While Meta has a "newsworthiness allowance" which allows content that violates its policies to remain on the platform if the company considers the content "newsworthy and in the public interest," Meta explained that it did not apply this allowance to the posts at issue in this case.

The Board upholds the initial decision to suspend the account but reverses the decision to leave the duration indefinite.

On May 5, we published our decision and issued a set of recommendations. The Board found that, given the seriousness of the violations and the ongoing risk of violence, Meta was justified in suspending Mr. Trump's accounts on January 6, 2021, and extending that suspension on January 7. The Board found that the two posts in question violated Facebook's Community Standards and Instagram's Community Guidelines. Facebook's rules prohibit the posting of content that expresses support or praise for groups, leaders, or individuals

involved in activities such as terrorism, organized violence or criminal activity.

In its decision, the Board noted that:

> *At the time that the posts were made, the violence at the Capitol was underway. Both posts praised or supported people engaged in violence. The words "We love you. You're very special" in the first post and "great patriots" and "remember this day forever" in the second post amounted to praise or support of the individuals involved in the violence and the events at the Capitol that day.*

However, the decision also found that it was not appropriate for Meta to impose the "indeterminate and standardless penalty of indefinite suspension." It stated that a penalty of indefinite duration "finds no support" in the Community Standards. Furthermore, the vagueness of the penalty imposed "violates principles of freedom of expression… It is unclear," the decision further stated, "what standards would trigger this penalty or… will be employed to maintain or remove it."

We rejected Meta's request for the Board to "endorse indefinite restrictions, imposed and lifted without clear criteria." We insisted that "within six months of this decision, [Meta] must reexamine the arbitrary penalty that it imposed on January 7 and decide the appropriate penalty. This penalty must be based on the gravity of the violation and the prospect of future harm. It must also be consistent with [its platforms'] rules for severe violations, which must, in turn, be clear, necessary, and proportionate."

> What we are telling [Meta] is that they can't invent penalties as they go along. They have to stick to their own rules."
>
> Helle Thorning-Schmidt
> *Board Co-Chair*

Meta responds to the Board's decision

On June 4, Meta responded to the Board's decision, including our call to review its indefinite penalty and decide upon a proportionate response consistent with the rules applied to other Facebook users. The company announced a new set of enforcement protocols "to be applied in exceptional cases such as this." In accordance with those protocols, Meta announced a new suspension of Mr. Trump's account for two years, effective from January 7, 2021. "Given the gravity of the circumstances that led to Mr. Trump's suspension," Meta maintained, "we believe his actions constituted a severe violation of our rules, which merit the highest penalty available under the new enforcement protocols." As the due date of the penalty's expiration approached, Meta promised to:

> *look to experts to assess whether the risk to public safety has receded. We will evaluate external factors, including instances of violence, restrictions on peaceful assembly and other markers of civil unrest. If we determine that there is still a serious risk to public safety, we will extend the restriction for a set period of time and continue to re-evaluate until that risk has receded. When the suspension is eventually lifted, there will be a strict set of rapidly escalating sanctions that will be triggered if Mr. Trump commits further violations in future, up to and including permanent removal of his pages and accounts.*

DECISION SPOTLIGHT

Ensuring Respect for Freedom of Expression

Of the 20 decisions the Board made in 2021, a significant number required Meta to restore content we found to have been wrongly removed. Meta's actions in these cases were found to be inconsistent with the company's rules, values, and international human rights commitments. In the following five representative cases, we found that Meta's decision to restrict users' expression was not justified under the circumstances.

Punjabi Concern Over the RSS in India

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
April 29, 2021	India	Dangerous Individuals and Organizations

In November 2020, a user shared a video post from the Punjabi-language online media company Global Punjab TV, featuring a 17-minute interview with a professor described as a social activist and supporter of the Punjabi culture. In an accompanying text, the user asserted that the Hindu nationalist organization Rashtriya Swayamsevak Sangh (RSS) and India's ruling party Bharatiya Janata Party (BJP) were threatening to kill Sikhs, a minority religious group in India. The post further alleged that Prime Minister Modi, a former RSS leader, was threatening "Genocide of the Sikhs" on the advice of the RSS President.

After being reported by one user, a human moderator removed it on the grounds that it violated Facebook's Dangerous Individuals and Organizations Community Standard. This triggered an automatic restriction on the user's account. The user then appealed to the company. Meta told the user it could not review this appeal, citing a temporary reduction in capacity caused by COVID-19. As a result of the Board selecting the case, Meta belatedly restored the content, conceding that its initial decision was wrong.

As the Board noted in its decision, nothing in the video violated Meta's policies. The post highlighted the concerns of minority and opposition voices in India that are allegedly being discriminated against by the government. In response to Meta's claim that it lacked the capacity to review the user's appeal due to COVID-19, we repeated our call from our *Breast cancer symptoms and nudity* decision for all cases to be appealed to Meta before they come to the Board.

Report on the timeliness of Meta's implementation of and response to our decisions

- Under our Bylaws, Meta must implement our decisions **within seven days** of publication.
- For the 20 decisions we published in 2021, Meta restored or removed the content within this 7-day timeframe, except in cases where the content had already been restored.
- For nearly all our decisions requiring action, Meta restored or removed the post on the same day we published our decision.
- Through our Implementation Committee, currently made up of five Board Members, we continue to urge Meta to provide greater transparency about its processes for identifying and taking enforcement action on pieces of content that are both identical to those featured in our decisions and presented in a parallel context. This would ensure our decisions are addressed outside of the specific case, and generalized to relevant content across similar contexts.

"Two buttons" Meme

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
May 20, 2021	United States, Armenia, Turkey	Hate Speech; Cruel and Insensitive

On December 24, 2020, a Facebook user in the U.S. posted a comment that adapted the 'two buttons' meme from the comic "Daily Struggle" to a Turkish political context. The meme (published in 2014) features a cartoon character attempting to push one of two buttons labeled with potentially contradictory statements. The post substituted a Turkish flag for the cartoon character's face, above which appeared two red buttons with statements in English: "The Armenian Genocide is a lie" and "The Armenians were terrorists that deserved it." After a human content moderator found that the meme violated Facebook's Hate Speech Community Standard, and another determined that it violated its Cruel and Insensitive Community Standard, Facebook removed it and informed the user. They appealed to Meta and then to the Board.

Once again, context was key. Meta had based its decision to take down the post on the statement "Armenians were terrorists that deserved it." However, we found that, when understood in context, the post actually criticized the Turkish government's denialist position on the Armenian Genocide. As such, the post was covered by an exception to the Hate Speech Community Standard which permits sharing hate speech "to condemn it or raise awareness." We also concluded that the content was covered by Facebook's satire exception — which we called on the company to include in its public-facing Community Standards.

Öcalan's Isolation

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
July 8, 2021	Turkey, United States	Dangerous Individuals and Organizations

The Board's July 2021 decision highlighted the importance of independent review of Meta's actions by uncovering a piece of policy guidance that had had gone missing for three years.

On January 25, 2021, an Instagram user in the U.S. posted a picture of Abdullah Öcalan, a founding member of the Kurdistan Workers' Party (PKK).

A caption in English encouraged readers to talk about ending Öcalan's isolation in prison on Imrali island in Turkey, and the inhumanity of solitary confinement. The PKK and Öcalan are designated as "dangerous entities" under Facebook's Dangerous Individuals and Organizations policy, which also prohibits "praise" or "support" of designated entities.

After a moderator removed the post for violating the Dangerous Individuals and Organizations policy, the user appealed that decision. After Meta rejected that appeal, the user appealed to the Oversight Board.

An important aspect of this decision that only emerged after the Board selected the case underscored the importance of our independent review of Meta's actions. The company disclosed that it had located a piece of internal guidance, developed in 2017, which expressly permitted discussion of the conditions of confinement for designated individuals. That policy had been crafted, in part, to respond to concerns about the conditions of Öcalan's imprisonment and to permit people to talk about that on the platform. In overturning Meta's decision to remove the post, we conveyed our concern about Meta's lack of transparency: "[Its] policy of defaulting towards removing content showing 'support' for designated individuals, while keeping key exceptions hidden from the public, allowed this mistake to go unnoticed for an extended period."

Which Community Standards did our decisions examine most?

Hate Speech
9 decisions

Dangerous Individuals and Organizations
5 decisions

Violence and Incitement
2 decisions

Shared Al Jazeera Post

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
September 14, 2021	Israel, Egypt	Dangerous Individuals and Organizations

On May 10, 2021, a Facebook user in Egypt shared a post by the verified Al Jazeera Arabic Page consisting of text in Arabic and a photo. The photo portrayed two men in camouflage fatigues with faces covered, wearing headbands with the insignia of the Al-Qassam Brigades.

The text stated: "The resistance leadership in the common room gives the occupation a respite until 18:00 to withdraw its soldiers from Al-Aqsa Mosque and Sheikh Jarrah neighborhood, otherwise he who warns is excused. Abu Ubaida – Al-Qassam Brigades military spokesman." The user shared Al Jazeera's post and added a single-word caption: "Ooh" in Arabic. The Al-Qassam Brigades and their spokesperson Abu Ubaida are both designated as dangerous under Facebook's Dangerous Individuals and Organizations Community Standard. Meta removed the post under this policy but restored it as a result of the Board selecting the case.

Meta was unable to explain why two human reviewers originally judged the content to violate its policy, noting that moderators are not required to record their reasoning for individual content decisions. The Board's decision highlighted that while Facebook's Dangerous Individuals and Organizations policy prohibits "channeling information or resources, including official communications, on behalf of a designated entity," it also contains an exception for content published as "news reporting." The content in this case was a reprint of a widely-republished news report in Al Jazeera about a May 2021 armed conflict between Israeli forces and Palestinian militant groups in Israel and Gaza, a Palestinian territory governed by Hamas, with no alteration other than the addition of the non-substantive comment "ooh." We therefore concluded that the post fell under Facebook's news reporting exception.

The case was also about treating users fairly, because this post, consisting of a republication of a news item from a legitimate outlet, was treated differently from content posted by the news organization itself. As we noted in our decision, individuals have as much right to repost news stories as media organizations have to publish them in the first place. We also noted that while the Al Jazeera page benefited from the cross-check system Meta applies to certain high-profile accounts, the user sharing that content did not.

Colombia Protests

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
September 27, 2021	Colombia	Hate Speech

In June 2021, the Facebook page belonging to a regional news outlet in Colombia shared a video that showed protesters marching behind a banner emblazoned with the slogan "SOS COLOMBIA." The video's audio contained chanting and singing in Spanish, with the protesters denouncing the Colombian President, Ivan Duque, using homophobic slurs that prompted Meta to remove the video (which had originated on Tik-Tok) for violating its Hate Speech Community Standard. This prohibits content that "describes or negatively targets people with slurs…used as insulting labels." The policy specifically bans "insulting labels" that refer negatively to protected characteristics, including sexual orientation.

In appealing Meta's decision to remove the video, the user stated that they were a local journalist reporting on news from their province. The user further maintained that the outlet's sharing of the video was not intended to inflict harm, but simply to show a group of young people protesting peacefully and demanding rights using typical language.

In deciding that Meta should restore the video, the Board acknowledged that while the content on its face violated Facebook's Hate Speech Community Standard, as sharing a slur can contribute to an environment of intimidation and exclusion for LGBT people, Meta had erred in its removal by failing to apply its "newsworthiness allowance" to the post. This provision may allow standard-violating content to stay up depending on its level of public interest and risk of harm.

DECISION SPOTLIGHT

Protecting Users from Harmful Content

In certain key decisions handed down in 2021, the Board upheld Meta's decision to take content down, finding that limiting user expression can be justified under certain circumstances. These decisions also made clear that Meta's removal of content can be consistent with the company's human rights responsibilities, especially when the removal of one user's expression was necessary and proportionate to respect other peoples' rights.

