Oversight Board

ANNUAL REPORT 2022

Contents

Co-Chairs' Foreword




Evelyn Aswad, Catalina Botero-Marino, Michael McConnell, Helle Thorning-Schmidt
CO-CHAIRS OF THE OVERSIGHT BOARD

In 2022, many of our recommendations to Meta became a reality, improving how the company treats people and communities around the world. Our work led Meta to review its content moderation policies, state its rules more clearly, and apply them more consistently. Meta is now telling more users which specific policy area was violated when their posts are removed and is better aligning its content moderation with human rights principles.

In response to our recommendations, Meta introduced a Crisis Policy Protocol to make its responses to crisis situations more consistent, launched a review of its Dangerous Individuals and Organizations policy, and created a new Community Standard on misinformation. In response to a recommendation in our "breast cancer symptoms and nudity" decision, Meta also enhanced its techniques for identifying breast cancer context in content on Instagram, which contributed to thousands of additional posts being sent for human review that would previously have been automatically removed.

In 2022, we made over half of our 91 policy recommendations as part of our first policy advisory opinions, including one on how Meta treats its most powerful users in its cross-check program. In response, Meta committed to extend greater protections to those at particular risk of over-enforcement, including journalists and human rights defenders.

We also protected the voice of users, especially during political and social transformations and crises. For example, in early 2023, as part of a specific board decision, we urged Meta to better protect political speech in Iran, where historic, widespread protests have been violently suppressed. In response, Meta allowed the term "Marg bar Khamenei" (which literally translates as "Death to [Iran's supreme leader] Khamenei") to be shared in the context of ongoing protests in Iran. Meta has also now changed its system of strikes and penalties to be fairer towards users.

Each decision and policy advisory opinion brought further transparency to otherwise frequently opaque content moderation processes, including by revealing the number of newsworthiness exceptions that the company applies in administering its rules. Our policy recommendations trigger public discourse about how digital platforms can approach some of the most complex challenges in content moderation.

To increase our impact, we adopted seven priority areas where we want to work with stakeholders to improve people's experiences online. These are elections and civic space, crisis and conflict situations, gender, hate speech against marginalized groups, government use of Meta's platforms, treating users fairly, and automated enforcement of policies and curation of content. We also prepared to take on a higher caseload and render decisions more quickly in 2023.

In 2022, we also saw a growing recognition of the idea that defining decisions on content moderation should not be made by companies alone. From the outset, the Board was designed to test an independent approach to content moderation, which, if successful, could also be applied to other companies. Independent oversight is about firms opening themselves up and inviting outsiders to challenge how they work. In the last three years, we have acquired a wealth of experience on independent oversight that can help companies make more robust decisions based on respect for freedom of expression and other human rights. As new regulation brings new requirements, there are also specific areas, such as transparency and user notifications, where we believe we can provide part of the solution.

We would like to take this opportunity to thank the Oversight Board Trustees, the Administration staff, and our fellow Board Members for their expertise and support on our journey so far. In particular, we would like to recognize the contribution of Jamal Greene, who stepped down as an Oversight Board Member and Co-Chair in December 2022. Jamal's leadership has been fundamental to our success in holding Meta accountable, and we would like to thank him for all he has done to establish the Board and advance our mission. We would also like to thank the many stakeholders who have submitted public comments, engaged with our work, and helped to make our achievements to date possible.

Given the uncharted path we are walking, the Oversight Board continues to adapt, and find new ways to fulfil our mission. While we have made good progress so far, we are under no illusions about the scale of the challenge ahead. Together, we can help to surmount the pitfalls of social media and help people connect with confidence.

Foreword by the Chair of the Trust

Stephen Neal
CHAIRPERSON OF THE OVERSIGHT BOARD TRUST

In 2022, my first full year as Chair of the Oversight Board Trust, I was hugely impressed with the Board's work. Board Members continued to deliberate the most difficult, significant cases and issue-defining decisions on a range of issues. These included the publication of the Board's first policy advisory opinion, which examined the sharing of private residential information, and the "Russian poem" decision related to the invasion of Ukraine. As Trustees, we helped appoint three new Board Members from Egypt, Mexico, and the United States.

Another crucial aspect of our role is overseeing the Oversight Board Administration, the full-time staff that supports Board Members with their work. In 2022, the Administration completed hiring across all teams and now comprises approximately 80 people based in London, Washington D.C., and San Francisco. We have attracted some excellent new colleagues, many of whom have unique expertise in free speech and human rights. The Administration, like social media itself, is global, with staff members speaking 40 languages between them.

In July 2022, we announced an additional $150 million commitment from Meta, on top of the $130 million announced in 2019 when the Trust was first established. By making an ongoing financial commitment, Meta issued a vote of confidence in the work of the Board and its efforts to apply Facebook and Instagram content standards in a manner that protects freedom of expression and pertinent human rights standards.

In 2023, we will continue to oversee the Board's operations and safeguard the Board's independence, both of which are critical to its success. Through its case decisions and policy advisory opinions, the Board will continue to improve Meta's products and policies, leading to a better experience for those using Facebook and Instagram. Through working with civil society groups, regulators, and other platforms, the Board aims to build its legitimacy. Meta's employees are another crucial constituency with a big say in the company's future, which the Board will look to harness as advocates for its work.

Meta deserves credit for its vision in setting up the Board as a new form of social media governance and for its ongoing commitment to this endeavor. As Trustees, we will support the Board's continued success with Meta, and help the Board share its approach with other companies and partners across the industry.

> "By making this ongoing financial commitment, Meta issued a vote of confidence in the work of the Board."
>
> **Stephen Neal**
> CHAIR OF THE OVERSIGHT BOARD TRUST

Executive Summary

In 2022, the Oversight Board made 91 recommendations to Meta

In response to our recommendations so far, Meta:

Started telling people which specific policy their content violated when removing their content.

Began systematically measuring the transparency of its enforcement messaging to users.

Enhanced how it identifies breast cancer context in content on Instagram, which contributed to thousands of additional posts being sent for human review that would previously have been automatically removed.

Completed global rollout of new messaging telling users whether human or automated review led to their content being removed.

Created a new section in the Community Standards on misinformation.

Introduced a new Crisis Policy Protocol

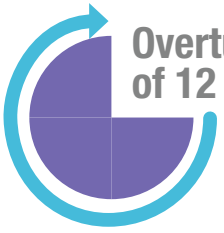
In 2022, the Oversight Board:

Received nearly 1.3m cases from users around the world

Around a quarter more than in 2021

Issued first policy advisory opinions

- Sharing private residential information
- Meta's cross-check program



Overturned Meta in three-quarters of 12 case decisions

Overturning its content moderation decisions 9 times and upholding them 3 times

12 decisions published in 2022

on topics ranging from Russia's invasion of Ukraine to the influence of law enforcement on content removals

Expanded our scope to include the ability to add warning screens to eligible content

Caused Meta to reverse its original decision in 32 cases considered for selection

where its original decision on a post was incorrect

In 2023, we will:

- Publish our first summary decisions on cases where Meta reversed its original decision on a piece of content.
- Issue our first expedited decisions where we publish a decision on a case within days.
- Reach our updated full Board membership goal for maximum efficiency.
- Deepen engagement around our seven strategic priorities.
- Pursue long-term plans for scope expansion.
- Monitor how Meta is implementing our recommendations and push the company to provide evidence of implementation and impact.

We believe in the value of independent oversight and would explore the possibility of new partnerships with companies, and how our work can best complement emerging regulation.

Meet the Board

Oversight Board Members

Afia Asantewaa Asare-Kyei
Director for Accountability & Justice, Open Society Foundations-Africa

Evelyn Aswad
Professor and Chair, University of Oklahoma College of Law

Endy Bayuni
Senior Editor and Board Member, The Jakarta Post

Catalina Botero-Marino
Chairholder, UNESCO Chair on Freedom of Expression, Universidad de Los Andes

Paolo Carozza
Professor, University of Notre Dame

Katherine Chen
Professor, National Chengchi University

Nighat Dad
Founder, Digital Rights Foundation

Tawakkol Karman
Nobel Peace Prize Laureate

Sudhir Krishnaswamy
Vice Chancellor and Professor of Law, National Law School of India University

Ronaldo Lemos
Professor, Rio de Janeiro State University's Law School

Khaled Mansour
Writer

Michael McConnell
Professor and Director of the Constitutional Law Center, Stanford Law School

Suzanne Nossel
Chief Executive Officer, PEN America

Julie Owono
Executive Director, Internet Sans Frontières

Emi Palmor
Advocate and Lecturer, Interdisciplinary Center Herzliya, Israel

Alan Rusbridger
Principal, Lady Margaret Hall Oxford

András Sajó
University Professor, Central European University

John Samples
Vice President, Cato Institute

Pamela San Martín
Former Electoral Councilor at the National Electoral Institute (INE) in Mexico

Nicolas Suzor
Professor, School of Law at Queensland University of Technology

Helle Thorning-Schmidt
Former Prime Minister, Denmark

Kenji Yoshino
Chief Justice Earl Warren Professor of Constitutional Law and Faculty Director of the Meltzer Center for Diversity, Inclusion, and Belonging

Oversight Board Trustees

Kristina Arriaga
Trustee

Cherine Chalaby
Trustee

Stephen Neal
Chairperson of the Trust

Kate O'Regan
Trustee

Robert Post
Trustee

Marie Wieck
Trustee

Oversight Board Administration

Thomas Hughes
Director

Introducing our seven strategic priorities

In October 2022, we announced seven strategic priorities based on an extensive, in-depth analysis of the issues raised by user appeals to the Board. As these priorities are now guiding the cases we select, we encourage users to take them into account when submitting appeals.

1. Elections and civic space

Social media companies face challenges in consistently applying their policies to political expression in many parts of the world, including during elections and large-scale protests. We highlighted the importance of protecting political expression in our "pro-Navalny protests in Russia" decision, while our "mention of the Taliban in news reporting" decision touched upon issues of media freedom. As a Board, we would like to explore Meta's responsibilities in elections, protests, and other key moments for civic participation.

2. Crisis and conflict situations

In times of crisis, such as armed conflict, terrorist attacks, and health emergencies, social media can help people exchange critical information, debate important public issues, and stay safe, but it can also create an environment where misinformation and hatred can spread. Our "alleged crimes in Raya Kobo" and "Tigray Communication Affairs Bureau" decisions examined posts related to the conflict in Ethiopia, while our decision on former President Trump led Meta to adopt a Crisis Policy Protocol. As a Board, we would like to explore Meta's role in protecting freedom of expression in such circumstances, as well as its preparedness for potential harms its products can contribute to during armed conflicts, civil unrest, and other emergencies.

3. Gender

Women, non-binary, and trans people experience obstacles to exercising their rights to freedom of expression on social media. In our "breast cancer symptoms and nudity" decision, for example, Meta's automated systems failed to apply exceptions for breast cancer awareness, which led to important health information being removed from Instagram. Our "gender identity and nudity" decision, which was published in early 2023, also found that Meta's policies on adult nudity result in greater barriers to expression for women, trans, and non-binary people on Facebook and Instagram. As a Board, we would like to explore gendered obstacles women and LGBTQIA+ people face in exercising their rights to freedom of expression, including gender-based violence and harassment, and the effects of gender-based distinctions in content policy.

4. Hate speech against marginalized groups

Hate speech creates an environment of discrimination and hostility towards marginalized groups. It is often context-specific, coded, and with harm resulting from effects which gradually build up over time. Our "depiction of Zwarte Piet" decision found that allowing images of blackface to accumulate online would create a discriminatory environment for Black people, while our "wampum belt" and "reclaiming Arabic words"

decisions examined 'counter speech,' which references hate speech to resist discrimination. As a Board, we would like to explore how Meta should protect members of marginalized groups, while ensuring its enforcement does not incorrectly target those challenging hate. At the same time, we are aware that restrictions on hate speech should not be over-enforced or used to limit the legitimate exercise of freedom of expression, including the expression of unpopular or controversial points of view.

5. Government use of Meta's platforms

Governments use Facebook and Instagram to convey their policies and make requests to Meta to remove content. In response to our "Öcalan's isolation" decision, Meta agreed to provide information on content removed for violating its Community Standards following a report by a government. Our "UK drill music" decision also made proposals for how Meta should respond to requests from national law enforcement. As a Board, we would like to explore how state actors use Meta's platforms, how they might influence content moderation practices and policies (sometimes in non-transparent ways), and the implications of the state's involvement in content moderation.

6. Treating users fairly

When people's content is removed from Facebook and Instagram, they are not always told which rule they have broken. In other instances, users are not treated equally, or they are not given adequate procedural guarantees and access to remedies for mistakes made. As a Board, we would like to explore how Meta can treat its users better, through providing more specific user notifications, ensuring that people can always appeal Meta's decision to the company, and being more transparent in areas such as 'strikes' and cross-check.

7. Automated enforcement of policies and curation of content

While algorithms are crucial to moderating content at scale, there is a lack of transparency and understanding around how Meta's automated systems work and how they affect the content users see. Our "Colombian police cartoon" decision showed how automation can amplify the impact of incorrect content moderation decisions. In response to our "breast cancer symptoms and nudity" decision, Meta has rolled out new messaging globally telling users whether human or automated review led to their content being removed. As a Board, we would like to explore how automated enforcement should be designed and reviewed, the accuracy and limitations of automated systems, and the importance of greater transparency in this area.

WORKING WITH STAKEHOLDERS TO INCREASE OUR IMPACT

As a Board, our achievements so far have been made possible by listening to and collaborating with researchers, civil society groups and others who have worked for many years on the issues we are dealing with. To find practical solutions to our strategic priorities, and the enormously challenging issues they raise, the subject-matter expertise and local knowledge of these stakeholders is essential.

For all strategic priorities, we will continue to work with a broad range of stakeholders who reflect the diversity of the people who use Meta's platforms. This will help us understand the policies and enforcement practices Meta most urgently needs to improve, and what kinds of cases could provide the opportunity to address them. We want to partner with organizations across the world to do this - through our public comments process, roundtables, and individual conversations. To discuss how your organization can get involved, please contact engagement@osbadmin.com.

How the Board Considers User Appeals

This graphic presents the appeals process as it applied to decisions on user appeals in 2022.

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.

ANNOUNCEMENT

A summary of the case is posted on the Oversight Board's website, inviting public comments. The Board will aim to issue a decision within 90 days of when the selection of a case is published.

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.

DECISION

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

PUBLICATION

Our decision is published on the Oversight Board website. Meta has to implement our decision within seven days of publication and respond to any recommendations within 60 days.

APPROVAL

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

IMPLEMENTATION

The Board monitors how Meta is implementing recommendations, providing updates in quarterly transparency reports.

2022 Key Events

FEBRUARY 8

The Board publishes its first policy advisory opinion on the sharing of private residential information, urging Meta to impose tighter restrictions on the sharing of such information.

MAY 11

Meta withdraws its request for a policy advisory opinion related to Russia's invasion of Ukraine, citing security concerns. In its response, the Board notes that Meta's decision "does not diminish Meta's responsibility to carefully consider the ongoing content moderation issues which have arisen from this war."

MAY 19

The Board announces the appointment of three new Board Members from Egypt, Mexico, and the US, bringing the total number of Members to 23.

JUNE 27-30

Board Members meet in person for the first time in California.

JULY 22

The Oversight Board Trust announces a new $150 million financial contribution from Meta.

JULY 26

The Board accepts Meta's request for a policy advisory opinion on removing COVID-19 misinformation.

OCTOBER 20

The Board announces seven strategic priorities focused on areas where it can make the greatest impact on people's experiences of Facebook and Instagram.

OCTOBER 20

The Board gains ability to apply warning screens marking posts as 'disturbing' or 'sensitive' when restoring or leaving up eligible content.

NOVEMBER 22

The Board publishes its "UK drill music" decision, the first time it has examined a post removed after a request from national law enforcement.

DECEMBER 6

The Board publishes policy advisory opinion on Meta's cross-check program. It finds that cross-check is flawed in key areas and makes 32 proposals to Meta.

Recommendations and Impact

91 recommendations made to Meta in 2022

In response to our recommendations so far, Meta:

- Launched **new notifications** globally telling users the specific policy they violated for its Hate Speech, Dangerous Individuals and Organizations, and Bullying and Harassment policies.
- Started **systematically measuring** the transparency of its enforcement messaging to users.
- **Enhanced how it identifies breast cancer context** in content on Instagram, which contributed to thousands of additional posts being sent for human review that would have previously been automatically removed.
- Completed global rollout of user messaging telling people whether **human or automated review** led to their content being removed.
- Launched new notifications telling users when their access to content has been restricted due to local law following a **government request.**
- Created a new section in the **Community Standards** on misinformation.
- Launched a **Crisis Policy Protocol.**
- Started an in-depth review of its Dangerous Individuals and Organizations policy to **prioritize designations based on risk.**

Overview

In our case decisions and policy advisory opinions, we offer specific recommendations for how Meta can improve the policies it applies to the content of billions of users. While our recommendations are non-binding, Meta must respond to them publicly within 60 days. Meta has publicly recognized how our recommendations are changing its behavior. In August 2022, the company stated that the Board "continues to push us to be more thoughtful about the impact of our global content moderation and more equitable in our application of policies and use of resources. Crucially, they also push us to be more transparent, as external voices can help to hold us accountable to our promises."

As a Board, we hold Meta accountable by publishing transparency reports each quarter. These apply a rigorous, independent, data-driven approach to assessing Meta's progress in implementing our recommendations over time. By publicly making these recommendations, and publicly monitoring Meta's responses and implementation, we have opened a space for transparent dialogue with the company that did not previously exist. This kind of openness helps to build legitimacy and trust with users and civil society.

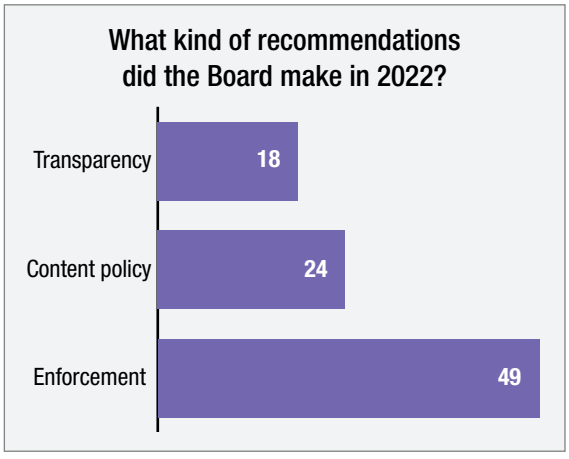
> **"[The Board's recommendations] also push us to be more transparent, as external voices can help to hold us accountable to our promises."**
>
> Meta's Q2 2022 Quarterly Update on the Oversight Board

The role of civil society groups in developing our recommendations also cannot be overstated. In many cases, these organizations submit specific ideas for recommendations as part of our public comments process. In other cases, our proposals echo, or build upon, calls that these groups have been making for many years — forcing Meta to consider and respond publicly to longstanding calls for action. While we explicitly mention these influences in our decision texts, we would like to reiterate our gratitude to these organizations for sharing their ideas and expertise.

RECOMMENDATIONS AND IMPACT IN 2022

The Board made 91 recommendations to Meta in 2022, up slightly from the 86 proposals we made to the company in 2021. By early April 2023, we had made a further 14 recommendations, giving a total of 191 recommendations to Meta.