Depiction of Zwarte Piet

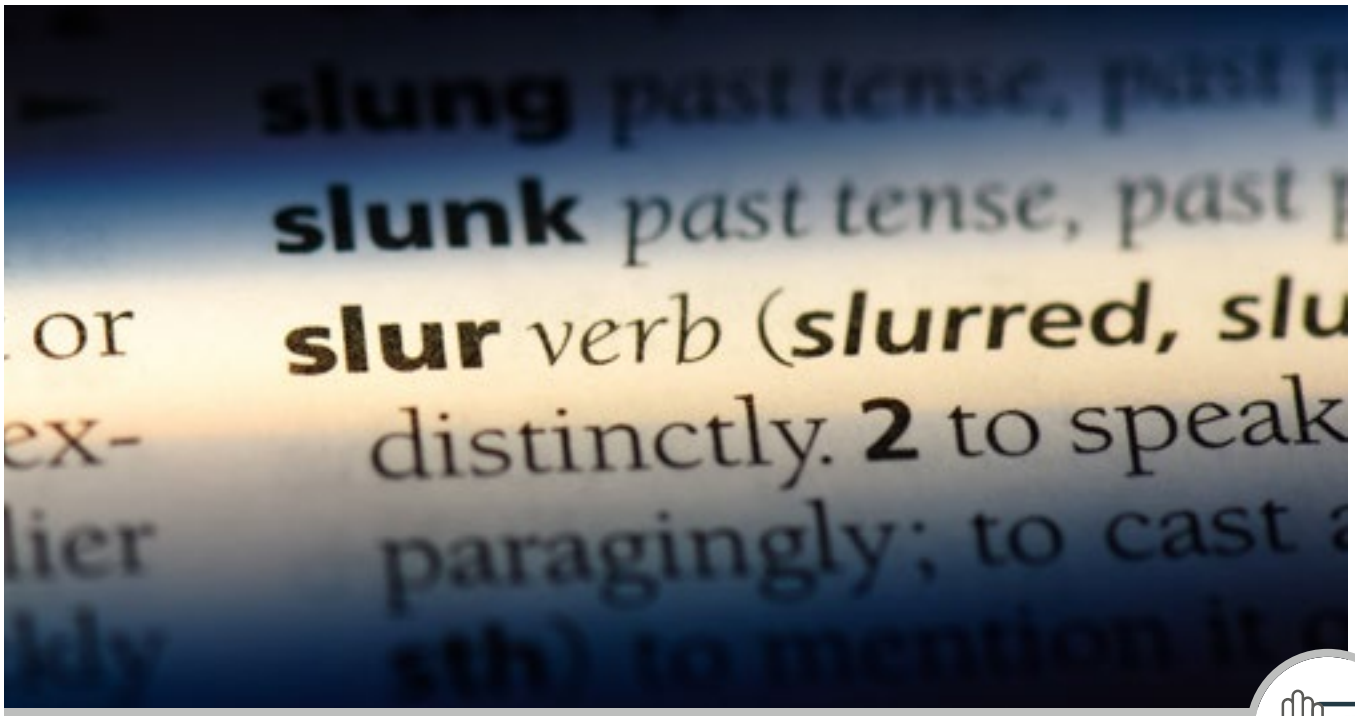
DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
April 13, 2021	The Netherlands	Hate Speech

On December 5, 2020, a Facebook user in the Netherlands shared a post on their timeline, which included text in Dutch and a 17-second-long video depicting a young child meeting three adults. One was dressed up as "Sinterklaas," or Santa Claus, while the two others appeared in blackface, wearing Afro wigs under hats and colorful renaissance-style clothes, portraying a Dutch caricature known as "Zwarte Piet." In English, the name translates as "Black Pete."

We upheld Meta's decision to take down the post for violating its Hate Speech Community Standard. A majority of the Board found that the content consisted of "caricatures that are inextricably linked to negative and racist stereotypes and are considered by parts of Dutch society to sustain systemic racism in the Netherlands." In reaching this decision, the majority of the Board considered studies that showed "Black children felt scared and unsafe in their homes and were afraid to go to school during the Sinterklaas festival."

While a minority of the Board saw insufficient evidence to directly link this piece of content to the harm supposedly being reduced by removing it, the majority of the Board found that allowing content like this to accumulate on the platform would help create a discriminatory environment for Black people that would be degrading and harassing. The majority noted that:

> *At scale, [Meta's] policy is clear and ensures Black people's dignity, safety and voice on the platform. Restricting the voice of people who share depictions of blackface in contexts where it is not condemning racism is acceptable to achieve this objective.*



South Africa Slurs

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
September 28, 2021	South Africa	Hate Speech

In May 2021, a Facebook user posted a statement in English to a public group that described itself as focused on "unlocking minds." The post discussed "multi-racialism" in South Africa, and argued that poverty, homelessness, and landlessness have increased for Black people in the country since 1994. The user stated that white people hold and control the majority of the wealth, and that while wealthy Black people may have ownership of some companies, they do not have control. The post challenged people with contrary views to "have their heads examined," describing a group of Black people using racial slurs that Meta explained violated its Hate Speech policy.

We upheld Meta's decision to remove this content, finding that one of the terms used is a "particularly hateful and harmful word in the South African context." We emphasized that open discussion about socio-economic inequality is key to advancing equality. Nevertheless, we considered it was possible for the user to engage in such discussion and appeal to the emotions of their audience without referencing this slur. Moreover, one slur in particular has a specific connection to the country's history of apartheid, and its use is still associated with the degradation and exclusion of Black people. The decision reiterated the importance of Meta providing people who have their content removed under the Hate Speech policy with a specific explanation of the rule violated.

Alleged Crimes in Raya Kobo

DATE PUBLISHED	COUNTRY	COMMUNITY STANDARD
December 14, 2021	Ethiopia	Hate Speech

A December 2021 decision found that a post including unverifiable rumors targeting an ethnic group violated Meta's Community Standards and that its removal was consistent with human rights principles.

In late July 2021, a Facebook user posted on his timeline in the Ethiopian language Amharic, allegations that the Tigray People's Liberation Front (TPLF) had killed and raped women and children and looted properties in the Ethiopian district of Raya Kobo and nearby towns in the nation's Amhara region. In the post, the user claimed that ethnic Tigrayan civilians had assisted the TPLF in committing these atrocities. The post concluded with the following words: "We will ensure our freedom through our struggle." Moreover, the content, which had been viewed by thousands of Amharic speakers in the 24 hours that it remained online, contained comments from some of those viewers that included calls for vengeance.

A content moderator from Meta's Amharic content review team determined that the post violated Facebook's Hate Speech Community Standard and removed it. The policy prohibits "violent speech" targeting a person based on race, ethnicity, or national origin. The user appealed the decision to Meta, which, following a second review by another moderator from the Amharic content review team, confirmed that the post had violated Facebook's policies.

As a result of the Board's selection of the case for review, Meta identified its removal as an "enforcement error" and restored the content. Meta based this decision to restore the post on its assessment that the statements in the post did not rise to the level of hate speech.

We disagreed, finding Meta's explanation "lacking in detail and incorrect." We applied the Violence and Incitement Community Standard to this post, finding it in violation of Facebook's prohibition on "misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm." The Board found that "rumors alleging the complicity of an ethnic group in mass atrocities are dangerous and significantly increase the risk of imminent violence during an ongoing violent conflict" such as the one presently occurring in Ethiopia.

International Human Rights Norms in the Board's Decision-Making Process

Part of our responsibility is to ensure that Meta's approach to content moderation is grounded in respect for human rights. This is reflected in our Charter, and in Meta's own corporate human rights commitments. Our decisions therefore look beyond whether Meta's content decisions followed the company's content rules and values, but also examine whether decisions by Meta comport with international human rights standards.

International human rights standards provide a universal framework for protecting free expression and other human rights. They also provide a globally consistent framework for holding Meta to account for its impacts throughout the world. Meta's Corporate Human Rights Policy sets out its commitments to respect human rights - including through how it develops its content policies. The policy also notes that the UN Guiding Principles on Business and Human Rights (UNGPs) informed the Board's creation.

> "It is about time that we have a conversation about how we create technology that is by design respectful to human rights."
>
> Julie Owono
> *Board Member*

We believe that placing human rights at the heart of our work, and pushing Meta to respect them, will make Facebook and Instagram better places for users. Over time, we hope that our decisions will bring Meta's approach to content moderation into greater alignment with international human rights standards.

A new challenge

Applying international human rights standards to content moderation by a corporation like Meta is a relatively new challenge. No other entity is engaged in the type of exercise the Board is currently undertaking.

Using a legal framework largely created by states for states, to interpret the human rights responsibilities of a social media company, raises many questions. To what extent should Meta's moderation of content be subject to the same strictures applied to governments regulating speech in other places? How should this approach consider the challenges

of moderating content at scale? How should we explore harms that may result from Meta's design choices, including algorithmic amplification of content? Our decisions have explored how international human rights standards can be used to define Meta's human rights responsibilities, which may differ from those applicable to states. In our *Depiction of Zwarte Piet* and *Armenians in Azerbaijan* decisions, for example, we upheld Meta's decision to remove the content, while recognizing that international human rights standards would likely not permit a state to punish the same speech through criminal or administrative measures.

Building an organization rooted in human rights

As a body whose decisions affect people around the world, it is vital that we respect human rights in all aspects of our work. The Board's Charter requires that the Board pay particular attention to human rights when making decisions. We seek to follow the UNGPs by making our work legitimate, accessible, equitable, transparent, rights-compatible, a continuous source of learning and based on engagement and dialogue. In 2020, the Board commissioned a follow-up report from non-profit organization Business for Social Responsibility (BSR) which stated that the Board had made "good progress" on recommendations it had made in a prior report.

How human rights underpin our approach to freedom of expression

Among the sources of authority guiding the Board's decisions are international human rights standards, with Article 19 of the ICCPR among the most cited. The ICCPR is a global human rights treaty that Meta cites in its Corporate Human Rights Policy, voluntarily pledging to respect the human rights it describes. The work of the UN Special Rapporteur

Organizations React to the Board's Work

Academics, civil society, and national and international institutions have started to analyze the Board's work and decisions. Highlights include:

- In December 2020, the UN Special Rapporteur on minority issues, Fernand de Varennes, urged us to consider minority rights, calling the Board "an innovative and ambitious initiative to regulate online expression, in particular hate speech, which is essential to the effective protection of vulnerable minorities worldwide."
- In March 2021, a legal research analyst at the Law Library of Congress examined how the Board's first five decisions applied international human rights law to social media companies.
- In September 2021, organizations, including Access Now, Article 19 and 7amleh welcomed the Board's Shared Al Jazeera post decision as a "step in the right direction" saying the case was "emblematic of Facebook's systematic arbitrary and non-transparent overenforcement" of the Dangerous Individuals and Organizations Community Standard.
- In December 2021, the United States Commission on International Religious Freedom mentioned our work applying human rights norms to content moderation.
- 'Lawfare' also set up a comprehensive Oversight Board blog compiling commentary and analysis on our decisions so far.
- Our decisions are also carried by the international legal databases Westlaw, and Lexis/Lexis+.

on freedom of opinion and expression, in particular a 2018 report on content moderation which explored how private companies can best respect human rights, has proven especially valuable to the Board.

Article 19, paragraph 2, of the ICCPR states that "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." However, in line with international human rights standards, the Board does not consider the right to freedom of expression to be absolute. While the right is important, it can be restricted in certain circumstances. ICCPR Article 19, paragraph 3 provides a three-part test for assessing a restriction on freedom of expression, considering legality, legitimate aim and necessity and proportionality. These principles are similar, though not identical, to freedom of expression standards found in the domestic and constitutional law of many nations where Meta operates, including the First Amendment of the United States Constitution. We also consider restrictions on expression required by international human rights law when relevant, including the prohibition on advocacy of hatred that constitutes incitement to hostility, discrimination, or violence (Article 20, paragraph 2, ICCPR).

In our *legality* analysis in decisions, we examine how clear Meta's rules are to the people who use Facebook and Instagram. Rules should be easy for users to find and understand. In numerous decisions, we have pushed Meta to publish parts of its rules that are hidden from public view, including the lists of individuals and organizations which the company designates as "dangerous."

In our *legitimate aim* analysis in decisions, we look at whether Meta's rules on Facebook and Instagram pursue aims that are considered legitimate reasons for restricting expression under international human rights law. Most of our decisions identify respect for the "rights of others" as a legitimate purpose behind Meta's content policies that restrict speech.

In our *necessity and proportionality* analysis, we examine whether Meta's action in a particular case was necessary to achieve the legitimate aim, and whether the measures imposed, such as content removal, were the least restrictive means to achieve that aim.

In our first 20 decisions, we have raised concerns that some of Meta's content rules are too vague, too broad, or unclear, prompting recommendations to clarify rules or make secretive internal guidance on interpretation of those rules public. The Board has attached particular importance to political content and content that raises awareness of human rights violations, including political satire, as well as artistic expression and discussion of health issues. We also have questioned whether content removal is always a proportionate response to content that may be linked to harm but does not directly cause imminent harm. For example, in two decisions related to COVID-19 (*Claimed COVID cure* and *COVID lockdowns in Brazil*) we asked whether other measures short of removing the posts (such as warning screens, labelling, or downranking content) could mitigate risk while also protecting expression.

Beyond the three-part test above, we have also integrated principles on equality and discrimination into our decision-making, recognizing that different people in different situations may be differently impacted by Meta's approach to content moderation. This informed recommendations that Meta assess its human rights impact in areas affected by conflict, in languages such as Arabic and Amharic, as well as calls for improvements in its moderation of Burmese. The Board's attention to equality and non-discrimination issues in the *breast cancer symptoms and nudity* decision also showed how Meta's reliance on automation, which led to the wrongful removal of health information specific to women, can have disproportionate and discriminatory impacts on users. For similar reasons, we are attentive to the importance of non-discrimination on the basis of political point of view.

Rights versus rules – pro-Navalny protests in Russia

In most of our decisions, our human rights analysis has supported the same outcome as our interpretation of Meta's rules. However, in a milestone for the Board, the *pro-Navalny protests in Russia* decision was the first case where we overturned Meta's decision based on its human rights responsibilities, even though the removal was in line with Meta's rules.

On January 24, 2021, a user in Russia posted pictures, a video, and text describing the protests that had broken out the day before in support of opposition leader Alexei Navalny in Saint Petersburg, Moscow and other communities across the country. Another user, the 'Protest Critic,' responded that while they did not know what had happened in Saint Petersburg, the protesters in Moscow were all "school children," were "mentally "slow," and had been "shamelessly used." Yet another user, who claimed to have participated in the protest in Saint Petersburg, concluded by calling the Protest Critic a "cowardly bot." "The Protest Critic" reported the Protester's comment to Meta for bullying and harassment.

Meta took down the post for violating its Bullying and Harassment standard, under which a private individual can get Meta to take down a Facebook post containing a "negative comment on their character." Deciding in favor of the Protester's request to restore the post, the Board observed:

> *This case highlights the tension between policies protecting people against bullying and harassment and the need to protect freedom of expression. This is especially relevant in the context of political protest in a country where there are credible complaints about the absence of effective mechanisms to protect human rights.*

Our decision stated that even though the content's removal was in line with Meta's Bullying and Harassment Community Standard, it was an unnecessary and disproportionate restriction on free expression under international human rights standards.