In total, 41 recommendations fell into the "implementation demonstrated" or "partial implementation demonstrated" categories, and we assessed a further 84 recommendations as "progress reported." In 2022, it was particularly encouraging to see that, in implementing our recommendations, Meta made several changes that had a systemic impact on the company's approach. These included rolling out more specific notifications to users, which we have been calling for since January 2021. As a result of repeated



Board recommendations, Meta also updated how it measures the specificity and transparency of messaging it provides to users when it takes enforcement action against content for violating its policies. This systemic change is part of the company's wider efforts to be more specific with users. Meta is also conducting an in-depth review of the definition of "praise" in relation to its praise, substantive support, and representation (PSR) framework within the Dangerous Individuals and Organizations policy. Meta uses this framework to assess how dangerous individuals and organizations are positively depicted in user content.

LESSONS LEARNED

One area where co-operation with Meta could have been improved in 2022 was data access. In late 2021 and early 2022, we spent eight months attempting to gain access to Meta's CrowdTangle tool to give us more information when selecting cases and assessing recommendation impact. After encountering several roadblocks, we escalated this issue to Meta leadership in early 2022 and were eventually granted access. Meta took several positive steps on data sharing later in 2022, including sharing ongoing research on user appeals and hiring a data scientist to validate the implementation of our recommendations. We look forward to continuing our partnership with Meta's data science teams to obtain meaningful data demonstrating both proof of implementation and impact.

From commitments to action: getting results for users

Given the ambition of our recommendations, and the technical changes they often require, we understand that they take time to implement. In 2022, we saw progress on many of the recommendations we made in 2021, as well as new commitments in response to more recent proposals. It was encouraging to see that, for the first time, Meta enacted systemic changes to what its rules are and how they are enforced, including on user notifications, and its rules on dangerous organizations. The examples below illustrate the impact of our recommendations on how the company treats users and communities around the world.

SYSTEMIC CHANGES TO META'S RULES AND ENFORCEMENT

- As a Board, the recommendation we have made most often is for Meta to **tell people what they have done wrong** when their content is removed. Since we first made this recommendation in January 2021, Meta has gradually been making progress towards this goal. In response to this recommendation, Meta introduced new messaging globally telling users the specific policy they violated for its Hate Speech, Dangerous Individuals and Organizations, and Bullying and Harassment policies. In response to our recommendation, Meta is also now systemically measuring the level of detail of its user communications for all content removals.

RECOMMENDATION
SPOTLIGHT

A new chapter in understanding our impact

In our "breast cancer symptoms and nudity" decision, published in January 2021, we recommended that Meta **"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."** In response, Meta's implementation team enhanced Instagram's techniques for identifying breast cancer context content via text and deployed these in July 2021. These enhancements have been in place since, and in the 30 days between February 26 and March 27, 2023, these enhancements contributed to an additional 2,500 pieces of content being sent for human review that would have previously been removed.

While it is challenging to contextualize 2,500 pieces of content without a denominator, we can see that Meta's implementation of the recommendation followed our framing closely, and that it successfully reduced over-enforcement on the platform. This is a win for independent governance, and the beginning of a new chapter for collaboration between Meta and the Board on understanding our impact on Meta's systems.

- In response to a recommendation from our decision about former President Trump, Meta has made **systemic changes to its response to crises and conflicts.** In August 2022, following consultation with more than 50 global experts, Meta published its Crisis Policy Protocol. This will help provide a more consistent, transparent basis for how Meta responds to crisis situations. On January 25, 2023, Meta noted that it used the Crisis Policy Protocol to evaluate the current environment, including looking at the conduct of the US 2022 midterm elections, ahead of its decision on former President Trump's accounts.

- Many of our proposals have urged Meta to provide users with **far greater transparency** around its rules and exceptions. In response, in August 2022, for the first time, Meta revealed the number of **'newsworthiness allowances,'** it applied to violating content it considered to be in the public interest. From June 1, 2021, through June 1, 2022, Meta documented 68 allowances, of which 13 were issued for posts by politicians. This kind of openness helps to build legitimacy and trust with users and civil society.

- In response to our recommendations on its rules on **dangerous individuals and organizations**, Meta initiated an in-depth review of this policy area. This review is focused on taking a risk-based approach to designating individuals or organizations as dangerous, where the entities assessed as being the highest risk would be prioritized for enforcement. In several of our decisions, we also found that Meta's definition of "praise" in this policy was too limiting of user expression. In response, Meta is reviewing how it assesses whether content amounts to praise, substantive support, or representation of a designated individual or organization.

TREATING USERS FAIRLY

- In 2022, Meta **updated its user notifications for content removed following a government request** based on local law. These tell users that the content has been restricted and explain how Meta processes such requests.

- Meta completed the global rollout of new user messaging telling people whether **human or automated review** led to their content being removed.

- In response to our "claimed COVID cure" decision, **Meta created a new section in the Community Standards on misinformation,** consolidating and clarifying the rules in one place.

RECOMMENDATION
SPOTLIGHT

Independent human rights report on Meta's impact in Israel and Palestine in May 2021

After reading claims in public comments submitted for our 2021 "shared Al Jazeera post" case that Meta had disproportionately removed posts from Palestinians, we urged Meta to engage an independent entity to examine whether its content moderation related to the conflict in Israel and Palestine in May 2021 was biased.

Meta agreed to do so, commissioning the non-profit organization Business for Social Responsibility (BSR) to undertake this review, which it published in September 2022. The report concluded that Meta's content moderation during the May 2021 Israel and Palestine conflict appeared to have had an adverse human rights impact on the rights of Palestinian users to freedom of expression, and on their ability to share information and insights about their experiences as they occurred. Much of the bias identified in the BSR report related to a lack of in-language proficiency and guidance among content moderators at Meta. As a Board, we had urged Meta in three separate decisions to translate its internal guidance for moderators into the language of the content they were reviewing. Meta, however, consistently declined to implement this recommendation on the grounds that English-language guidance is sufficient as its moderators are fluent in English. The BSR report showed that Meta's lack of language capabilities and knowledge of cultural context among moderators led to the over-enforcement of content in Palestinian Arabic, and the under-enforcement of antisemitic content. As we have previously pointed out, English-only guidance may cause reviewers to miss context and nuance across languages and dialects. In addition, because human review data is used to train classifiers, this bias is amplified across Facebook and Instagram.

> "Meta's actions in May 2021 appear to have had an adverse human rights impact… on the rights of Palestinian users to freedom of expression."
>
> Human rights due diligence of Meta's impacts in Israel and Palestine – report by Business for Social Responsibility (BSR)

RECOMMENDATIONS THAT REQUIRE FURTHER ATTENTION FROM META

Meta's implementation of many of our previous recommendations is already improving users' experiences on Facebook and Instagram. However, there are three areas where the company has, so far, not implemented key recommendations.

1. Bring Facebook and Instagram's rules into alignment.
2. Translate internal guidance for moderators into the languages in which they moderate content.
3. Provide more information on how newsworthy posts are escalated within Meta.

Bring Facebook and Instagram's rules into alignment. We first recommended that Meta align Facebook and Instagram's rules in our "breast cancer symptoms and nudity" decision in January 2021. While the company initially committed to this recommendation, it has repeatedly pushed back the deadline for implementation. Meta also has not yet informed users, as per our recommendation, that Meta enforces the Facebook Community Standards on Instagram, and that if content is considered violating on Facebook, it is also considered violating on Instagram.

Translate internal guidance for moderators into the languages in which they moderate content. In three decisions relating to content in Punjabi, Burmese, and Arabic, we recommended that Meta translate its internal guidelines for moderators – the Internal Implementation Standards – into the languages of the content in question. Despite this, and similar concerns being raised by BSR in their recent report, Meta has repeatedly said it will take 'no further action' on this recommendation, as its moderators are all fluent in English.

Provide more information on how newsworthy posts are escalated within Meta. This year, Meta revealed, for the first time, the number of newsworthiness allowances it applied. However, little is known about the process it uses to decide whether content is newsworthy. In our 2021 "Colombia protests" decision, we called on Meta to develop and publicize clear criteria for content reviewers for how to escalate content that violates Meta's rules but could be eligible for the newsworthiness allowance. Meta's response to this recommendation seemed to misunderstand its purpose by focusing on the fact that something *can* be escalated to receive the newsworthiness allowance, rather than what the recommendation requested: a description of *when and why* something might be escalated. As such, we encourage the company to prioritize sharing more information in this area.

REPORT ON THE TIMELINESS OF META'S IMPLEMENTATION OF AND RESPONSE TO OUR RECOMMENDATIONS

- Under our Bylaws, Meta has to respond to our recommendations publicly within 60 days of the Board publishing a decision or policy advisory opinion.
- For the 12 decisions and two policy advisory opinions published in 2022, Meta responded to our recommendations within this timeframe in all cases apart from our cross-check policy advisory opinion. In this case, due to the large number of recommendations we provided (32 in total), Meta responded to our proposals within 90 days, instead of the usual 60.

Impact timeline: tell users what they have done wrong

In response to our proposals, Meta now tells more users globally *which specific policy* their content violated when removing their content.

JANUARY 2021

In three of its first five decisions, Board recommends that "users should always be notified of the reasons for any enforcement of the Community Standards agains them, including the specific rule [it] is enforcing."

FEBRUARY 2021

Meta responds, saying it is assessing the feasibility of this recommendation.

JULY 2021

Meta develops additional classifier that can predict what kind of hate speech is contained in content. People using Facebook in English now receive more specific messaging if content removed for hate speech.

NOVEMBER 2021

Meta begins testing more spedific messaging for hate speech violations on Facebook in non-English langauages, beginning with Arabic, Spanish, and Portuguese. It also starts testing this message on Instagram.

MARCH 2022

Meta begins testing more specific messaging for users on content removed for violating its Bullying and Harassment policy.

MARCH 2022

Meta says that more specific notifications have caused a "statistically significant increase in perceptions of its transparency and legitimacy across multiple markets."

MAY 2022

Meta states that its research has demonstrated that, to ensure people feel heard, its content review process needs to function more like a dialogue that promotes mutal understanding.

AUGUST 2022

Meta launches new messaging explaining to people exactly which policy caused it to take an enforcement action. This is available globally, in English, with translations into other languages underway.