The Board further found, in favor of the user's right to expression, that while Meta's original decision to remove the content "may have been consistent with a strict application of the Community Standards," an overly strict application of those standards had caused it "to fail to consider the wider context and disproportionately restricted freedom of expression."

This case showed that where conflict emerges between Meta's content policies and its human rights responsibilities, **the Board is prepared to prioritize human rights.**

Case studies on specific standards

The Board's first 20 decisions considered a wide range of international human rights standards. Every decision published in 2021 referred to Article 19 of the ICCPR, the UN Guiding Principles on Business and Human Rights and the UN Human Rights Committee's General Comment 34 on the freedoms of opinion and expression. The application of these and other international human rights standards to our decisions has deepened our understanding and analysis of the underlying issues driving many content moderation decisions. It has also helped us to arrive at more principled conclusions. While each decision references several human rights standards, we provide a few examples below of standards that made a crucial contribution to our decisions.

UN Guiding Principles on Business and Human Rights

In a decision about Meta's removal of a post that showed a wampum belt, a North American Indigenous art form, with the title "Kill the Indian / Save the Man," we urged Meta to fulfil its human rights responsibilities to marginalized communities under Principle 17 of the UNGPs. The cultural context surrounding the quote was provided in a public comment from the Association of American Indian Affairs, which identified the author of the

quote as Richard Henry Pratt, the founder of the first federal Indian boarding school in the United States. Taken in context, the quote constituted a clear example of "counter speech," a form of expression that deliberately references hate speech to resist, not support or promote, oppression and discrimination.

In response to this case, we urged Meta to fulfil its human rights responsibilities to marginalized communities under Principle 17 of the UNGPs. This requires Meta to undertake human rights due diligence to ensure that its systems are operating fairly and not exacerbating historical and ongoing oppression. This includes identifying, preventing, and mitigating adverse impacts of content moderation on the expression of Indigenous peoples countering discrimination. Finally, we noted our expectation that Meta prioritize risks to marginalized groups and provide evidence for how the company is improving.

General Comment 34

In the *Myanmar bot* case, which focused on content posted following an undemocratic coup in the country, we clarified to Meta that a user's criticism of a government must be distinguished from discriminatory speech against people based on their nationality. In doing so, the Board drew on the UN Human Rights Committee's General Comment 34, which provides guidance on how states should interpret Article 19 of the ICCPR, which protects freedom of expression at the global level. The *Myanmar bot* decision specifically refers to the parts of the General Comment addressing the importance of political speech, and in particular the value of permitting uninhibited discussion of public institutions. While freedom of expression may be limited in certain circumstances to protect individuals against discriminatory hate speech that leads to particular harms, people should be free to say what they want about nation states and public institutions.

Right to remedy (Article 2, ICCPR)

The *Öcalan's isolation* decision quoted the UN Special Rapporteur on freedom of expression's statement that the remedy process "should include a transparent and accessible process for appealing platform decisions, with companies providing a reasoned response that should also be publicly accessible."

In this case, however, Meta's actions fell short of these standards. Initially, the company informed the user that they could not appeal to the company due to COVID-19. We also expressed concern "at what may be a significant number of removals that should not have happened" because Meta had lost a piece of policy guidance for three years that clearly allowed the user's content. The company claimed that it was not technically feasible to determine how many pieces of content were removed when the policy guidance was not available to reviewers. For these concerns, and others, we criticized Meta for "failing to respect the right to remedy, in contravention of its Corporate Human Rights Policy."

A difficult, but valuable, task

We recognize that using international human rights standards to interpret the human rights responsibilities of a social media company is a new and challenging exercise.

Yet while this is a difficult task, it is a valuable one. Applying international human rights to content moderation can help us ensure Meta respects freedom of expression alongside other human rights, and provide a consistent, global framework for holding social media companies to account.

Building a body of work that shows what rights-based content moderation looks like is central to the Board's mission. In 2022, we will look to the precedential value of our earlier decisions as we continue to ensure respect for the human rights of people who use Meta's platforms.

Questions the Board asked Meta as Part of Our Decisions

To assist with making our decisions and to push Meta to be as transparent as possible, we send questions to the company before deliberating cases. By asking specific questions and including the details in our final decision, we can hold Meta to account and provide researchers with new information about how the company works.

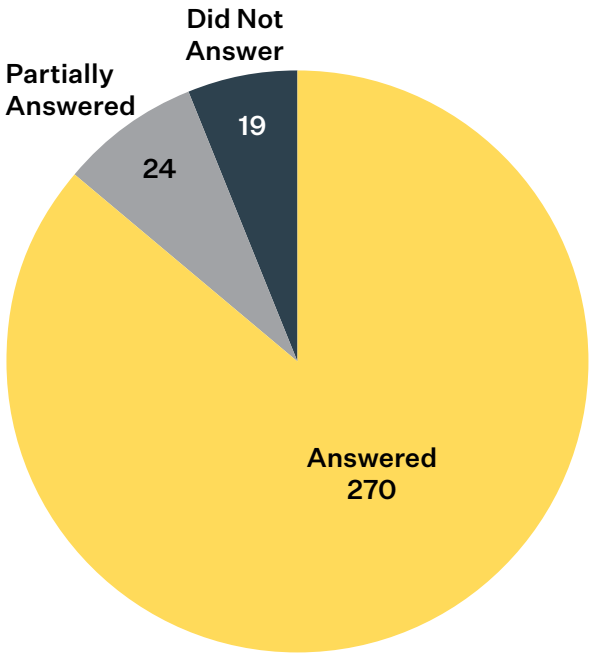
For the 20 decisions we published in 2021, our questions covered many different areas. In many cases our questions explored specific concerns highlighted by users in their statement to the Board, as well as issues raised in public comments. Many of our questions related to how Meta treated the user who appealed the content, such as the reason they were given for why their content was removed and whether the company reviewed their original appeal. For other cases, we requested more detail on Meta's rules, automated systems, and regional teams.

As part of our first 20 decisions, we asked Meta 313 questions. Of these, Meta answered 270, partially answered 24 and did not answer 19. We present these numbers in the interest of transparency and make no claim here as to whether Meta had good reason not to answer in any case.

In our first year, Meta has generally been responsive to the Board's inquiries, answering the vast majority (86%) of our questions. Where Meta did not answer our questions, we raised this both in our quarterly transparency reports and in certain decisions. A detailed description of the questions which Meta did not answer and the reasons it gave are included later in this section.

After receiving some late responses to our questions on early cases, we worked with Meta to agree on a tiered system where the company's deadline depends on how many questions the Board asks.

Meta's response to Oversight Board questions

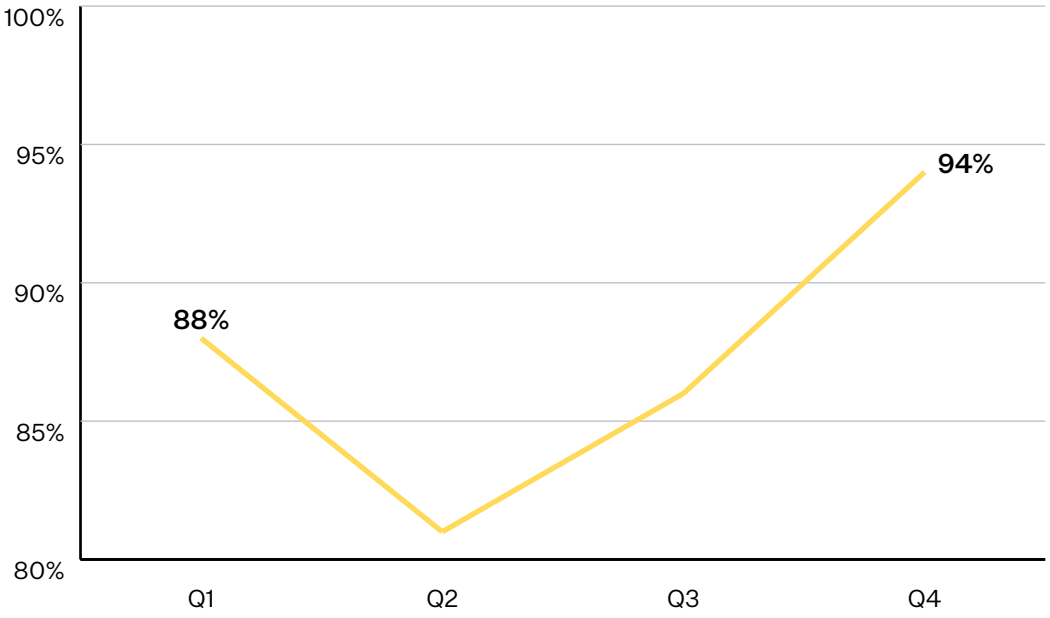


We also appreciate that Meta made itself available to give briefings on issues related to its rules, including the cross-check system.

The share of questions which Meta answered increased for decisions published in Q4 2021

The share of questions Meta answered for decisions published between Q1 and Q3 2021 started at 88% of questions answered for Q1, compared with 81% for Q2 and 86% for Q3. For the three decisions published in Q4 2021, however, this increased to 94%, with Meta answering 51 of our 54 questions, and partially answering three questions. Q4 was the first quarter where none of Meta's responses fell into the "did not answer" category. This progress is encouraging, and we hope this trend continues into 2022.

Percentage of questions Meta answered by quarter



Details on questions which Meta did not answer

Meta did not answer 19 of the 313 questions we asked as part of our decisions published in 2021. Below we provide more detail on these questions, including the case they were related to, what they asked, and the reason Meta gave for not answering.

- Meta did not answer **two questions** in the *Myanmar post about Muslims* case. The first question asked whether Meta had previously enforced hate speech violations against the user and group in question. Meta responded that information about the user's previous behavior on the platform was irrelevant to the Board's determination. The second question requested Facebook Market Team English translations of the comments made on the post. Meta responded that it did not automatically translate the comments as part of its content review process, and that it did not have manual translations of the comments.
- Meta did not answer **one question** in the *Armenians in Azerbaijan* case. The Board asked why one of the user's previous posts had been removed. Meta responded that information about the user's previous behavior on the platform was irrelevant to the Board's determination.
- Meta did not answer **one question** in the *Nazi quote* case. The Board asked whether the user had a record of posts being banned or suspended. Meta responded that information about the user's previous behavior on the platform was irrelevant to the Board's determination.
- Meta did not answer **one question** in the *protest in India against France* case. The Board's question asked whether Meta had previously enforced violations under the Violence and Incitement Community Standard against the user or group. Meta responded that information about the user's previous behavior on the platform was irrelevant to the Board's determination.
- Meta did not answer **two questions** in the *Punjabi concern over the RSS in India* case. The first question asked what specific language in the content caused Meta to remove it under the Dangerous Individuals and Organizations Community Standard. Meta responded that it was unable to identify the specific language that led to the erroneous conclusion that the content violated the Dangerous Individuals

and Organizations policy. The second question included asking how many "strikes" users need for Meta to impose an account restriction, and how many violations of the Dangerous Individuals and Organizations policy are required for account-level restrictions. Meta responded that this information was not reasonably required for decision-making in accordance with the intent of the Charter.

- Meta did not answer **seven questions** in the *former President Trump's suspension* case. The questions Meta did not answer asked how Facebook's News Feed and other features impacted the visibility of Mr. Trump's content; whether Meta had researched, or had plans to research, those design decisions in relation to the events of 6 January 2021; and information about violating content from followers of Mr. Trump's accounts. The Board also asked questions related to the suspension of other political figures and removal of other content; whether Meta had been contacted by political officeholders or their staff about the suspension of Mr. Trump's accounts; and whether account suspension or deletion impacts the ability of advertisers to target the accounts of followers. Meta stated that this information was not reasonably required for decision-making in accordance with the intent of the Charter; was not technically feasible to provide; was covered by lawyer/client privilege; and/or could not or should not be provided due to legal, privacy, safety or data protection concerns.
- Meta did not answer **three questions** in the *Öcalan's isolation* case. The first question asked whether Meta could determine how many pieces of content were wrongly taken down while a policy was not being enforced. The second question asked how much content mentioning Öcalan Meta had removed or left up in the last five years. For both questions, Meta responded that it was not technically feasible to determine this information. The third question asked, in part, why moderators considered the post a call to action to support Öcalan and the Kurdistan Workers' Party (PKK). Meta responded that, as the company does not require its at-scale content reviewers to document their reasoning for each content decision, it was unable to provide a detailed description of the basis for the post's removal.
- Meta did not answer **two questions** in the *South Africa slurs* case. The first question asked Meta to provide metrics on three offensive terms used in the post. Meta responded that it was not technically feasible to provide the requested information. The second question asked whether Meta's Public Policy Team based in South Africa included people from South Africa. Meta declined to provide the requested information, citing the Board's Bylaws.

Of the 19 questions Meta did not answer, around two-thirds (13) related to cases where the content was removed under the Dangerous Individuals and Organizations Community Standard. These were the *Nazi quote* case, *Punjabi concern over the RSS* in India case, *former President Trump's suspension* case and the *Öcalan's isolation* case. Five of the remaining questions related to cases where the content was removed under the Hate Speech policy, while one question related to content removed under Meta's Violence and Incitement policy.