NOVEMBER 2022

Meta says that this new messaging will apply to the vast majority of violation types on Facebook by the end of 2022, and expand to Instagram in 2023. It will also assess feasibility of making notifications even more specific.

Meta's implementation of our recommendations

To ensure that Meta delivers on its commitments, we monitor its progress towards implementing our recommendations. To do this, we look at whether the criteria for a given recommendation have been met. We measure Meta's implementation according to seven categories, updating our assessments on a quarterly basis.

'Implementation demonstrated through published information'
Meta has provided sufficient data for the Board to verify the recommendation has been implemented.

'Partial implementation demonstrated through published information'
Meta has implemented a central component of the recommendation and has provided sufficient data to verify this to the Board.

'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 declined after feasibility assessment'
Meta engaged with the recommendation and then decided to decline its implementation after providing information on the decision.

'Recommendation omitted, declined, or reframed'
Meta will not take any further action on our proposal.

'Awaiting first response from Meta'
The Board has made the recommendation, but Meta has not yet responded publicly (as it has 60 days from publication to respond).

This data-driven approach means that our assessment of whether Meta has implemented a recommendation may differ from the company's reports. We believe, however, that this kind of independent validation is crucial to accountability and to ensuring that users feel the impact of our recommendations.

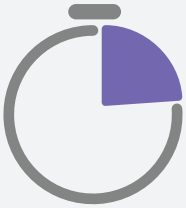
The chart on this page provides a breakdown of our assessment of Meta's implementation of the 191 recommendations we made up to early April 2023, when this report was finalized. We have also published an accompanying document to this Annual Report which sets out our 191 recommendations in full, providing our assessment of Meta's response and implementation. This reflects Meta's quarterly reports on the Board up to Q4 2022.

IMPLEMENTATION CATEGORY	NO. OF RECOMMENDATIONS
 **Implementation demonstrated through published information**	27
 **Partial implementation demonstrated through published information**	14
 **Progress reported**	84
 **Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation**	29
 **Recommendation declined after feasibility assessment**	10
 **Recommendation omitted, declined, or reframed**	23
 **Awaiting first response from Meta**	4
TOTAL NUMBER OF RECOMMENDATIONS	**191**

Case Selection

1,290,942 cases were submitted to the Board in 2022

An increase of around a quarter compared to 2021



ON AVERAGE, the Board received a case every 24 seconds in 2022

MORE THAN 2/3 of user appeals to restore content concerned just two Community Standards

Violence and Incitement

Hate Speech

IN NEARLY 2/3 OF CASES SHORTLISTED BY THE BOARD IN 2022, Meta identified its original decision on the content as incorrect

Overview

With hundreds of millions of posts shared on Facebook and Instagram every day, and tens of thousands of content moderators making split-second decisions on what content stays and goes, one of our biggest challenges is deciding which cases to review. In 2022, we received nearly 1.3 million requests from users to independently review Meta's content moderation decisions. Meta also referred 21 cases to the Board. In 2022 we received, on average, 3,537 cases a day.

> "We try to choose cases that have real impact and touch upon problems the company frequently faces."
>
> **Michael McConnell**
> OVERSIGHT BOARD CO-CHAIR

We aim to focus on cases that pose challenging and consequential questions, reflecting a breadth of geographic regions and subject areas. Through selecting cases that allow us to look at policies that matter most to users, and applying thoughtful, principled review, we can improve policies for all users, and not just those whose cases are heard. Selecting the "Colombian police cartoon" case, for example, gave us the opportunity to make recommendations on Meta's media matching banks, which can automatically remove images that might violate Meta's rules. Similarly, the "UK drill music" case provided an opportunity to make proposals on how Meta should respond to requests from law enforcement agencies around the world.

To address issues that are relevant to people globally, we continued to select cases from different regions around the world. We also chose cases that are critically important to public discourse, such as the "Russian poem" case, which examined questions of expression in the context of Russia's invasion of Ukraine, and the "reclaiming Arabic words" case on the use of derogatory terms affecting LGBTQIA+ people.

LESSONS LEARNED

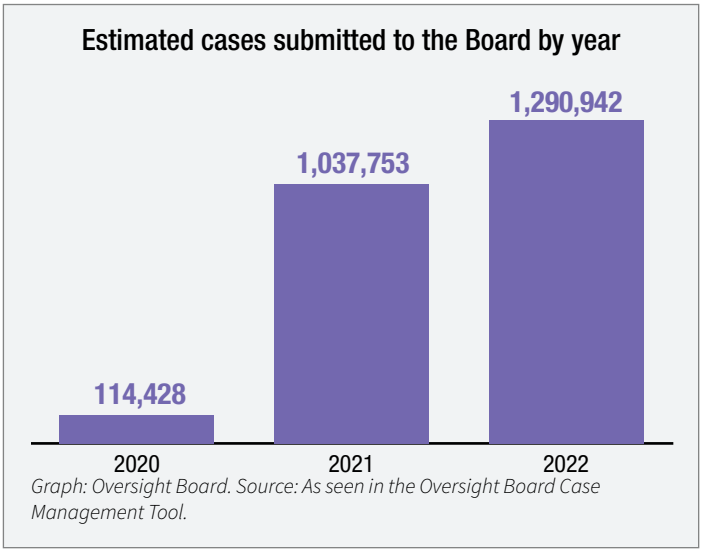
In late 2022, when preparing our "UK drill music" decision and policy advisory opinion on Meta's cross-check program, we identified an issue with the company's appeals process where, for certain escalated appeals, users had not been offered the opportunity to appeal to the Board, despite these cases being eligible for review according to our Bylaws. This raised serious concerns around transparency and users' right to redress. In both the "UK drill music" decision and our policy advisory opinion on Meta's cross-check program, we urged the company to rectify this mistake.

In January 2023, Meta responded, saying that people in the EU, UK, and India would soon be able to appeal eligible content decisions made on escalation to Meta and to the Board. Decisions made "at escalation" are actioned by Meta's internal specialist teams, rather than external contractors. For Facebook and Instagram users in other countries, Meta announced plans to develop an alternate pathway allowing users to appeal eligible escalation takedown decisions that are not internally appealable, directly to us. The company noted that it hoped to implement this solution by the second half of 2023.

Cases submitted to the Board

In 2022, the number of appeals we received from users increased by around a quarter – from around one million in 2021 to around 1.3 million in 2022. While Q1 and Q2 2022 saw higher appeal numbers than the corresponding quarters in 2021, Q3 and Q4 2022 saw fewer appeals than one year previously.

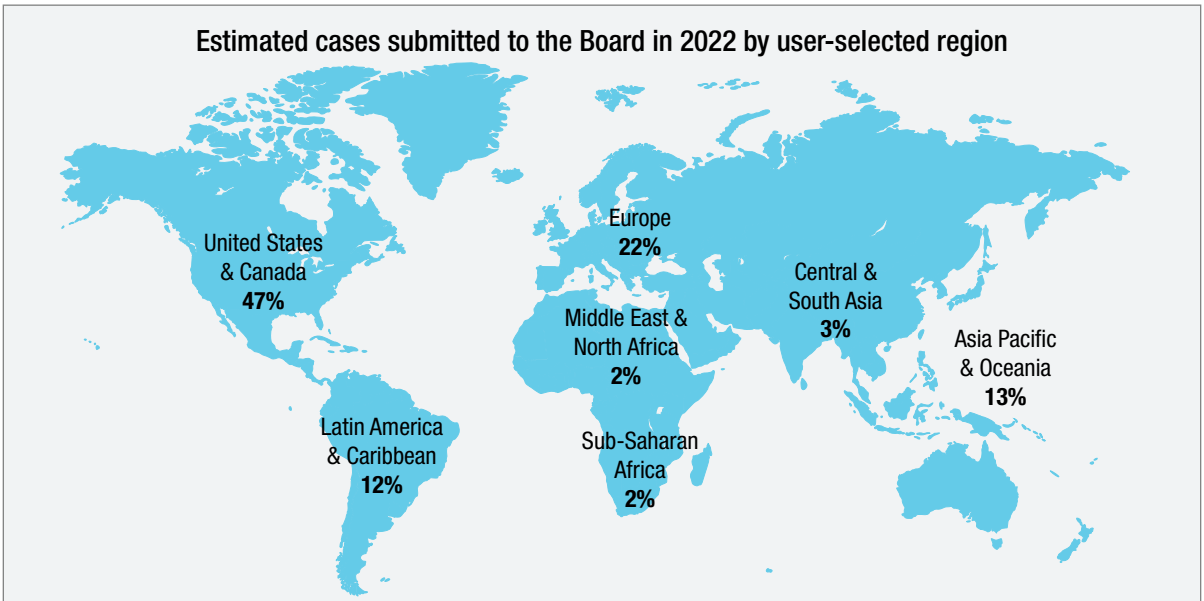
In total, between October 2020, when we started accepting cases, and December 2022, we received more than 2.4 million appeals, reflecting the ongoing demand from users to appeal Meta's content moderation decisions to an independent body. While we can only review a small number of cases, we continue to select cases that raise underlying issues facing large numbers of users around the world and make recommendations to address them.



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

2022 saw a significant increase in the share of cases about content on Instagram. While these constituted just 1% of appeals in every quarter up to Q1 2022, by Q4 2022 this share had grown to more than 10% of appeals.