Public Comments

As part of our decisions process, individuals and organizations can submit public comments. In our first year, these have given people a voice in our decisions, providing crucial expertise on language, culture, politics, and human rights. This input is crucial to achieving our goal of improving how Meta treats people and communities around the world. On numerous occasions, public comments have shaped our decisions and our recommendations to Meta.

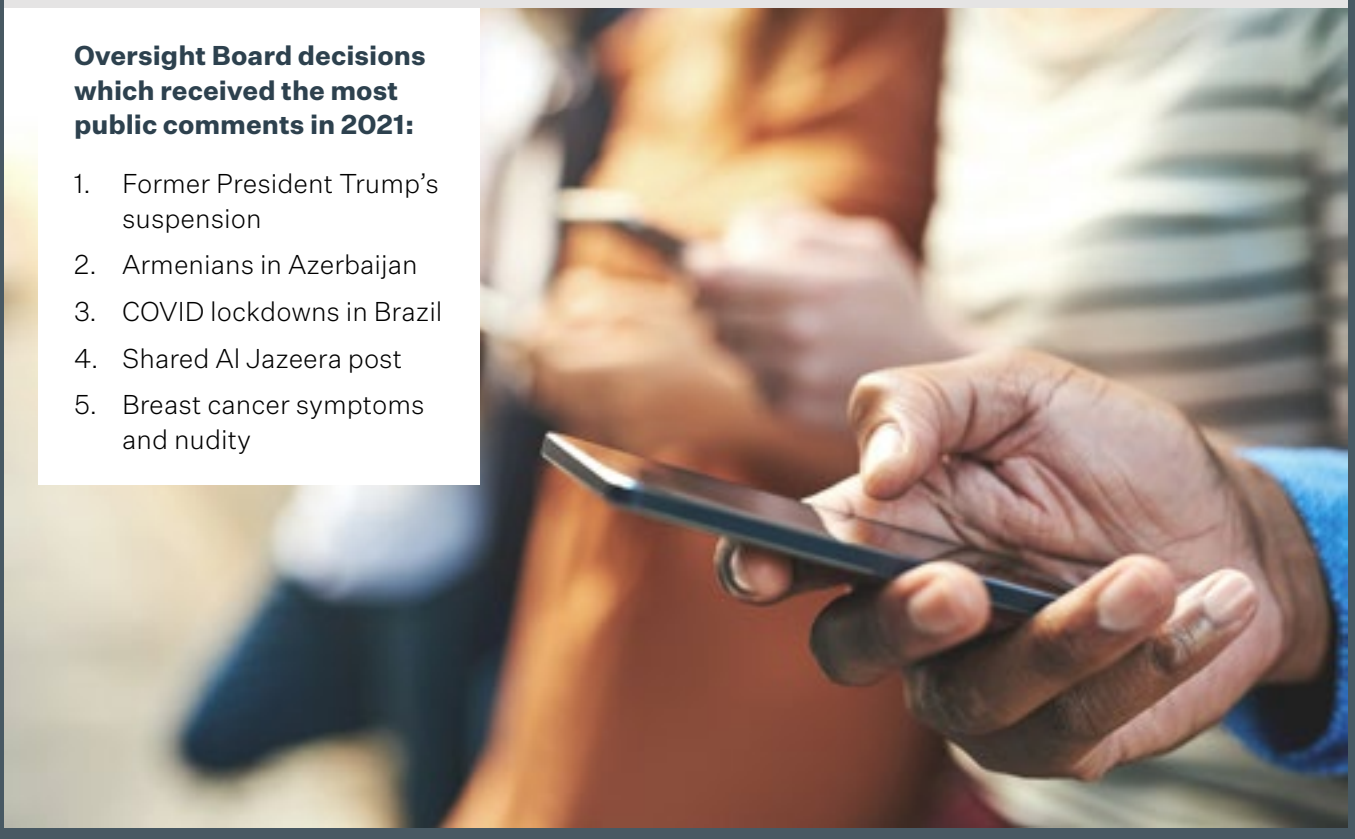
For the 20 decisions published in 2021, we received 9,986 public comments from individuals and organizations around the world. The overwhelming majority of these (97%) were submitted for our decision on *former President Trump's suspension,* but it is important to stress the quality and value of the expertise we received in many cases, even when the absolute number of comments may be limited. Some issues are deeply complex and require additional specialist expertise that public comments can provide.

96% of public comments came from the U.S. and Canada, largely due to the number of comments from the United States on the *former President Trump's suspension* decision. 95% of public comments came from individuals, while 5% came from organizations.

In one third of published public comments, people submitted their comment anonymously. It is interesting that many chose to make use of this option, sharing their expertise while also protecting their right to privacy. The three decisions where more than half of published comments were submitted anonymously (*Punjabi concern over the RSS in India*, *"Two Buttons" meme*, and *Colombia protests*) all related to content criticizing governments which are currently in power.

Oversight Board decisions which received the most public comments in 2021:

1. Former President Trump's suspension
2. Armenians in Azerbaijan
3. COVID lockdowns in Brazil
4. Shared Al Jazeera post
5. Breast cancer symptoms and nudity



How public comments have shaped our decisions

Public comments have played a vital role in the 20 decisions the Board published in 2021. Comments have raised users' concerns related to cases, and provided linguistic, cultural and context which has enriched our decisions. During 2021, the Board noticed an increase in the quality of public comments submitted, with more comments being explicitly referenced or quoted in decisions published later in the year.

> Through public comments, the Board can stand with the voices of civil society and raise them, and this time Meta has to respond to them."
>
> Afia Asantewaa Asare-Kyei
> *Board Member*

Providing linguistic expertise

In several cases, public comments provided crucial linguistic context. In the *Myanmar bot* case, public comments agreed with the Board's linguistic experts that in Burmese the same word is used to refer to states and people from that state. This was essential context for the Board's conclusion that as the profanity in the post targeted "China" the state and not the "Chinese" as a people, the content should be restored.

In the *South Africa slurs* case, public comments also helped the Board understand the local significance of the slurs used in the post. This contributed to our conclusion that one of the slurs was a particularly hateful and harmful word in the South African context. This, in turn, underpinned our decision not to restore the post to Facebook.

Offering greater context

In our decision about *COVID lockdowns in Brazil* public comments emphasized the politicization of measures to counter the spread of COVID-19 in the country. This contributed to one of the Board's recommendations that, given the pandemic, Meta should prioritize the fact-checking of potential health misinformation shared by public authorities, taking into consideration the local context.

Voicing users' concerns

Public comments also provided an opportunity for people to raise wider issues related to the case. Comments submitted for the *shared Al Jazeera* post case, for example, included allegations that Meta disproportionately removed or demoted content from Palestinian users and content in Arabic, especially compared to its treatment of posts threatening anti-Arab or anti-Palestinian violence within Israel. These comments contributed to a recommendation in this case that urged Meta to conduct an examination into whether its content moderation in Arabic and Hebrew had been applied without bias.

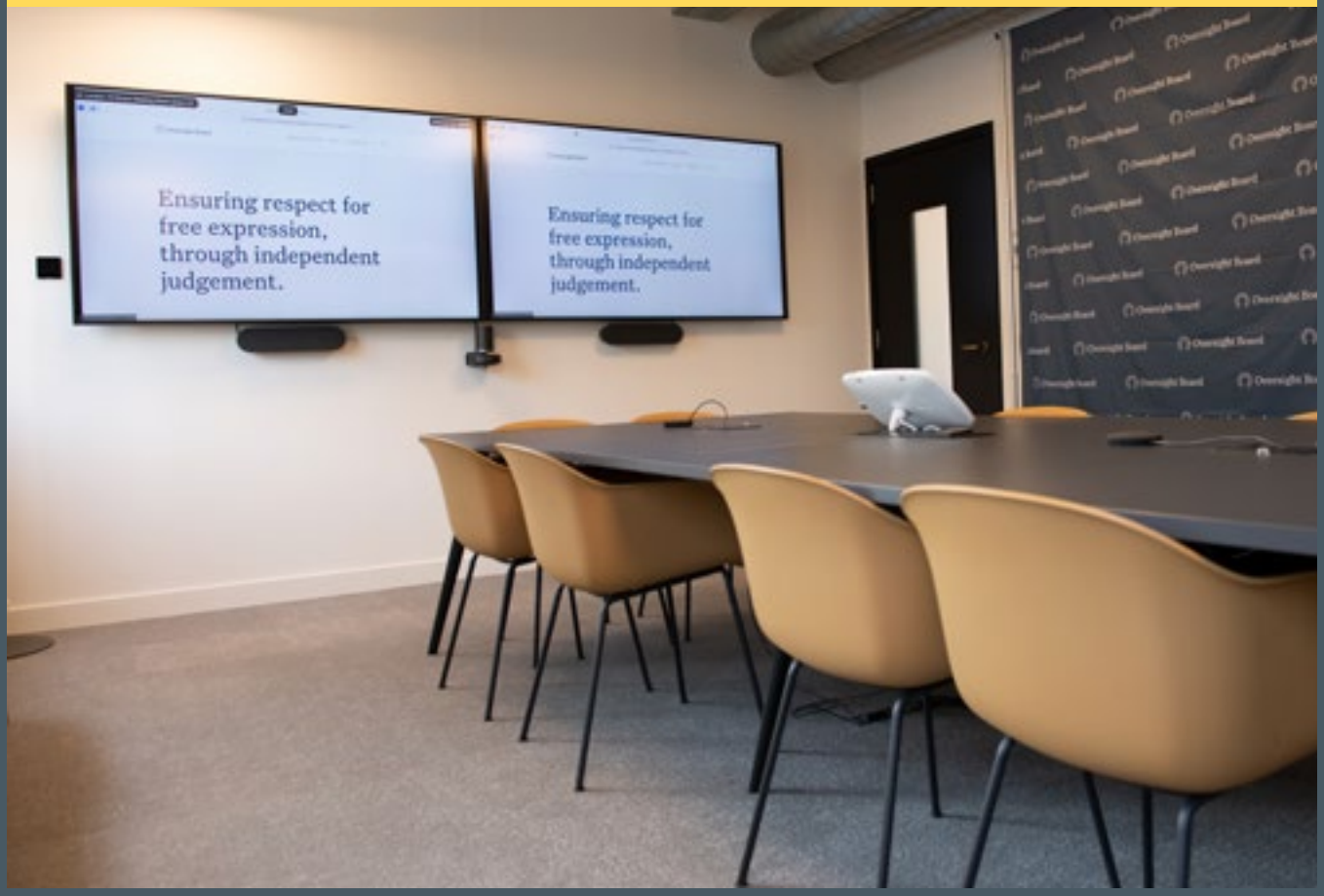
Assessing Meta's human rights commitments

The Board received three public comments from UN Special Rapporteurs, as well as other human rights experts from around the world. In the *Öcalan's isolation* case, the UN Special Rapporteur on human rights and counter terrorism noted that Meta's Community Guidelines and Community Standard on Dangerous Individuals and Organizations "are insufficiently consistent with international law and may function in practice to undermine certain fundamental rights, including but not limited to freedom of expression, association, participation in public affairs and non-discrimination."

LESSONS LEARNED

Since we announced our first cases in December 2020, we've consistently listened to users and improved our public comments process.

In response to stakeholder feedback, we provided more detail in case summaries and added 'prompt' questions to highlight areas where we would particularly welcome stakeholder input. To help as many people and organizations as possible to engage with our work, we extended the original seven-day window for submitting public comments to 10 days, and then 14 days. We also increased the page limit to allow more detailed submissions. While we've always accepted comments in the language of the post and English, in 2022 we plan to offer users the option to submit comments in more languages.



Recommendations

86 Recommendations were made to Meta in 2021

Following our recommendations, Meta committed to:

- **Be more specific** with users when removing hate speech posts
- Translate its rules into **languages spoken by 400+ million people**
- Provide new info on **government requests**
- Roll out new messaging in certain locations telling users whether **automation** or **human review** resulted in their content being removed
- Adopt a new **Crisis Policy Protocol** to govern its response to crisis situations

This report takes a **new, data-based approach** to track implementation and ensure Meta honors its commitments on recommendations.

of our 86 recommendations, Meta either demonstrated implementation or reported progress, with recommendations on **transparency** most likely to fall into these categories.

Meta's responses to recommendations **improved** over time.

Overview

In the 20 decisions we published in 2021, we made 86 recommendations to Meta for how it can improve the policies and procedures it applies to content posted by its billions of users.

While our recommendations are advisory, Meta must respond to them publicly. In 2021, Meta had 30 days from the publication of a decision to respond to our recommendations. This has since been extended to 60 days.

Our recommendations have repeatedly urged Meta to follow some central tenets of transparency: make your rules easily accessible in your users' languages; tell people as clearly as possible how you make and enforce your decisions; and, where people break your rules, tell them exactly what they've done wrong.

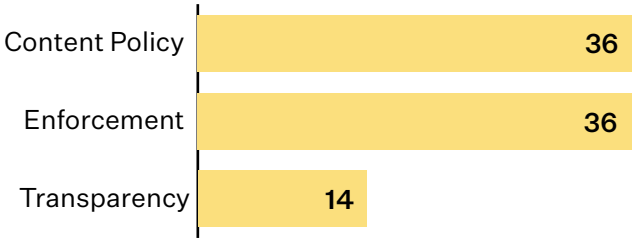
Early commitments, growing impact

In 2021, we saw Meta make commitments in response to our recommendations which, taken together, have real potential for improving how the company treats Facebook and Instagram users around the world.

Tell users what your rules are and what they've done wrong:

- In 2021, we made one recommendation to Meta more times than any other, repeating it in six decisions: when you remove people's content, **tell them which specific rule they broke.** A result of Meta's positive response is that people using Facebook in English now receive specific messaging when their content is removed for hate speech. Meta is now testing this specific messaging to apply to content in Arabic, Spanish, and Portuguese, as well as for posts removed for bullying and harassment.

What kind of recommendations did the Board make in 2021?



- In our three decisions about content on Instagram, we urged Meta to tell users that **Facebook's Community Standards apply to Instagram.** The company has said it is "fully implementing" this recommendation.
- In response to our recommendation to explain and provide examples of key terms in its **Dangerous Individuals and Organizations policy**, Meta added definitions and examples of "praise," "substantive support," and "representation." These clarifications are crucial for users to understand what is and isn't allowed.

Be transparent with users about how you moderate their content:

- In response to a recommendation in our *breast cancer symptoms and nudity* decision, Meta is rolling out new messaging in certain locations telling users whether automation or human review resulted in their content being removed.
- Adopting another of our proposals, Meta agreed to provide information on content removed for violating its Community Standards **following a formal report by a government.** This will increase transparency about how governments put pressure on Meta and how the company responds to that pressure.

- Responding to another recommendation, Meta has begun necessary preparations to report instances where it applies its **newsworthiness allowance** in its quarterly Community Standards Enforcement Report. The company expects to implement this by the end of 2022.

Treat users fairly, wherever they are:

- Following another recommendation, Meta **translated Facebook's Community Standards into Punjabi and Urdu.** It also committed to translate Facebook's rules into Marathi, Telugu, Tamil, Gujarati, and other languages in early 2022. Once completed, **over 400 million more people,** primarily in South Asia, will have Facebook's Community Standards in their native language.
- After hearing claims from stakeholders that Meta had disproportionately removed posts from Palestinians, we urged Meta to examine whether its **content moderation in Hebrew and Arabic was biased.** It agreed to do so.

Improve how you enforce your rules, and apply them consistently,

- In response to a recommendation in our *breast cancer symptoms and nudity* decision, which dealt with a post raising awareness about the illness that had been wrongly removed, Meta updated its automatic nudity detection models to account for health contexts.
- In our *Öcalan's isolation* decision, Meta first removed the content because, it claimed, the user broke its rules on dangerous individuals. It then disclosed, however, that it had found **a policy exception that it had lost for three years,** which clearly allowed the post. The exception specifically allowed people to discuss the conditions of confinement of people who appear on its list of individuals who belong to "dangerous organizations." As a result of our recommendation in this case, Meta says it has now restored this guidance. This will mean that, in the future, users will have a clearer knowledge of how Meta moderates content related to dangerous individuals on its platforms.
- In response to a recommendation in our decision on *former President Trump's suspension*, and following consultation with more than 50 global experts from outside the company, Meta adopted a Crisis Policy Protocol to govern its responses to crises. This will provide a more consistent, transparent basis for how the company responds to crisis situations.

LESSONS LEARNED

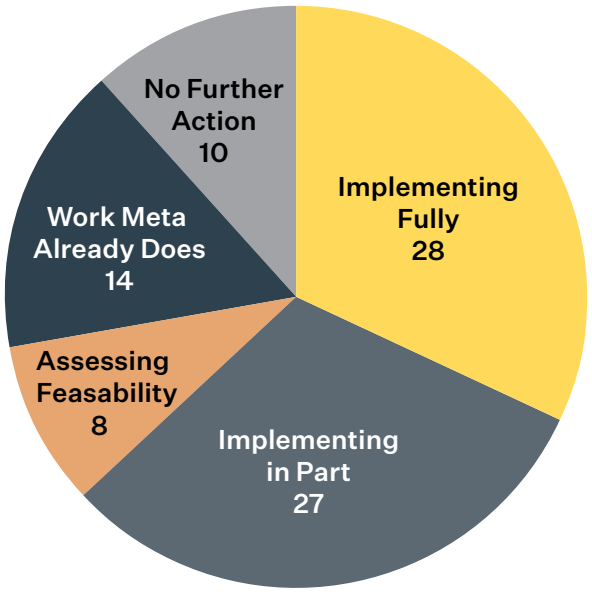
Despite issues with Meta's responses to our first recommendations, both Meta and the Board have taken action to improve the recommendations process during 2021. While this work is already producing results, key questions remain. How can we make our recommendations more meaningful? How can we work with Meta in this area without compromising our independence? And, crucially, how can we ensure that Meta honors its commitments, through actions that can be measured by the Board and felt by people across the world? In response to these points, we established a new, more rigorous approach to monitoring how Meta implements our recommendations, described later in this chapter.

How Meta responded to our recommendations

The chart on this page sets out Meta's responses to recommendations made by the Board in 2021.

Meta committed to implement most of our recommendations, with 55 out of Meta's 87 responses falling into "implementing fully," or "implementing in part."

For eight recommendations, Meta responded that it was "assessing feasibility," while it claimed that 14 recommendations represented work it already does. Meta also said it would take no further action in response to 10 recommendations. These figures reflect the responses set out in Meta's Q4 Quarterly Update on the Board, published in March 2022. The number of responses from Meta (87) does not match the total number of recommendations we report here (86) due to differences in counting the Board's first set of recommendations.



Improving the recommendations process

During 2021, we saw progress on recommendations. The Board got better at making recommendations and Meta improved at responding to them. Despite some issues raised by Meta's responses to our first recommendations, both sides iterated their approach. We entered 2022 with a clearer understanding of the recommendations process and how we can improve it.

In February 2021, Meta responded to the recommendations we made in our first five decisions. While it committed to pursue most of our proposals, in some cases the company reframed or reworded our recommendations. In other cases, Meta only responded to certain parts of our proposals or did not seem to fully understand the point we were trying to make.

Throughout its first year of work, we have also made changes to our processes to strengthen the Board's recommendations.

In our *Depiction of Zwarte Piet* decision, we created a new 'policy advisory statement' section in our decisions, which clearly numbered our recommendations and stated that "the Board requests that [Meta] provides an individual response to each as drafted." From October, we also started using our quarterly transparency reports to note instances where Meta had misrepresented or misunderstood our recommendations.

We also recognize that, for its part, Meta took steps in 2021 to improve its response to our recommendations.

Meta made its response categories more specific over time, replacing "committed to action," with "implementing fully" or "implementing in part," for example. The company also asked BSR (Business for Social Responsibility) to explore options for how it should interact with us on implementing our recommendations.

Report on the timeliness of Meta's implementation of and response to our recommendations

- Under our Bylaws which applied in 2021, Meta had to respond to our recommendations publicly within 30 days of the Board publishing a decision.
- In 2021, Meta responded to our recommendations within this timeframe.

Overall, by the close of the year the company was providing more comprehensive responses to our recommendations, compared to at the start of the year.

Two examples of responses to recommendations made in late 2021 illustrate this progress. When responding to a recommendation from our *Wampum belt* decision, Meta described two experiments it had conducted in areas relevant to our proposal. This shows how the process of making and responding to recommendations is increasing transparency and revealing new information about Meta's internal processes. In a separate response to a recommendation from our *Ayahuasca brew* decision, Meta noted that it would assess feasibility in its Product Policy Forum, which publishes minutes of its discussions. We welcome more of these types of responses from Meta.

Meta has also become more transparent with the Board about empirical evidence of its impact. In a recent feasibility assessment, Meta learned that notifying users of which hate speech violation subtype resulted in their content being removed led to users better understanding how and why the rules were applied. Meta shared that it found a statistically significant increase in perceptions of its transparency and legitimacy across multiple markets. In light of this, Meta is now experimenting with providing users with more specific notifications for bullying and harassment violations.

On several occasions in 2021, Meta raised issues around the level of resourcing required to respond to our recommendations. In its first quarterly transparency report on the Board, published in July 2021, the company said:

> *"[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."*

Meta unpacked these concerns in its next quarterly transparency report on the Board published in November 2021:

> *"[T]he pace and volume of the recommendations do not allow us adequate time to initially assess and implement the recommendations. [...] [O]n average, our teams assess and respond to anywhere from 5 to 35 recommendations at any one point in time. The majority of these recommendations require over a dozen people to assess feasibility, which we cannot easily complete in 30 days. This difficulty is further compounded by our need to incorporate the recommendations into existing initiatives and priorities, such as those reflected in our Integrity and related product teams' 'roadmaps.'"*

In 2021, our recommendations pushed for Meta to make ambitious changes to how users experience Facebook and Instagram, some of which might require it to make significant technical changes. We know that it takes time first to assess whether

these changes are feasible, and then provide us with a comprehensive response. In response to those concerns, we subsequently agreed to extend Meta's deadline for responding to our recommendations from 30 to 60 days.

Not counting the *former President Trump's suspension* and *Öcalan's isolation* cases, where Meta specifically requested detailed policy guidance, decisions published in 2021 included, on average, around three recommendations each. We have thus sought to ensure that the Board's recommendations are both implementable and implemented, and engaged with Meta on the resourcing required to action them.

Improvement in Meta's responses to our recommendations

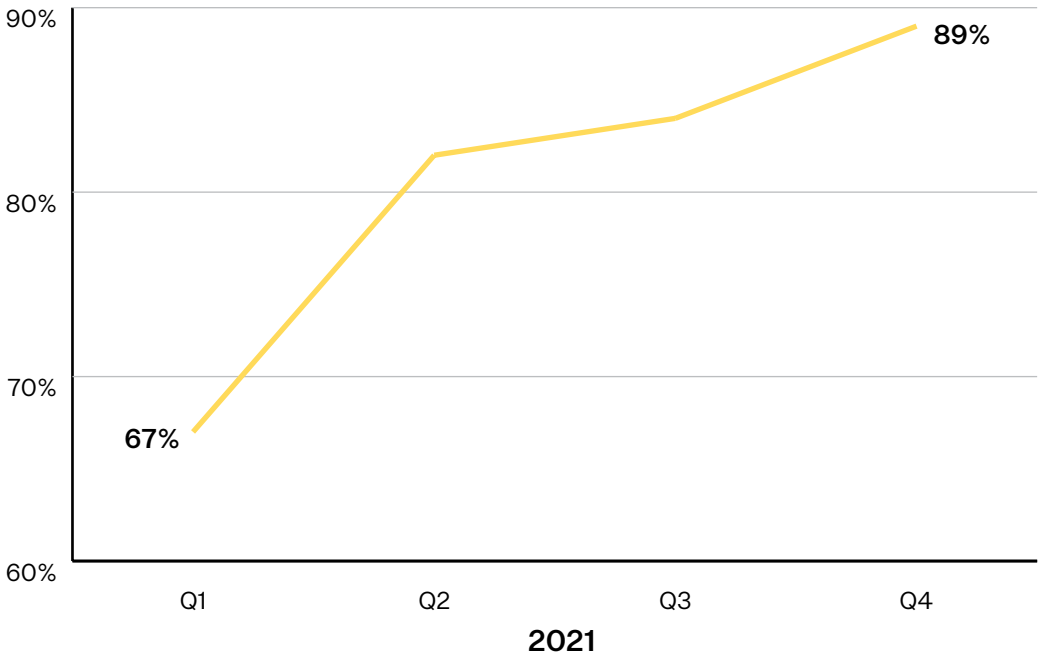
This report contains new analysis from the Board on Meta's responses to recommendations we made during 2021. Our 'response' measurement assesses whether Meta provided a comprehensive response to the Board's recommendation. These statistics reflect Meta's quarterly updates on the Board through Q3 2021.

In assessing each response, we asked three questions: (1) Does it address all parts of our recommendation? (2) Does it provide a commitment to action? (3) Does it provide a timeline? Where Meta met one of these criteria we deemed it to have provided a 'somewhat comprehensive,' and where it met at least two we deemed it to have provided a 'comprehensive' response.

When we examine Meta's responses to recommendations published in different quarters, we can see that the share of responses deemed either 'comprehensive' or 'somewhat comprehensive' increased in each period, rising from 67% in Q1 to 89% in Q4 2021.

While there is still room for improvement, this seems to suggest that changes to the recommendations process in 2021 have led to more comprehensive responses from Meta.

How often did Meta provide a 'comprehensive' or 'somewhat comprehensive' response to our recommendations?



From Commitments to Action: Getting Results for Users

> It's important to ensure the commitments Meta has made, some of which will take time, are not kicked into the long grass and forgotten about."
>
> Thomas Hughes
> *Oversight Board Director*

While Meta committed to implement most of our recommendations in 2021, our next task is to ensure that Meta keeps its promises. The Board's top priority is making sure that Meta turns its commitments into actions that improve how the company treats people around the world.

As our recommendations grew in number during 2021, we expanded our work on implementation. This report sets out a rigorous, data-driven approach to tracking how Meta implements our recommendations and more clearly assessing their impact on users.

Building our ability to hold Meta to account

From mid-2021, we started expanding our ability to monitor how Meta was implementing our recommendations. In July, we created and staffed the Case Implementation and Monitoring Team to support the Board. This team, which has since expanded, monitors and measures Meta's responses and actions, to understand the impact of our recommendations on Facebook and Instagram users.

In mid-2021, we also established an Implementation Working Group made up of Board Members and senior Meta staff. This group has met several times to discuss and improve the recommendations process, with Meta answering our questions about its processes to implement our proposals. Later in the year, we also established an Implementation Committee currently made up of five Board Members to sit alongside our Case Selection and Membership Committees. This represented a clear choice to place implementation on par with our organization's most critical functions. The Implementation Committee has led efforts on sharpening the Board's recommendations, and ensuring they are focused on specific, measurable impacts.