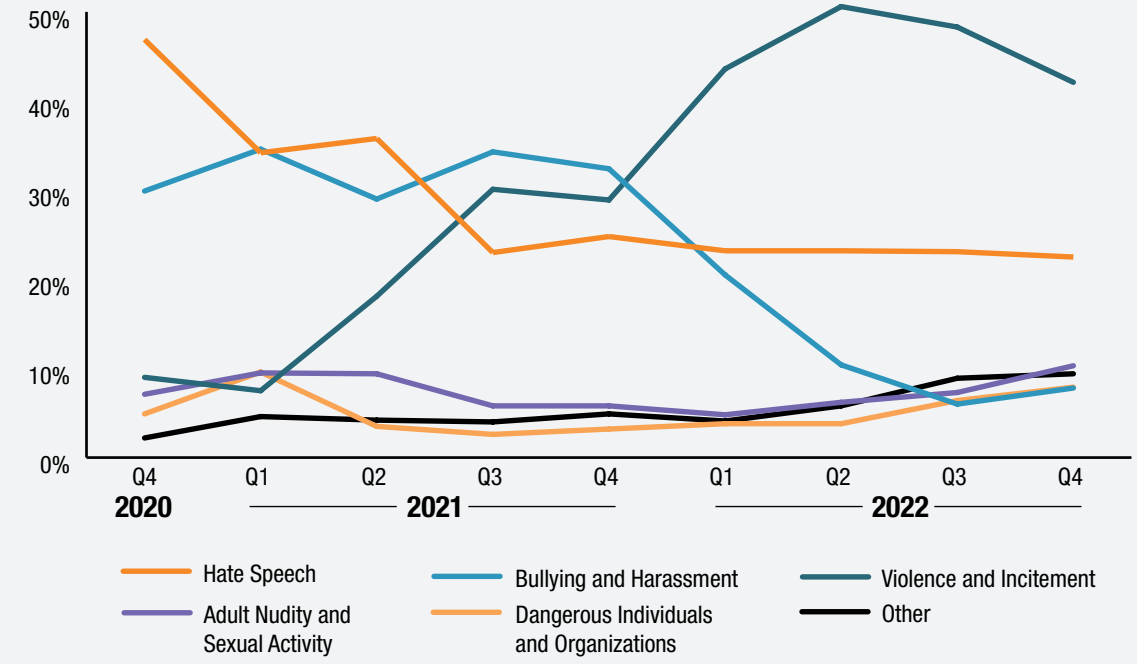
In 2022, the breakdown of cases by user-selected region remained broadly the same as from October 2020 to December 2021, with more than two-thirds of all appeals coming from the US & Canada and Europe. The share of cases coming from the US & Canada fell slightly from 49% in 2020/21 to 47% in 2022. The share of cases from Europe increased from 20% in 2020/21 to 22% in 2022. There was a notable increase in the share of cases from



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

Asia Pacific and Oceania, which increased from 9% in 2020/21 to 13% in 2022. 12% of cases came from Latin America and the Caribbean in 2022, followed by 3% from Central and South Asia, 2% from the Middle East and North Africa, and 2% from Sub-Saharan Africa. We recognize that these figures do not reflect the spread of Facebook and Instagram users worldwide, or the actual distribution of content moderation issues around the world. In 2023, we will be increasing outreach and engagement in regions outside the US and Europe to raise awareness of our work and encourage users in these regions to submit cases to the Board.

In 2022, more than two-thirds (69%) of user appeals to restore content related to just two Community Standards: Violence and Incitement (46% of appeals to restore content) and Hate Speech (23% of appeals to restore content). These figures only concern user appeals to *restore* content that Meta deemed to have broken its rules (which represented 92% of appeals in 2022) and not appeals to *remove* other people's content still live on Facebook or Instagram (which represented 8% of appeals in 2022), because the latter has supposedly not violated a Community Standard.



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

As the graph shows, in 2022 the share of user appeals to restore content removed under Meta's Violence and Incitement Community Standard increased, reaching a high of 51% in Q2 2022. User appeals to restore content removed under its Bullying and Harassment Standard decreased, however, falling from around 30% in Q4 2021 to below 10% in the latter half of 2022.

Meta's Community Standards Enforcement Reports noted a similar increase in the amount of violence and incitement content actioned on Facebook in 2022, rising from 12.7 million in Q4 2021 to 21.7 million in Q1 2022 and 19.3 million in Q2 2022. This may partly explain the increase in appeals to the Board. Meta's report also showed that, in Q3 2022, the amount of bullying and harassment content Meta actioned on Facebook fell to 6.6 million, its lowest level in nearly two years. This may partly explain the reduction in the share of appeals to the Board to restore content removed under this Community Standard.

Cases considered by the Case Selection Committee

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.

Meta deemed its original decision in 32 out of 50 cases shortlisted in 2022 (64%) to have been incorrect, compared to 51 out of 130 shortlisted cases (39%) in 2020/21. This represents an increase of 25 percentage points.

While this is only a small sample, and the Board intentionally seeks out the most challenging and difficult cases, it is concerning that in nearly two-thirds of shortlisted cases, Meta found its original decision to have been incorrect. This raises wider questions both about the accuracy of Meta's content moderation and the appeals process Meta applies before cases reach the Board. In 2023, we will start publishing 'summary decisions.' These will examine cases which we do not select for full review, but which, nonetheless, result in Meta reversing its original enforcement action.

Case Decisions and Policy Advisory Opinions

12 decisions published in 2022

9 of which overturned Meta

7 cases from users
5 cases from Meta

8 cases from Facebook
4 cases from Instagram

Published the Board's first policy advisory opinions

- Sharing private information
- Meta's cross-check program

Gained ability to direct Meta to leave up or restore content with a warning screen

356 questions asked to Meta
as part of our case review in 2022

Of which Meta answered 86% fully

Overview

The decisions we make in response to user appeals and Meta referrals are at the heart of the Board's work. In 2022, we also published our first policy advisory opinions, on sharing private residential information and Meta's cross-check program.

CASE DECISIONS

Our case decisions examine whether Meta's choices to remove or leave up content are in line with the company's rules, values, and stated human rights commitments. Our decisions are binding on Meta, and the company must implement them within seven days of publication, unless doing so would violate the law. In 2022, we published 12 case decisions, overturning Meta's decision on the content in question 75% of the time – up slightly from 70% in 2021.

We also made several changes to our review process to increase our impact. In July 2022, we announced that we would be considering two cases about gender identity and nudity in one decision text for the first time. Considering cases together will help us explore a wider variety of user experiences and compare differences between borderline posts. In October 2022, we also expanded our scope, gaining the ability to make binding decisions to apply a warning screen marking content as "disturbing" or "sensitive" when leaving up or restoring qualifying content. In November, we published our "UK drill music" decision. As part of this decision, we used a freedom of information request to gain new details about how London's Metropolitan Police makes requests to social media companies to remove content.

POLICY ADVISORY OPINIONS

Beyond reviewing individual cases to remove or restore content, we can also accept policy advisory opinions from Meta. Through these, we provide the company with detailed recommendations on changes that Meta should make to its policies on a given topic. We believe that these recommendations may also be helpful to other companies, which often face content moderation issues similar to those faced by Meta.

In 2022, we issued our first policy advisory opinions on sharing private residential information and Meta's cross-check program. Our policy advisory opinion on cross-check shared an unprecedented level of detail on the program, and made 32 recommendations to the company.

HOW META RESPONDED TO OUR QUESTIONS

To assist with our decisions and policy advisory opinions, we send questions to the company ahead of deliberations. For the 12 decisions and two policy advisory opinions we published in 2022, we asked Meta 356 questions. Meta answered 306 of these questions fully, 37 partially, and did not answer 13.

The share of questions that Meta answered fully in 2022 – 86% – was identical to 2021. The share of questions Meta did not answer, however, fell from 6% (19 out of 313 questions) in 2021, to 4% (13 out of 356 questions) in 2022. We believe that Meta's answers in 2022 were more comprehensive than in 2021. When it could not answer a question, Meta explained why this was the case more often than in 2021.

LESSONS LEARNED

In 2022, our Bylaws provided a target timeframe of publishing decisions within 90 days of announcing the case on our website. The Bylaws also provided for exceptions to this timeframe, for example in exceptional circumstances or in the event of technical or operational incidents. In 2022, there were more exceptional extensions than cases delivered on time. In some instances, negotiations with Meta about how much information, which was originally provided by the company on a confidential basis, we could include in our final decision continued for longer than anticipated. In other cases, translation issues delayed publication. On other occasions, Board scheduling constraints and other issues prevented us from publishing within 90 days. It is still our goal to deliver cases within 90 days (as reflected in our February 2023 Bylaws update) and to be transparent around case timelines.

While the Board published two policy advisory opinions in 2022, in May of that year we also announced that Meta had informed the Board that the company would be withdrawing an earlier request for policy guidance concerning content moderation issues related to Russia's ongoing war with Ukraine. In taking this action, the company cited specific ongoing safety and security concerns. While understanding these concerns, we believed that the request raised important issues and were disappointed by Meta's decision to withdraw it. We did, however, later select and publish our "Russian poem" decision, which covered several content moderation issues related to Russia's invasion of Ukraine.

Decisions and policy advisory opinions issued in 2022

CASE DECISIONS

	PLATFORM	SOURCE	COMMUNITY STANDARD	COUNTRIES	BOARD'S DECISION
Asking for Adderall® Case no.:2021-015-FB-UA		User	Restricted Goods and Services	United States	**Overturned** Meta's decision to remove
Swedish journalist reporting sexual violence against minors Case no.: 2021-016-FB-FBR		Meta	Adult Nudity and Sexual Activity	Sweden	**Overturned** Meta's decision to remove
Knin cartoon Case no.: 2022-001-FB-UA		User	Hate Speech	Croatia	**Overturned** Meta's decision to leave up
Sudan graphic video Case no.: 2022-002-FB-MR		Meta	Violent and Graphic Content	Sudan	**Upheld** Meta's decision to leave up
Reclaiming Arabic words Case no.: 2022-003-IG-UA		User	Hate Speech	Morocco, Egypt, Lebanon	**Overturned** Meta's decision to remove
Colombian police cartoon Case no.: 2022-004-FB-UA		User	Dangerous Individuals and Organizations	Colombia	**Overturned** Meta's decision to remove
Mention of the Taliban in news reporting Case no.: 2022-005-FB-UA		User	Dangerous Individuals and Organizations	Afghanistan	**Overturned** Meta's decision to remove
Tigray Communication Affairs Bureau Case no.: 2022-006-FB-MR		Meta	Violence and Incitement	Ethiopia	**Upheld** Meta's decision to remove
UK drill music Case no.: 2022-007-IG-MR		Meta	Violence and Incitement	United Kingdom	**Overturned** Meta's decision to remove
Russian poem Case no.: 2022-008-FB-UA		User	Hate Speech	Latvia, Ukraine, Russia	**Overturned** Meta's decision to remove
Video after Nigeria church attack Case no.: 2022-011-IG-UA		User	Violent and Graphic Content	Nigeria	**Overturned** Meta's decision to remove
India sexual harassment video Case no.: 2022-012-IG-MR		Meta	Adult Sexual Exploitation	India	**Upheld** Meta's decision to leave up

POLICY ADVISORY OPINIONS

Sharing private residential information Case no.: PAO-2021-01	**Meta's cross-check program** Case no.: PAO-2021-02

Summaries of decisions and policy advisory opinions

In October 2022, we announced seven strategic priorities based on an in-depth analysis of the issues and concerns most widely raised by user appeals. This section groups the decisions and policy advisory opinions we published in 2022 according to these priorities.

Elections and civic space

Mention of the Taliban in news reporting

OVERTURNED

In January 2022, a popular Urdu-language newspaper in India posted on its Facebook page an announcement from an official spokesperson for the Taliban regime in Afghanistan, stating that schools and colleges for women and girls would be reopening in two months. Meta removed the post for violating Facebook's Dangerous Individuals and Organizations Standard, which prohibits "praise" of entities, including terrorist organizations, deemed to "engage in serious offline harms." Meta also imposed "strikes" against the page administrator and limited the account's access to some Facebook features.

After the user appealed Meta's decision, a second human reviewer confirmed the company's initial assessment that the post was violating. But after we selected the case for review, Meta conceded that the original decision to remove had been an "enforcement error," restored the content, removed the "strikes," and canceled the account restrictions. Despite Meta's change of course, we overturned Meta's original decision to remove this positive announcement from the Taliban regime. We cited a specific exception to the Dangerous Individuals and Organizations Community Standard permitting the posting of newsworthy reporting on the activities of terrorist groups. We also found that the relationship between the "reporting" allowance in the Dangerous Individuals and Organizations policy and the overarching newsworthiness allowance was unclear.

> "The Board decided that, even though the Taliban is a dangerous organization, news reporting about it should not be restricted because of the public interest in the question."
>
> Paolo Carozza
> OVERSIGHT BOARD MEMBER

Which Community Standards did our decisions examine most in 2022?

Hate Speech
3 decisions

Violence and Incitement
2 decisions

Dangerous Individuals and Organizations
2 decisions

Violent and Graphic Content
2 decisions

Swedish journalist reporting sexual violence against minors — OVERTURNED

In August 2019, a user in Sweden posted a stock photo on their Facebook page of a young girl seated with her head in her hands, obscuring her face. The photo was accompanied by detailed descriptions of the rapes of two unnamed minors. The post gave the ages of two convicted unnamed perpetrators and included graphic details of the harmful impact of the crime on one victim. It also contained quotes, attributed to one perpetrator, bragging about the rape, and referring to one minor victim in sexually explicit terms. In September 2021, after leaving it up on the platform for two years, Meta removed the post for violating the Child Sexual Exploitation, Abuse and Nudity policy.

Our decision held that Meta's decision to remove the post was wrong. Conveying a clinical description of the aftermath of a rape, and the perpetrator's sexually explicit statement about it, did not amount to a message that "sexually exploited children or depicted a minor in a sexualized context." The broader context of the post clearly confirmed that the user's intent had been to report on an issue of obvious public interest while condemning the sexual exploitation of a minor. Among our recommendations in this case, we urged Meta to clearly distinguish between content that endorses or promotes child sexual exploitation and content that raises awareness of it.

Crisis and conflict situations

Tigray Communication Affairs Bureau — UPHELD

In early 2022, Meta referred a case about a post that appeared on the official Facebook page of the Tigray Regional State's Communication Affairs Bureau, an agency operated by the authorities of an Ethiopian province. Posted during the ongoing conflict between federal and Tigrayan forces in the region, the post referred to recent losses sustained by federal forces and urged the national army to "turn its gun" towards the "Abiy Ahmed group," a reference to Ethiopia's Prime Minister. It went on to state that if the federal forces refused to abide by this warning, they would die.

After being reported by several users and identified by Meta's automated systems, two Amharic-speaking reviewers decided that the post did not violate Meta's policies. Two days later, after being escalated for expert review to an Integrity Product Operations Center (IPOC) set up by the company to moderate content emerging from the conflict, Meta reversed its original decision and removed the post for violating Facebook's Violence and Incitement policy.

While acknowledging that Meta has taken positive steps - including the establishment of the IPOC system - to monitor content for abuses in high-risk conflict situations, our decision highlighted that the violating content had remained on the platform for two days and been viewed more than 300,000 times before being removed. At a minimum, that alone underscored the fact that IPOCs are "not intended to be a sustainable, long-term solution to dealing with a years-long conflict." In response, we recommended that "Meta may need to invest in a more sustained mechanism" to fulfill its human rights responsibilities in conflict zones and crisis situations.

Sudan graphic video

UPHELD

In the final weeks of 2021, Meta referred a case about a graphic video posted to a Facebook profile page after the October 2021 military coup in Sudan. The video showed a person lying beside a car with a significant head wound and a visibly detached eye. In the background, voices could be heard saying in Arabic that someone had been beaten and left in the street. A caption, also in Arabic, called on people to not trust the military, citing a longstanding pattern of abuses.

After Meta removed the post for violating Facebook's Violent and Graphic Content Community Standard, the company issued a "newsworthiness allowance." Following a delay of nearly five weeks, Meta restored the content with a warning screen that restricted its viewership.

Our decision upheld Meta's decision to restore the content with a warning screen. We also held that applying a "newsworthiness allowance" was not an effective means of moderating graphic content shared on Facebook at scale. Underscoring that conclusion, we cited Meta's own admission that it had issued 17 newsworthy allowances in connection with the Violent Graphic Content policy during the first three quarters of 2021. By contrast, it had removed more than 90 million pieces of content for violating the same Standard over the same period. As such, we recommended that Meta revise its Violent and Graphic Content Community Standard to permit the sharing of graphic videos when "intended…to raise awareness or document abuses."

Russian poem

OVERTURNED

In April 2022, a Facebook user in Latvia posted an image of a dead body lying in a street, which Meta confirmed was of a person shot in Ukraine. An accompanying text in Russian argued that during the Second World War, alleged atrocities committed by Soviet soldiers in Germany had been excused as vengeance for crimes allegedly committed by Nazi soldiers in the USSR. Citing contemporary alleged atrocities committed by Russian soldiers in Ukraine, it quoted from a poem by Soviet poet Konstantin Simonov: "Kill the fascist... Kill him!"

After Meta removed the post for violating its Hate Speech Community Standard, we selected the case for review. Meta then restored it with a warning screen. In our decision, we drew a distinction between content that might have targeted Russian soldiers'

nationality, which would have been violating, and content that draws a *historical parallel* to actions taken by Nazis, which is not. Quoting lines from the Soviet-era poem, we found, was an artistic and cultural reference employed to *describe*, not *encourage*, a state of mind. While acknowledging the complexities of evaluating violent speech in conflict situations where international law permits combatants to be targeted, we recommended that Meta revise its policies to clearly account for a context of "unlawful military interventions."

Hate speech against marginalized groups

Knin cartoon

OVERTURNED

In December 2021, a public Facebook page posted a video based on the Disney cartoon "The Pied Piper" containing a caption, in Croatian, which Meta translated as "The Player from Čavoglave" – a village in Croatia – "and the Rats from Knin," a city in Croatia. The narrator stated that when a population of rats decided they wanted to live in a "pure rat country," they began harassing and persecuting people, permitting them to take over. But after a piper from Čavoglave appeared and played a melody on his "magic flute," he lured the rats out of the city and into a tractor, which disappeared. The narrator concluded: "the rats disappeared forever from these lands…[and] everyone lived happily ever after."

Even after users reported the content nearly 400 times, Meta did not remove it. After the case was appealed to the Board, Meta conducted an additional human review, which also found that the content was not violating. Once we selected the case for review, Meta reversed that decision, determining that while the post did not violate the *letter* it did violate the *spirit* of the Hate Speech policy, and removed it. While drafting an explanation for the Board, Meta changed its mind again, this time determining that the post did, in fact, violate the *letter* of the policy.

In finding that the content did violate the Hate Speech and Violence and Incitement Community Standards, we cited comments containing references to a 1995 Croatian military operation, "Operation Storm." This operation led to the forcible displacement, execution, and disappearance of ethnic Serb civilians in Croatia. In the context of those references, we expressed concern that nearly 40 Croatian-speaking moderators decided the content was not violating because they thought the standard required an *explicit* derogatory comparison between ethnic Serbs and rats. We recommended that Meta clarify the Hate Speech Community Standard to clearly convey that the policy prohibits *implicit* as well as *explicit* hostile references to protected groups.

Reclaiming Arabic words

OVERTURNED

In November 2021, a public Instagram account described as a space for "discussing queer narratives in Arabic culture," posted images with a caption in Arabic explaining that each picture contained a word used in the Arabic-speaking world to denigrate men with "effeminate mannerisms." The user stated that their intention was "to reclaim [the] power of such hurtful terms."

After Meta removed the content for violating its Hate Speech policy, restored it after the user appealed, and removed it again after another user reported it, we selected the case. Meta escalated the content for additional review, which concluded that the content did not violate the policy. Meta then restored it, stating that its initial decisions had been based solely on reviews of the pictures that contained the derogatory terms, not the content itself or explanation of intent.

> "Meta has values and standards, but where it is lacking is enforcement."
>
> **Endy Bayuni**
> OVERSIGHT BOARD MEMBER

In finding that the content's removal did not align with the Hate Speech policy, we noted that the content is covered by an exception permitting the sharing of speech "used self-referentially or in an empowering way." Quoting hate speech with the intent of condemning it or raising awareness" is permitted. Meta's back-and-forth decision-making in this case reflected, we held, an inconsistent application of "the exemptions in the Hate Speech policy to expression from marginalized groups." To address this issue, we recommended that Meta translate the "Internal Implementation Standards" and "Known Questions" it gives its moderators into Modern Standard Arabic.