A new, data-driven approach to implementation

This report marks another milestone in our journey to hold Meta to account. In the annex to this report we provide, for the first time, our assessment of how Meta has implemented each of the 86 recommendations we made in 2021. These statistics reflect Meta's updates through Q3 2021.

To measure Meta's progress on implementation, we looked at whether certain criteria for a given recommendation have been met. These vary depending on the recommendation. For example, if we proposed that Meta add extra detail to its public-facing Community Standards or issue a human rights due diligence report on a certain topic, then publishing these externally would demonstrate implementation. For other recommendations, Meta would need to provide data which isn't publicly available to demonstrate implementation. This could mean providing user notification data to prove that it is telling users which specific rule

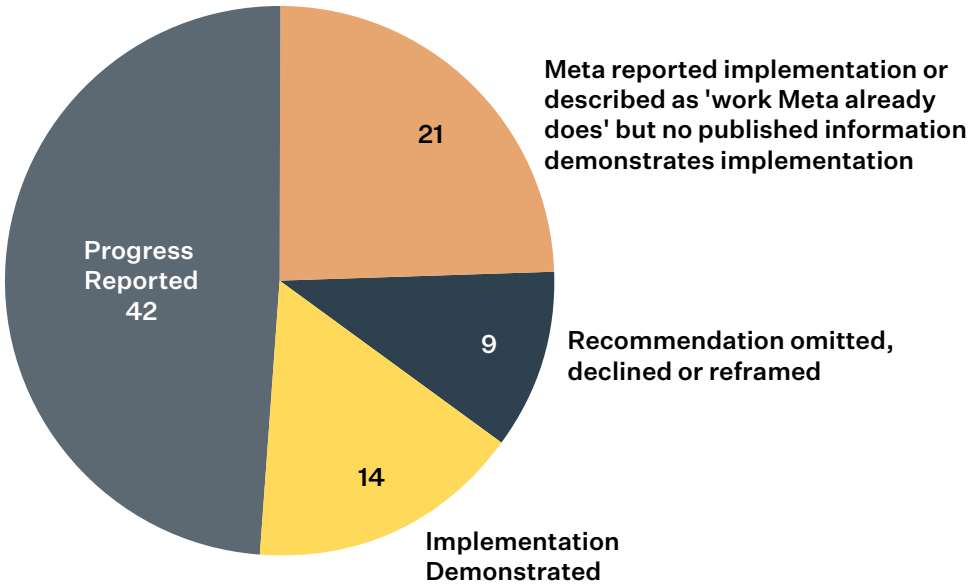
they broke. Going forward, we will measure Meta's implementation of our recommendations according to four categories, updating our assessments on a quarterly basis:

- *'Implementation demonstrated'* – Meta has provided sufficient data for us to say that this recommendation has been completed.
- *'Progress reported'* – Meta has made a commitment to implementing this recommendation but has not yet completed all necessary actions.
- *'Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation'* – Meta says it has implemented this recommendation but has not provided sufficient evidence for us to verify this.
- *'Recommendation omitted, declined or reframed'* – Meta will not take any further action on our proposal.

This new, data-driven approach means that our assessment of whether Meta has implemented a recommendation may at times differ from the company's reports. We believe, however, that this kind of independent validation is crucial to hold Meta to account and ensure that users feel the impact of our recommendations. The chart on this page provides a breakdown of our assessment of how Meta has implemented our recommendations.

This shows that in 2021 Meta demonstrated implementation of 14 of our recommendations. The fact that the Board has only been making recommendations for a relatively short period of time, coupled with the high level of ambition of the proposals we made in our first year, means that it will take time for Meta to complete our recommendations and demonstrate that they have done so. In all, 42 recommendations fell into 'progress reported' category, which includes many of the high-impact recommendations set out in the introduction to this section. We expect that many recommendations in this category will move into 'implementation demonstrated' during 2022.

Meta's implementation of our recommendations



For 21 recommendations, while Meta reported that it had implemented our proposal, or described it as work it already does, it did not publish enough information for us to validate implementation. Going forward, we encourage Meta to share more data and provide more evidence to support statements that it has implemented our recommendations. We are also pushing Meta to grant greater access to the company's data to allow progress to be validated. Nine recommendations fell into the 'recommendation omitted, declined or reframed' category, meaning that Meta will take no further action on these proposals.

When we looked at the different types of recommendations assessed as 'implementation demonstrated' or 'progress reported,' we found that 79% of transparency recommendations and 67% of content policy recommendations fell into these categories, compared to 58% for enforcement recommendations. The lower rate for enforcement recommendations may reflect the fact that these proposals often require Meta to make broader changes on Facebook and Instagram.

Getting results for Facebook and Instagram users

While validating how Meta is implementing our recommendations is a hugely complex task, we have chosen to prioritize and invest in this area. The reason for these investments is simple: only by helping to turn Meta's commitments into actions can we improve how the company treats Facebook and Instagram users in the long-term, and honors its human rights responsibilities.

Holding Meta to Account on Cross-Check

On September 13, 2021, *The Wall Street Journal* reported that Meta's cross-check program was far more comprehensive in scale and scope than the company had previously disclosed. One of the first in a series of articles based on company documents obtained from a whistleblower stated:

> *Mark Zuckerberg has publicly said that [Meta] allows its more than three billion users to speak on equal footing with the elites of politics, culture, and journalism, and that its standards of behavior apply to everyone, no matter their status or fame. [But] in private, the company has built a system that exempted high-profile users from some or all its rules.*[7]

The article stated that cross-check had expanded from a program "initially intended as a quality-control measure for actions taken against high-profile accounts, including celebrities, politicians and journalists" into a program which shielded "millions of VIP users" from the company's normal enforcement processes. Some high-profile posters were, according to the whistleblower documents, "whitelisted." The article reported that part of this process enabled individual Meta employees to add certain entities (e.g. profiles, pages) to lists which allegedly enabled "rule-violating material" to stay up indefinitely, "pending Facebook employee reviews that often never come."[8]

On September 21, 2021, we published our response to the revelations:

> *At the Oversight Board, we have been asking questions about cross-check for some time. In our decision concerning former US President Donald Trump's accounts, we warned that a lack of clear public information on cross-check and [Meta's] 'newsworthiness exception' could contribute to perceptions that [the company] is unduly influenced by political and commercial considerations.*

Among the recommendations we sent to Meta as part of that decision, one had urged it to "report on the relative error rates and thematic consistency of determinations made through the cross-check process compared with ordinary enforcement procedures." In its response, Meta declined to implement this proposal on the grounds that "track[ing] this information is not feasible." Meta had backed up that claim with a link to a 2018 blog post that stated: "We remove content from Facebook, no matter who posts it, when it breaks our standards." *The Wall Street Journal,* however, reported that a 2019 internal review by Meta had found that blog post to be "misleading."[9]

The Board concluded, "In light of recent developments, we are looking into the degree to which [Meta] has been fully forthcoming in its responses in relation to cross-check, including the practice of whitelisting. The Board has reached out to [Meta] to… further clarify the information previously shared with us. We expect to receive a briefing in the coming days."

In that briefing, Meta's team conceded that it "should not have said that cross-check only applied to a 'small number of decisions'" in responding to our recommendation in the case related to former U.S. President Trump. On September 28, 2021, a week after our September 21 blog post, Meta sent its response to the Board's questions in the form of its own request for a policy advisory opinion. Meta asked us to include specific recommendations in our opinion on how to improve cross-check. In announcing this referral, Meta noted:

7 "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt," Jeff Horwitz, *The Wall Street Journal,* Sept. 13, 2021

8 Ibid.

9 Ibid.

Recently the Board has expressed interest in looking more into our cross-check system. This referral goes beyond the briefing we have already provided. We are proactively asking the Board for its input through a formal and transparent process... Specifically, we will ask the Board for guidance on the criteria we use to determine what is prioritized for a secondary review via cross-check, as well as how we manage the program.[10]

In early October, prior to publishing the Board's first quarterly transparency report, we announced that we would be meeting with Frances Haugen, the whistle-blower previously employed by Meta, to discuss the significant issues raised by her disclosures about Meta's content moderation procedures and policies.

In the last few weeks, new information about [Meta's] approach to content moderation has come to light as a result of the actions of a former [Meta] employee, Frances Haugen. In light of the serious claims made about [Meta] by Ms. Haugen, we have extended an invitation for her to speak to the Board over the coming weeks, which she has accepted. Board members appreciate the chance to discuss Ms. Haugen's experiences and gather information that can help push for greater transparency and accountability from [Meta] through our case decisions and recommendations.

Our blog concluded:

The choices made by companies like [Meta] have real-world consequences for the freedom of expression and human rights of billions of people across the world. In this context, transparency around rules is essential.

As a Board, we will continue to ask Facebook difficult questions and push the company to commit to greater transparency, accountability and fairness. Ultimately, only this can give users the confidence that they are being treated fairly.[11]

In our first quarterly transparency report, published later that month, we stated that, in our view, Meta's internal team tasked with providing information had not been "fully forthcoming" about cross-check. "On some occasions," we elaborated, "[that team] failed to provide relevant information to the Board… [and] in other instances, the information it did provide was incomplete." In light of this, we declared that the Board's future working relationship with Meta would depend on our ability to "trust that the information provided by Meta is accurate, comprehensive, and paints a full picture of the topic at hand." In keeping with that goal, on October 21, 2021 we announced that we had accepted Meta's request to formulate a policy advisory opinion that would review its cross-check program and make recommendations on how it could be improved.

10 "Requesting Oversight Board Guidance on Our Cross-Check System," Meta Blog Post by Nick Clegg, VP, Global Affairs, September 28, 2021 (Updated October 21, 2021)

11 "Oversight Board to meet with Frances Haugen" Oversight Board Blog Post October 2021

Engagement and Outreach

In our first year, we have engaged with individuals and organizations across the world, explaining what the Board does, encouraging input to our decisions, and discussing why a principled, global approach to content moderation matters to users.

Julie Owono

Nighat Dad

Suzanne Nossel

Encouraging people to share their perspectives

Much of our external engagement has encouraged individuals and organizations to get involved with our work and to submit public comments for ongoing cases. We hosted 19 'Office Hours' sessions explaining how to submit public comments which were attended by more than 500 stakeholders. To encourage contributions from across the world, these sessions were held across multiple time-zones, with sessions offering interpretation in languages such as Arabic, Hebrew, Portuguese, and Spanish. We also held 12 events and roundtables in the two weeks after announcing the case to encourage public comments for the case on *former President Trump's suspension*. These were attended by more than 1,000 people.

Driving the debate on content moderation and human rights

Our approach to applying human rights to content moderation has not just been shaped by internal deliberations between Board Members, but also through public discussions with representatives from academia, journalism, civil society, and international institutions.

In June, we held events at RightsCon, including a session on aligning content moderation with human rights. In September, during the UN General Assembly, Board Member Julie Owono joined two UN panels on countering disinformation and hate speech, and promoting transparency. In December, Board Member and former UN Special Rapporteur Maina Kiai spoke at an event discussing lessons the Board learned about human rights protection and digital platform governance. Speaking alongside Peggy Hicks, Director at UN Human Rights, Maina Kiai said, "At the Oversight Board, we understand that we are tackling only a part of the problem — experts in many other fields must take action as well." We also hosted an event at the Internet Governance Forum in December. This featured the Board's Director Thomas Hughes, Board Member Afia Asantewaa Asare-Kyei, and Board Trustee Cherine Chalaby, discussing questions surrounding digital rights and self-regulation.

Making the debate on content moderation global

Social media affects events, and billions of people, around the world. Yet, too often, the debate around content moderation only considers users in Europe and North America. To open this debate to everyone, we prioritized events examining content moderation challenges faced by those in the Global South.

In May, Board Members Catalina Botero-Marino and Jamal Greene participated in an Inter-American Dialogue event discussing what the Board's decisions mean for the Global South, and in particular Latin America. In August, Board Member Endy Bayuni took part in an event with FORUM Asia on human-centric cyber security. In September, Board Member Julie Owono discussed lessons in content moderation from Western and Central Africa, at an event attended by the Senegalese Commission of Personal Data Protection.

> "What is proposed in the Global North can damage communities in the Global South."
>
> Nighat Dad
> *Board Member*

Finally, in November, Board Members Nighat Dad, Ronaldo Lemos and Maina Kiai spoke at the Paris Peace Forum, with Nighat Dad noting that "Some of the regulations that we see in the Global South are copy-pasted from the Global North and don't consider the local context and realities of our region. What is proposed in the Global North can damage communities in the Global South."

What's Next: 2022 and Beyond

In 2021, the Oversight Board delivered 20 decisions and 86 recommendations to Meta. Taken together, they show our commitment to holding the company to account and steering it toward greater transparency.

As we enter 2022, on top of adding new Board Members, we are looking to deepen our impact, improving how Meta treats people and communities around the world.

Broadening our scope

We are in dialogue with Meta on expanding the Board's scope, including to review user appeals against the company's decisions in areas such as groups and accounts, as opposed to just individual posts, and expect to report progress on this in the next year. We are also looking at our role in assessing Meta's content moderation plans for the emerging 'Metaverse.' In 2021, the company spent $10 billion on the Metaverse, indicating that this will be a major area of focus in years to come.