Government use of Meta's platforms

UK drill music

OVERTURNED

In January 2022, an Instagram account described as promoting British music posted a clip from a recently released music video of a "UK drill music" track by the rapper Chinx (OS) called "Secrets Not Safe." UK drill music is a localized, grassroots, subgenre of rap, popular among urban young Black people in the UK. Shortly after, the Metropolitan Police, a law enforcement agency with responsibilities in Greater London, emailed Meta a request for the company to review all content associated with "Secrets Not Safe." The police sent additional information related to the prevalence of gang violence in London, conveying concern that the track might increase the risk of retaliatory gang violence.

In response to this request, a Meta specialist team, relying on the context provided by the police, determined that the post constituted a "veiled threat." The company based that conclusion on a reference to a 2017 shooting, which in its view raised the potential of the track to incite violence. Meta not only removed the content from the account for violating its Violence and Incitement policy, but also removed more than 50 additional pieces of

content containing the track from other accounts, including the artist's. Meta's automated systems later removed content with the track another 112 times.

After Meta referred the case to the Board, we asked to review Chinx (OS)'s original post. Meta responded that the actions it took to remove the video and song from across the platform had caused the artist's account to be deleted. In considering whether Meta's actions aligned with its standards, values and human rights responsibilities, the context of the post was key:

> "[Rap] artists often speak in granular detail about ongoing violent street conflicts, using a first-person narrative with imagery and lyrics that depict or describe violent acts. Potential claims of violence and performative bravado are considered to be part of the genre – a form of artistic expression where fact and fiction can blur.
>
> Through these claims, artists compete for relevance and popularity. Whether drill music causes real-world violence or not is disputed, particularly the reliability of evidential claims made in the debate."

In assessing whether Meta's decision aligned with its rules, values, and human rights responsibilities, we cited a lack of evidence to support the idea that the content constituted a "credible threat." We further found that "in the absence of such evidence, Meta should have given more weight to the content's artistic nature." Our decision noted that "while law enforcement can sometimes provide context and expertise, not every piece of content that law enforcement would prefer to have taken down should be taken down."

When evaluating such requests in the future, we urged Meta to take tangible steps to "evaluate these requests independently, particularly when they relate to artistic expression from individuals in minority or marginalized groups for whom the risk of cultural bias against their content is acute."

Our decision also highlighted a lack of transparency about the processes Meta uses to respond to requests from government and law enforcement agencies. We cited information we received in response to a Freedom of Information request to the Metropolitan Police, revealing that all 286 requests that the police had made to social media companies and streaming services to review or remove musical content in the year to May 2022 involved UK drill music. On January 4, 2023, however, the Metropolitan Police contacted us to say that it had identified errors in its response and corrected them. The Metropolitan Police had actually made 992, not 286, such requests in the year to May 2022, all involving drill music.

Treating users fairly

Policy advisory opinion on sharing private residential information

In February 2022, we published our first policy advisory opinion (PAO). PAOs are used by the Board to review Meta's policies and make recommendations for how they can be improved. In this case, Meta asked the Board for a detailed assessment of changes it might make to Facebook's Privacy Violations Community Standard, which prohibits the sharing of "personally identifiable information about yourself or others" except under specific conditions. Under the terms of the policy, "personally identifiable information," including residential addresses, may not be shared except in cases when the information is "shared or solicited to promote charitable causes, find missing people, animals, or objects, or contact business service providers."

To protect users' privacy, the policy prohibits the sharing of "imagery that displays the external view of private residences," particularly if the residence is a single-family home, its unit number is identifiable in the image/caption, the resident "objects to the exposure of their private residence," or if "there is a context of organizing protests against the resident." The one exception is if the residence serves as a government embassy, where it may be a focus of protest.

An underlying issue Meta asked us to address is the fact that having access to such information can be useful to journalists and civic activists. Conversely, the exposure of such information without residents' consent could "create a risk to residents' safety and infringe on an individual's privacy." Among the potential real-world harms to users' safety and privacy Meta cited was "doxing," a term for the unauthorized release of documents (abbreviated as "dox") with the intention of revealing personal information to people who may abuse it. As Meta noted, doxing has negative real-world consequences, including harassment, stalking, violence, and death.

The policy contains exceptions under which private residential information may be publicly posted, including when it is already "publicly available through news coverage, court filings, press releases, or other sources." In its internal guidance to content reviewers, Meta states that information previously published by "at least five news outlets" is no longer considered private under the policy.

We agreed with Meta's assertion that serious real-world harms may result from violations of the right to residential privacy, and that such harms disproportionately affect women, children and LGBTQIA+ people. We also agreed with Meta that while physically accessing public records and other sources of "publicly available" information requires resources and effort, gaining access to such information is much easier on digital platforms and is more easily shared, on a larger scale. In view of the gravity and severity of such harms, we advised Meta to limit the circumstances in which sharing "publicly available" information is allowed.

Policy advisory opinion on Meta's cross-check program

In October 2021, in the wake of disclosures by the *Wall Street Journal* about Meta's cross-check program, we accepted a request from the company to review the program and make recommendations for how it might be improved. In the context of that request, Meta shared that it was performing about 100 million content enforcement attempts every day. At that rate, even if 99% of its moderation decisions were accurate, the company would still find itself making around one million mistakes a day.

The sheer volume and complexity of content posted on Facebook and Instagram pose real challenges for building systems that fulfill Meta's human rights commitments. While recognizing this, we noted that, as currently structured, cross-check did not comply with the company's human rights commitments. For years, cross-check allowed content from a select group of politicians, business partners, celebrities, and others to remain on Facebook and Instagram for several days when it would have otherwise been removed quickly. Meta told us that, on average, it can take more than five days to reach a decision on content from users on its cross-check lists.

Meta, for its part, maintained that the original goal of the program had been to "advance its human rights commitments." Upon closer scrutiny, we found that the program appeared "more directly structured to satisfy business concerns." Among the most prominent flaws our review surfaced, which we urged the company to address, was that by granting selected users greater protection from content removal than others, it amplified widespread concerns that the company was privileging celebrities, and political and business leaders, over millions of others.

If a post from a user on Meta's cross-check lists is identified as violating the company's rules, it remains on the platform pending further review. Meta then applies its full range of policies, including exceptions and context-specific provisions, to the post, likely increasing its chances of remaining on the platform. Ordinary users, by contrast, are much less likely to have their content reach reviewers who can apply the full range of Meta's rules.

This unequal treatment was amplified by a lack of transparency around the criteria used to compile the list of entities whose content benefits from additional protection. While Meta has clear criteria for adding business partners and government leaders to cross-check lists, groups whose posts are important from a human rights perspective, such as journalists or civil society leaders, have less clear paths to the program.

This lack of transparency extends to Meta's failure to disclose the core metrics it uses to measure the program's effectiveness. Meta provided no evidence that the company compares the accuracy of decisions it makes under cross-check to those made by its normal quality control mechanisms. It extends to Meta's continued failure to inform users if they are on cross-check lists, failure to disclose the procedures for creating and auditing those lists, and failure to disclose cases when "entities that continuously post violating content are kept on…lists based on their profile."

We made 32 recommendations to address these issues. These urged Meta to prioritize expression important for human rights in its cross-check lists, including expressions of "special public importance." We also called on Meta to radically increase transparency about the program and how it operates, and made proposals to reduce the harm caused by violating content left up during "enhanced review." Meta responded to our recommendations in March 2023, committing to improve transparency and extend greater protections to those at particular risk of over-enforcement, including journalists and human rights defenders.

Asking for Adderall®

OVERTURNED

In June 2021, a Facebook user in the United States who identified as having attention deficit hyperactivity disorder (ADHD), asked in a post in a private group how to talk to a doctor about medication. The user stated that they had been given a Xanax prescription, but that Adderall had worked well for them in the past. The user expressed concern that if they asked their doctor for a different prescription, they might present as someone with "drug-seeking behavior."

In August 2021, Meta removed the post citing its alleged violation of Facebook's Restricted Goods and Services Community Standard. After we selected the case, Meta flagged the removal as an "enforcement error" and restored it.

In our decision, we found that Meta's initial decision to remove the post was wrong. Facebook's Restricted Goods and Services Community Standard does not prohibit users from seeking advice on specific pharmaceutical drugs in the context of "medical conditions." We also found that the public definitions of substances under the Standard are opaque to users because even under its internal definitions, Adderall and Xanax might or might not fall under "non-medical" or pharmaceutical drugs depending on the circumstances. In our recommendations, we urged the company to "review user appeals in a timely fashion when content-level enforcement measures trigger account-level penalties."

Automated enforcement of policies and curation of content

Colombia police cartoon

OVERTURNED

In September 2020, a Facebook user in Colombia posted a cartoon image resembling the official crest of the National Police of Colombia. It showed three figures in police uniforms holding batons over their heads, apparently kicking and beating another figure lying on the ground with blood under their head. In Spanish, the text of the crest read, in Meta's translation: "National Police – Republic of Colombia – Baton and Kick."

Sixteen months after the content was posted, Meta removed it after matching the image with a similar one in one of its Media Matching Service banks. These banks can automatically identify and remove images that have been identified by human reviewers as violating the company's rules. After we selected the case, Meta determined that the post did not violate its rules and restored it. The company also restored other content featuring the cartoon, which it acknowledged had been incorrectly removed by its Media Matching Service banks.

> "Content decisions and the algorithmic treatment of content cannot be separated."
>
> Catalina Botero-Marino
> OVERSIGHT BOARD CO-CHAIR

This case highlighted how such banks can amplify the impact of incorrect decisions to bank pieces of content. Despite 215 users appealing these removals, 98% of which succeeded, Meta did not remove the cartoon from its bank, and restore the comparable content, until we took up the case. We urged Meta to "urgently improve its procedures to quickly remove non-violating content from these banks."