Making a global impact

As Meta's impacts are felt across the world, the Board's work must be global in scope. In 2021, more than two-thirds of user appeals came from Europe, the U.S. and Canada, with 49% from the U.S. and Canada, and 20% from Europe. By contrast, just 14% came from Latin America and the Caribbean; 9% from Asia Pacific and Oceania; 4% from the Middle East and North Africa; 2% from Central and South Asia, and 2% from Sub-Saharan Africa. This distribution reflects an imbalance we are working to address.

In 2022 we are expanding our outreach and stakeholder engagement to work with policy makers and civil society leaders in Asia, Latin America, the Middle East, and Africa. We are building a global network of regional consultants to encourage people to submit appeals and submit public comments in their respective regions. We also plan to expand the number of languages in which we accept public comments beyond English and the language of the content in question.

Publishing policy advisory opinions

Beyond reviewing individual cases to remove or restore content, the Board can also accept requests for policy advisory opinions from Meta. These allow us to review the company's policies in a given area and make recommendations on how they can be improved. In 2021, we accepted two such requests from Meta on sharing private residential information and the company's cross-check system. Our opinion on sharing private residential information was published in February 2022, and our opinion on cross-check will be issued later this year.

Securing greater access to data

In 2022, as part of our work to ensure that Meta keeps its commitments, the Board is seeking access to Facebook data and research. This will better inform the cases the Board selects, the recommendations we make, and the impact we have. We also plan to hire data scientists as part of our efforts to understand and increase our impact.

Sharing what we have learned so far

In our first year, we developed, refined, and implemented a new approach to content moderation by applying a framework based on international human rights standards. In doing so, we developed a set of best practices and a wealth of experience.

In 2022, we look forward to sharing what we have learned in our common endeavor to improve social media. We will also further develop our role as a source of best practice for self-regulation within the wider evolving content moderation and regulatory landscape. We will share our expertise about self-regulatory solutions across tech policy conversations, as well as contribute to discussions about content moderation innovation with potential tech industry partners.

Creating an institution built for long-term success

Finally, the Co-Chairs and Board Members, Trustees, and the Administration will continue to build a high-performing and sustainable institution in line with a multi-year strategic plan. This will include autonomous processes for Board Member identification, selection and onboarding, as well as performance and conduct. We will also continue to ensure the efficient and responsible management of the Trust's resources, including setting out a long-term funding model.

Conclusion

The insights in our first Annual Report show that the Oversight Board is already having a tangible impact on the way Meta works and serves its users. Every decision we have taken, experience we have gained, and conclusion we have drawn, represents the first steps on what will be a much longer journey.

The anodyne term "content moderation" obscures the truth that content posted and shared on social media can have significant impacts – positive and negative – on people's lives offline and online. While social media has to a large degree fulfilled its early promise of bringing billions together, it has also created new ways for people to inflict harm on others.

More than the result of any specific content decision, the Board's impact will increasingly be felt by users who stand to benefit from a clear, fair, and transparent application of Meta's own policies. We hope that the Board's work provides a credible framework for other social media companies and platform publishers, regulators, and policymakers around the world as they grapple with the tensions and challenges inherent to content moderation.

At the Oversight Board, we present this first Annual Report as an accurate account of that work. We are proud of what we have accomplished thus far, and clear-eyed about the substantial amount of work ahead. This document reflects, above all, our collective commitment to advancing the values of transparency, accountability, and responsibility, which we hope will make social media platforms better places for users to spend time. We will continue to hold Meta accountable for improving how it treats people and communities around the world.

Annex

How Meta responded to and implemented our recommendations[12]

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of: Meta's response	Board's assessment of: Implementation
Armenians in Azerbaijan				
1	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. Doing so would enable Facebook to encourage expression that complies with its Community Standards, rather than adopting an adversarial posture towards users. In this case, the user was informed that the post violated the Community Standard on hate speech but was not told that the post violated the standard because it included a slur targeting national origin. Facebook satisfied the principle of legality in this instance, but Facebook's lack of transparency left its decision susceptible to the mistaken belief that it had removed the post because the user was addressing a controversial subject or expressing a viewpoint it disagreed with.	E	◐	
Breast cancer symptoms and nudity				
1	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.	E	●	
2	Ensure that users are always notified of the reasons for the enforcement of content policies against them, providing the specific rule within the Community Standard Facebook based its decision on.	E	✕	
3	Inform users when automation is used to take enforcement action against their content, including accessible descriptions of what this means.	E	◐	
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. Where Facebook is seeking to prevent child sexual exploitation or the dissemination of non-consensual intimate images, it should enforce based on its Community Standards on Sexual Exploitation of Adults and Child Sexual Exploitation, Abuse and Nudity, rather than rely on over-enforcing policies on adult nudity. Appeals should still be available in these cases, so incorrect removals of permitted consensual adult nudity can be reversed.	E	✕	
5	Implement an internal audit procedure to continuously analyze a statistically representative sample of automated content removal decisions to reverse and learn from enforcement mistakes.	E	✕	
6	Expand transparency reporting to disclose data on the number of automated removal decisions per Community Standard, and the proportion of those decisions subsequently reversed following human review.	T	◐	
7	Revise the "short" explanation of the Instagram Community Guidelines to clarify that the ban on adult nudity is not absolute.	CP	✕	

12 The assessments in this table take into account Meta's updates on the Board's recommendations through Q3 2021.

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Breast cancer symptoms and nudity				
8	Revise the "long" explanation of the Instagram Community Guidelines to clarify that visible female nipples can be shown to raise breast cancer awareness;	CP	●	
9	Clarify that the Instagram Community Guidelines are interpreted in line with the Facebook Community Standards, and where there are inconsistencies the latter take precedence.	CP	●	
Nazi quote				
1	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 (e.g. for support of a hate organization).	E	◐	
2	Explain and provide examples of the application of key terms used in the Dangerous Individuals and Organizations policy, including the meanings of "praise," "support" and "representation." These should align with the definitions used in Facebook's Internal Implementation Standards. The Community Standard should provide clearer guidance to users on how to make their intent apparent when discussing individuals or organizations designated as dangerous.	CP	●	
3	Provide a public list of the organizations and individuals designated "dangerous" under the Dangerous Individuals and Organizations Community Standard. At a minimum, illustrative examples should be provided. This would help users to better understand the policy and conduct themselves accordingly.	CP	●	
Claimed COVID cure				
1	The Board recommends that Facebook set out a clear and accessible Community Standard on health misinformation, consolidating and clarifying existing rules in one place (including defining key terms such as misinformation). This rule-making should be accompanied with "detailed hypotheticals that illustrate the nuances of interpretation and application of [these] rules" to provide further clarity for users (See report A/HRC/38/35, para. 46 (2018)). Facebook should conduct a human rights impact assessment with relevant stakeholders as part of its process of rule modification (UNGPs, Principles 18-19).	CP	●	
2	To ensure enforcement measures on health misinformation represent the least intrusive means of protecting public health, the Board recommends that Facebook: Clarify the particular harms it is seeking to prevent and provide transparency about how it will assess the potential harm of particular content; Conduct an assessment of its existing range of tools to deal with health misinformation; Consider the potential for development of further tools that are less intrusive than content removals; 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; 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; Make clear within the Community Standards what enforcement option applies to each rule.	E	●	
3	In cases where users post information about COVID-19 treatments that contradicts the specific advice of health authorities and where a potential for physical harm is identified but is not imminent, the Board strongly recommends Facebook to adopt a range of less intrusive measures. This could include labelling which alerts users to the disputed nature of the post's content and provides links to the views of the World Health Organization and national health authorities. In certain situations it may be necessary to introduce additional friction to a post - for example, by preventing interactions or sharing, to reduce organic and algorithmically driven amplification. Downranking content, to prevent visibility in other users' newsfeeds, might also be considered. All enforcement measures, including labelling or other methods of introducing friction, should be clearly communicated to users, and subject to appeal.	E	✕	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Claimed COVID cure				
4	Publish a transparency report on how the Community Standards have been enforced during the COVID-19 global health crisis. This should include: data in absolute and percentage terms on the number of removals, as well as data on other enforcement measures, on the specific Community Standards enforced against, including on the proportion that relied entirely on automation; a breakdown by content type enforced against (including individual posts, accounts, and groups); a breakdown by the source of detection (including automation, user flagging, trusted partners, law enforcement authorities); a breakdown by region and language; metrics on the effectiveness of less intrusive measures (e.g., impact of labelling or downranking); data on the availability of appeals throughout the crisis, including the total number of cases where appeal was withdrawn entirely, and the percentage of automated appeals; conclusions and lessons learned, including information on any changes Facebook is making to ensure greater compliance with its human rights responsibilities going forward.	T	✕	
Protest in India against France				
1	To ensure users have clarity regarding permissible content, the Board recommends that Facebook 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.	CP	●	
Depiction of Zwarte Piet				
1	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.	CP	●	
2	In line with the Board's recommendation in case 2020-003-FB-UA, 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 above mentioned resources that explain the harm this rule seeks to prevent. Facebook should provide a detailed update on its "feasibility assessment" of the Board's prior recommendations on this topic, including the specific nature of any technical limitations and how these can be overcome.	E	◐	
Punjabi concern over the RSS in India				
1	Facebook should translate its Community Standards and Internal Implementation Standards into Punjabi. Facebook should aim to make its Community Standards accessible in all languages widely spoken by its users. This would allow a full understanding of the rules that users must abide by when using Facebook's products. It would also make it simpler for users to engage with Facebook over content that may violate their rights.	CP	●	
2	In line with the Board's recommendation in case 2020-004-IG-UA, the company should restore human review and access to a human appeals process to pre-pandemic levels as soon as possible while fully protecting the health of Facebook's staff and contractors.	E	●	
3	Facebook should improve its transparency reporting to increase public information on error rates by making this information viewable by country and language for each Community Standard. The Board underscores that more detailed transparency reports will help the public spot areas where errors are more common, including potential specific impacts on minority groups, and alert Facebook to correct them.	T	◐	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Former President Trump's suspension				
1	The Board believes that it is not always useful to draw a firm distinction between political leaders and other influential users. It is important to recognise that other users with large audiences can also contribute to serious risks of harm. The same rules should apply to all users of the platform; but context matters when assessing issues of causality and the probability and imminence of harm. What is important is the degree of influence that a user has over other users [...] Facebook must assess posts by influential users in context according to the way they are likely to be understood, even if their incendiary message is couched in language designed to avoid responsibility, such as superficial encouragement to act peacefully or lawfully. Facebook used the six contextual factors in the Rabat Plan of Action in this case and the Board thinks that this is a useful way to assess the contextual risks of potentially harmful speech. The Board stresses that time is of the essence in such situations; taking action before influential users can cause significant harm should take priority over newsworthiness and other values of political communication.	CP	◐	
2	When posts by influential users pose a high probability of imminent harm, as assessed under international human rights standards, Facebook should take action to enforce its rules quickly.	E	◐	
3	While all users should be held to the same content policies, there are unique factors that must be considered in assessing the speech of political leaders. Heads of state and other high-ranking government officials can have a greater power to cause harm than other people. Facebook should recognize that posts by heads of state and other high officials of government can carry a heightened risk of encouraging, legitimizing, or inciting violence - either because their high position of trust imbues their words with greater force and credibility or because their followers may infer they can act with impunity. At the same time, it is important to protect the rights of people to hear political speech. Nonetheless, if the head of state or high government official has repeatedly posted messages that pose a risk of harm under international human rights norms, Facebook should suspend the account for a determinate period sufficient to protect against imminent harm.	CP	●	
4	Periods of suspension should be long enough to deter misconduct and may, in appropriate cases, include account or page deletion.	CP	●	
5	Restrictions on speech are often imposed by or at the behest of powerful state actors against dissenting voices and members of political oppositions. Facebook must resist pressure from governments to silence their political opposition. When assessing potential risks, Facebook should be particularly careful to consider the relevant political context.	E	◐	
6	In evaluating political speech from highly influential users, Facebook should rapidly escalate the content moderation process to specialized staff who are familiar with the linguistic and political context and insulated from political and economic interference and undue influence. This analysis should examine the conduct of highly influential users off the Facebook and Instagram platforms to adequately assess the full relevant context of potentially harmful speech. Further, Facebook should ensure that it dedicates adequate resourcing and expertise to assess risks of harm from influential accounts globally.	E	✕	
7	Facebook should publicly explain the rules that it uses when it imposes account-level sanctions against influential users. These rules should ensure that when Facebook imposes a time-limited suspension on the account of an influential user to reduce the risk of significant harm, it will assess whether the risk has receded before the suspension term expires. If Facebook identifies that the user poses a serious risk of inciting imminent violence, discrimination, or other lawless action at that time, another time-bound suspension should be imposed when such measures are necessary to protect public safety and proportionate to the risk.	CP	●	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Former President Trump's suspension				
8	When Facebook implements special procedures that apply to influential users, these should be well documented. It was unclear whether Facebook applied different standards in this case, and the Board heard many concerns about the potential application of the newsworthiness allowance. It is important that Facebook address this lack of transparency and the confusion it has caused.	T	●	
9	Facebook should produce more information to help users understand and evaluate the process and criteria for applying the newsworthiness allowance. Facebook should clearly explain how the newsworthiness allowance applies to influential accounts, including political leaders and other public figures.	CP	●	
10	For cross check review, Facebook should clearly explain the rationale, standards, and processes of review, including the criteria to determine which pages and accounts are selected for inclusion.	CP	◐	
11	Facebook should report on the relative error rates and thematic consistency of determinations made through the cross-check process compared with ordinary enforcement procedures.	T	✕	
12	When Facebook's platform has been abused by influential users in a way that results in serious adverse human rights impacts, it should conduct a thorough investigation into the incident. Facebook should assess what influence it had and assess what changes it could enact to identify, prevent, mitigate, and account for adverse impacts in future.	E	✕	
13	Facebook should undertake a comprehensive review of its potential contribution to the narrative of electoral fraud and the exacerbated tensions that culminated in the violence in the United States on January 6, 2021. This should be an open reflection on the design and policy choices that Meta has made that may enable its platform to be abused. Facebook should carry out this due diligence, implement a plan to act upon its findings, and communicate openly about how it addresses adverse human rights impacts it was involved with.	T	✕	
14	In cases where Facebook or Instagram users may have engaged in atrocity crimes or grave human rights violations, as well as incitement under Article 20 of the ICCPR, the removal of content and disabling of accounts, while potentially reducing the risk of harm, may also undermine accountability efforts, including by removing evidence. Facebook has a responsibility to collect, preserve and, where appropriate, share information to assist in the investigation and potential prosecution of grave violations of international criminal, human rights and humanitarian law by competent authorities and accountability mechanisms. Facebook's corporate human rights policy should make clear the protocols the company has in place in this regard. The policy should also make clear how information previously public on the platform can be made available to researchers conducting investigations that conform with international standards and applicable data protection laws.	T	◐	
15	Facebook should explain in its Community Standards and Guidelines its strikes and penalties process for restricting profiles, pages, groups and accounts on Facebook and Instagram in a clear, comprehensive, and accessible manner. These policies should provide users with sufficient information to understand when strikes are imposed (including any applicable exceptions or allowances) and how penalties are calculated.	CP	●	
16	Facebook should also provide users with accessible information on how many violations, strikes, and penalties have been assessed against them, as well as the consequences that will follow future violations.	E	●	
17	In its transparency reporting, Facebook should include numbers of profile, page, and account restrictions, including the reason and manner in which enforcement action was taken, with information broken down by region and country.	T	✕	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Former President Trump's suspension				
18	Facebook should develop and publish a policy that governs its response to crises or novel situations where its regular processes would not prevent or avoid imminent harm. While these situations cannot always be anticipated, Facebook's guidance should set appropriate parameters for such actions, including a requirement to review its decision within a fixed time.	CP	●	
"Two buttons" meme				
1	Make technical arrangements to ensure that notice to users refers to the Community Standard enforced by the company. If Facebook determines that (i) the content does not violate the Community Standard notified to user, and (ii) that the content violates a different Community Standard, the user should be properly notified about it and given another opportunity to appeal. They should always have access to the correct information before coming to the Board.	E	●	
2	Include the satire exception, which is currently not communicated to users, in the public language of the Hate Speech Community Standard.	CP	●	
3	Make sure that it has adequate procedures in place to assess satirical content and relevant context properly. This includes providing content moderators with: (i) access to Facebook's local operation teams to gather relevant cultural and background information; and (ii) sufficient time to consult with Facebook's local operation teams and to make the assessment. Facebook should ensure that its policies for content moderators incentivize further investigation or escalation where a content moderator is not sure if a meme is satirical or not.	E	●	
4	Let users indicate in their appeal that their content falls into one of the exceptions to the Hate Speech policy. This includes exceptions for satirical content and where users share hateful content to condemn it or raise awareness.	E	●	
5	Ensure appeals based on policy exceptions are prioritized for human review.	E	●	
Pro Navalny protests				
1	Facebook should amend and redraft the Bullying & Harassment Community Standard to explain the relationship between the Policy Rationale and the "Do nots" as well as the other rules restricting content that follow it.	CP	●	
2	Differentiate between bullying and harassment and provide definitions that distinguish the two acts. Further, the Community Standard should clearly explain to users how bullying and harassment differ from speech that only causes offense and may be protected under international human rights law.	CP	◐	
3	Clearly define its approach to different target user categories and provide illustrative examples of each target category (i.e. who qualifies as a public figure). Format the Community Standard on Bullying and Harassment by user categories currently listed in the policy.	CP	●	
4	Include illustrative examples of violating and non-violating content in the Bullying and Harassment Community Standard to clarify the policy lines drawn and how these distinctions can rest on the identity status of the target.	CP	●	
5	When assessing content including a 'negative character claim' against a private adult, Facebook should amend the Community Standard to require an assessment of the social and political context of the content. Facebook should reconsider the enforcement of this rule in political or public debates where the removal of the content would stifle debate.	CP	✕	
6	Whenever Facebook removes content because of a negative character claim that is only a single word or phrase in a larger post, it should promptly notify the user of that fact, so that the user can repost the material without the negative character claim.	E	●	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Öcalan's isolation				
1	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.	E	●	
2	Evaluate automated moderation processes for enforcement of the Dangerous Individuals and Organizations policy. Where necessary, Facebook should 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.	E	✕	
3	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.	T	◐	
4	Reflect in the Dangerous Individuals and Organizations "policy rationale" that respect for human rights and freedom of expression, in particular open discussion about human rights violations and abuses that relate to terrorism and efforts to counter terrorism, can advance the value of "Safety," and that it is important for the platform to provide a space for these discussions. While "Safety" and "Voice" may sometimes be in tension, the policy rationale should specify in greater detail the "real-world harms" the policy seeks to prevent and disrupt when "Voice" is suppressed.	CP	●	
5	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. This should not be limited to detained individuals. It should include discussion of rights protected by the UN human rights conventions as cited in Facebook's Corporate Human Rights Policy. This should allow, for example, discussions on allegations of torture or cruel, inhuman, or degrading treatment or punishment, violations of the right to a fair trial, as well as extrajudicial, summary, or arbitrary executions, enforced disappearance, extraordinary rendition and revocation of citizenship rendering a person stateless. Calls for accountability for human rights violations and abuses should also be protected. Content that incites acts of violence or recruits people to join or otherwise provide material support to Facebook-designated organizations should be excluded from protection even if the same content also discusses human rights concerns. The user's intent, the broader context in which they post, and how other users understand their post, is key to determining the likelihood of real-world harm that may result from such posts.	CP	●	
6	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.	CP	◐	
7	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. This engagement should also include human rights, civil society, and academic organizations with expert knowledge on those harms, as well as the harms that may result from over-enforcement of the existing policy.	CP	◐	
8	Ensure internal guidance and training is 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.	E	●	
9	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).	E	●	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Öcalan's isolation				
10	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.	CP	●	
11	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.	T	◐	
12	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.	T	◐	
Myanmar bot				
1	Facebook should ensure that its Internal Implementation Standards are available in the language in which content moderators review content. If necessary to prioritize, Facebook should focus first on contexts where the risks to human rights are more severe.	E	◐	
COVID lockdowns in Brazil				
1	Facebook should conduct a proportionality analysis to identify a range of less intrusive measures than removing the content. When necessary, the least intrusive measures should be used where content related to COVID-19 distorts the advice of international health authorities and where a potential for physical harm is identified but is not imminent. Recommended measures include: (a) labeling content to alert users to the disputed nature of a post's content and to provide links to the views of the World Health Organization and national health authorities; (b) introducing friction to posts to prevent interactions or sharing; and (c) down-ranking, to reduce visibility in other users' News Feeds. All these enforcement measures should be clearly communicated to all users, and subject to appeal.	E	✕	
2	Given the context of the COVID-19 pandemic, Facebook should make technical arrangements to prioritize fact-checking of potential health misinformation shared by public authorities which comes to the company's attention, taking into consideration the local context.	E	✕	
3	Facebook should provide more transparency within the False News Community Standard regarding when content is eligible for fact-checking, including whether public institutions' accounts are subject to fact-checking.	CP	●	
Shared Al Jazeera post				
1	Add criteria and illustrative examples to its Dangerous Individuals and Organizations policy to increase understanding of the exceptions for neutral discussion, condemnation and news reporting.	CP	●	
2	Ensure swift translation of updates to the Community Standards into all available languages.	CP	◐	
3	Engage an independent entity not associated with either side of the Israeli-Palestinian conflict to conduct a thorough examination to determine whether Facebook's content moderation in Arabic and Hebrew, including its use of automation, have been applied without bias. This examination should review not only the treatment of Palestinian or pro-Palestinian content, but also content that incites violence against any potential targets, no matter their nationality, ethnicity, religion or belief, or political opinion. The review should look at content posted by Facebook users located in and outside of Israel and the Palestinian Occupied Territories. The report and its conclusions should be made public.	T	●	
4	Formalize a transparent process on how it receives and responds to all government requests for content removal, and ensure that they are included in transparency reporting. The transparency reporting should distinguish government requests that led to removals for violations of the Community Standards from requests that led to removal or geo-blocking for violating local law, in addition to requests that led to no action.	T	●	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Colombia protests				
1	Publish illustrative examples from the list of slurs designated as violating under its Hate Speech Community Standard, including borderline cases with words which may be harmful in some contexts but not others.	CP	●	
2	Link the short explanation of the newsworthiness allowance provided in the introduction to the Community Standards to the more detailed Transparency Center explanation of how this policy applies. The company should supplement this explanation with illustrative examples from a variety of contexts, including reporting on large scale protests.	CP	●	
3	Develop and publicize clear criteria for content reviewers for escalating for additional review public interest content that potentially violates the Community Standards but may be eligible for the newsworthiness allowance. These criteria should cover content depicting large protests on political issues.	E	✕	
4	Notify all users who reported content which was assessed as violating but left on the platform for public interest reasons that the newsworthiness allowance was applied to the post. The notice should link to the Transparency Center explanation of the newsworthiness allowance.	E	●	
South Africa slurs				
1	Notify users of the specific rule within the Hate Speech Community Standard that has been violated in the language in which they use Facebook, as recommended in case decision 2020-003-FB-UA (Armenians in Azerbaijan) and case decision 2021-002-FB-UA (Depiction of Zwarte Piet). In this case, for example, the user should have been notified they violated the slurs prohibition. The Board has noted Facebook's response to Recommendation No. 2 in case decision 2021-002-FB-UA, which describes a new classifier that should be able to notify English-language Facebook users their content has violated the slur rule. The Board looks forward to Facebook providing information that confirms implementation for English-language users and information about the timeframe for implementation for other language users.	E	●	
Wampum belt				
1	Provide users with timely and accurate notice of any company 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. These actions should be taken with respect to all cases that are corrected at the eligibility stage of the Board's process.	E	●	
2	Study the impacts of modified approaches to secondary review on reviewer accuracy and throughput. In particular, the Board requests an evaluation of accuracy rates when content moderators are informed that they are engaged in secondary review, so they know the initial determination was contested. This experiment should ideally include 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 Board transparency report to demonstrate it has complied with this recommendation.	E	●	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

Recommendation identifier	Oversight Board Recommendation	Category	Board's assessment of	
			Meta's response	Implementation
Wampum belt				
3	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 assessment should also specifically investigate 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. The Board understands this analysis likely requires the development of appropriate and accurately labelled samples of relevant content. 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 Board transparency report to demonstrate it has complied with this recommendation.	E	◐	
Ayahuasca brew				
1	The Board reiterates its recommendation from case decision 2020-004-IG-UA and case decision 2021-006-IG-UA that Meta should explain to users that it enforces the Facebook Community Standards on Instagram, with several specific exceptions. The Board notes Meta's response to these recommendations. While Meta may be taking other actions to comply with the recommendations, the Board recommends Meta update the introduction to the Instagram Community Guidelines ("The Short" Community Guidelines) within 90 days to inform users that if content is considered violating on Facebook, it is also considered violating on Instagram, as stated in the company's Transparency Center, with some exceptions.	E	◐	
2	The Board reiterates its recommendation from case decision 2021-005-FB-UA and case decision 2020-005-FB-UA that Meta should explain to users precisely what rule in a content policy they have violated.	E	◐	
3	To respect diverse traditional and religious expressions and practices, the Board recommends that Meta modify the Instagram Community Guidelines and Facebook Regulated Goods Community Standard to allow positive discussion of traditional and religious uses of non-medical drugs where there is historic evidence of such use. The Board also recommends that Meta make public all allowances, including existing allowances.	CP	●	
Alleged crimes in Raya Kobo				
1	Meta should rewrite Meta's value of "Safety" to reflect that online speech may pose risk to the physical security of persons and the right to life, in addition to the risks of intimidation, exclusion and silencing.	CP	●	
2	Facebook's Community Standards should reflect that in the contexts of war and violent conflict, unverified rumors pose higher risk to the rights of life and security of persons. This should be reflected at all levels of the moderation process.	CP	✕	
3	Meta should commission an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumors that heighten the risk of violence in Ethiopia. The assessment should review the success of measures Meta took to prevent the misuse of its products and services in Ethiopia. The assessment should also review the success of measures Meta took to allow for corroborated and public interest reporting on human rights atrocities in Ethiopia. The assessment should review Meta's language capabilities in Ethiopia and if they are adequate to protect the rights of its users. The assessment should cover a period from June 1, 2020, to the present. The company should complete the assessment within six months from the moment it responds to these recommendations. The assessment should be published in full.	T	◐	

CATEGORY: E– Enforcement, T–Transparency, CP–Content Policy

META'S RESPONSE: ● Comprehensive, ◐ Somewhat Comprehensive, ✕ Not Comprehensive

IMPLEMENTATION

- Implementation demonstrated through published information
- Progress reported
- Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation
- Recommendation omitted, declined, or reframed

© 2022 Oversight Board LLC