Video after Nigeria church attack

OVERTURNED

On June 5, 2022, terrorists attacked a Catholic church in southwestern Nigeria. Hours later, an Instagram user in Nigeria posted a video showing motionless, bloodied bodies on the church floor, some with their faces visible. After being identified by one of Meta's Media Matching Service banks as resembling a video determined by a human reviewer to be violating, an automated content "classifier" decided that it could stay up, with a "disturbing content" warning screen.

Several days later, while the content was being reviewed by a second media matching bank, the user added a caption in English describing the attack as "sad." The user also added hashtags, some of which referenced the live-action video game "airsoft," and others commonly used to market firearms to collectors. Meta determined that while the video itself was not violating, the hashtags were, because they could be interpreted as "glorifying violence and minimizing the suffering of the victims." A majority of the Board disagreed with Meta's decision, finding that the original restoration of the content with a warning screen had been correct. The majority based its conclusion on a determination that the hashtags should not be classified as sadistic "merely because they are associated with users of firearms." A minority disagreed with this assessment, noting that the use of

shooting-related hashtags could be read as "sadistic, and could traumatize survivors or victims' families." Noting this ambiguity, we recommended that Meta review the language in the public Violent and Graphic Content policy.

Gender

India Sexual Harassment Video

UPHELD

In March 2022, an Instagram account described as a "platform for Dalit perspectives" posted a video from India showing a person identified as a "tribal woman" being assaulted by a group of men. "Dalit" people, previously known as "untouchables," face oppression under the country's caste system. Despite an absence of nudity and the fact that the woman in question was not identifiable, Meta removed it for violating the Adult Sexual Exploitation policy. Later, Meta's internal teams reversed that decision and restored the content after applying a "newsworthiness allowance" and a warning screen, in view of its intent to raise awareness of longstanding discrimination against "tribal women" in India.

While we found that Meta's decision to restore the content after applying a newsworthy allowance and then a warning screen was correct, we repeated our concern that applying the newsworthiness allowance to content that is otherwise violating is not the right way of "dealing with such cases at scale."

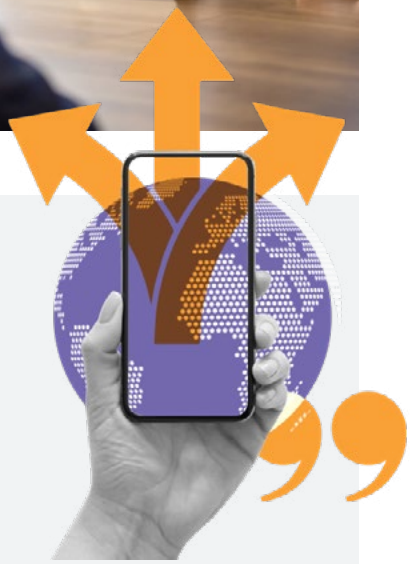
> "The newsworthiness allowance is vague, leaves considerable discretion to whoever applies it, and cannot ensure consistent application at scale. Nor does it include clear criteria to assess the potential harm caused by content that violates the Adult Sexual Exploitation policy. The Board finds that Meta's human rights responsibilities require it to provide clearer standards and more effective enforcement processes for cases such as this one."

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 12 decisions we published in 2022, Meta restored or removed the content within this seven-day timeframe, except in cases where the content had already been restored.
- Through our Implementation Committee, currently made up of five Board Members, we continue to urge Meta to provide greater transparency about how it is 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.

Applying international human rights standards to content moderation: Article 19

A defining theme of the Board's work is our conviction that Meta will make content moderation decisions in a fairer, more principled way if it bases them on the international human rights standards to which it has committed itself. To that end, our 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." Article 19 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?

We look at 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, and was the restriction proportionate to the interests being protected?

> "Companies must respect the right to freedom of expression, which includes imparting and receiving ideas of all kinds, regardless of frontiers."

Evelyn Aswad
OVERSIGHT BOARD CO-CHAIR

International human rights norms in the Board's decision-making process

In 2021, Meta published a Corporate Human Rights Policy that described its human rights commitments as rooted in the United Nations Guiding Principles on Business and Human Rights (UNGPs). This 2011 framework provided new guidance to private companies on their responsibilities to respect human rights.

In its first Human Rights Report,[1] published in 2022, Meta framed its decision to establish the Board as part of its efforts to provide access to remedy for human rights impacts. The UNGPs call on businesses to provide "access to remedy," including by establishing "effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted [by their operations]."

As an operational-level grievance mechanism, the Board seeks to embody the effectiveness criteria set out in the UNGPs, including by being legitimate, accessible, equitable, transparent, rights-compatible, and a source of continual learning. We choose cases, issue decisions, and provide recommendations to advance Meta's respect for the human rights of all people. In every decision we provide a detailed analysis of the human rights implications and concerns animating the case.

THE BOARD'S SUBMISSIONS TO THE UNITED NATIONS

Drawing on our analysis of Meta's human rights responsibilities in cases and policy advisory opinions, we provided extensive submissions to two United Nations bodies in 2022 describing the evolution of our approach to applying human rights principles to digital platforms' content policies and decisions.

> **"We constantly remind Meta that human rights principles should be at the center of their Community Standards."**
>
> Nighat Dad
> OVERSIGHT BOARD MEMBER

In our February submission to the UN Office of the High Commissioner for Human Rights on "the practical application of the UNGPs to the activities of technology companies"[2] and our July submission to the UN Special Rapporteur on freedom of opinion and expression,[3] we focused on the twin themes of "accountability" and "remedy" as set forth in the UNGPs.

Both submissions highlighted the grave human rights implications of social media decisions on content during conflict. This focus reflects the importance we ascribed to the UNGPs for companies to "increase efforts to minimize human rights risks where the consequences are 'most severe.'" In line with the UNGPs, for every case assigned to a panel we conduct a detailed Human Rights Impact Assessment (HRIA), which may include a "conflict sensitivity analysis" to ensure the Board is itself meeting its due diligence responsibilities. Moreover, our Case Selection Team has processes in place to alert Meta to "left up" content that may require "expeditious attention" to mitigate the risk of severe harm.

1 Meta Human Rights Report: Insights and Actions 2020 – 2021 (July 2022)," https://about.fb.com/wp-content/uploads/2022/07/Meta_Human-Rights-Report-July-2022.pdf

2 "Operationalizing the UN Guiding Principles on Business and Human Rights Submission to the Office of the High Commissioner for Human Rights, United Nations on the practical application of the UNGPs to the activities of technology companies" The Oversight Board February 2022

3 "Submission to the Special Rapporteur on freedom of opinion and expression: Challenges in Times of Conflicts and Disturbances," The Oversight Board July 2022.

CASE STUDIES FROM AFGHANISTAN, ETHIOPIA, AND SUDAN

In the submission to the UN Special Rapporteur, we cited three illustrative 2022 cases: "Sudan graphic video," "mention of the Taliban in news reporting," and "Tigray Communication Affairs Bureau." All reflected Meta's lack of progress in developing a "principled, transparent system for moderating content in conflict zones." Formulating an effective crisis response protocol, we contended, would significantly improve the company's ability to uphold its human rights responsibilities during crisis and conflict situations.

In our **"Sudan graphic video"** decision, we underscored the context of crisis and conflict at the time the content was posted: "Security forces…targeted journalists and activists, searching their homes and offices. Journalists [were] attacked, arrested, and detained…With the military takeover of state media and crackdown on Sudanese papers and broadcasters, social media became a crucial source of information and venue to document the violence carried out by the military. The military shut down the internet…with the arrest of civilian leadership."

Applying the Article 19 three-part test, we questioned Facebook's Violent and Graphic Content policy for failing to clarify how it permits users to "share graphic content to raise awareness of or document abuses." While Meta correctly applied its "newsworthiness allowance" as the primary rationale for restoring the content in question, we pointed out that the rule fails to clearly define the relevant term. As for the remedy of placing a warning screen on the content, we held that the placement of a warning screen on the content constituted a "necessary and proportionate restriction on freedom of expression," which "adequately protect[ed] the dignity of the individual depicted and their family."

> If freedom of expression is going to be suppressed, it must be clear why."
>
> Julie Owono
> OVERSIGHT BOARD MEMBER

In a second case, **"mention of the Taliban in news reporting,"** concerning the removal of a post relaying an announcement by an official Taliban spokesperson in Afghanistan of the imminent opening of schools for girls, we found Meta's initial decision to remove it violated users' freedom of expression. This was because it denied Facebook users in Afghanistan their right to "access information about events of public interest…especially when a designated dangerous group forcibly removed the recognized government." The constraints on media freedom imposed by the Taliban regime rendered "the role of international reporting even more important," we added, because "the information…was essential to people concerned about girls' and women's equal right to education."

A third case, **"Tigray Communication Affairs Bureau,"** concerning the removal of content containing a call from an official source within the Tigrayan provincial government for government forces to turn on central government forces, found Meta's decision to take down the post to be consistent with its human rights responsibilities. To support that decision, we applied the six-factor test from the Rabat Plan of Action, which provides specific guidance on how to protect freedom of expression while also protecting people from incitement of discrimination, hostility or violence.

The six Rabat factors were assessed as follows: (1) *Context*: The content was posted in the context of an ongoing and escalating civil war. (2) *Speaker:* The speaker was a regional government ministry affiliated with one of the parties to the conflict. (3) *Intent:* An explicit call to kill soldiers who did not surrender. (4) *Content:* The post could be read to advocate targeting combatants and political leaders, regardless of their participation in the hostilities. (5) *Extent of dissemination:* The content was posted on the public page of a body connected to one of the parties to the conflict with about 260,000 followers and remained on the platform for two days before being removed. (6) *Likelihood and Imminence:* The content was posted as Tigrayan forces were advancing beyond Tigray and the Prime Minister was declaring a nationwide state of emergency and calling on civilians to take up arms and fight.

We also employed the Rabat test to support our decision in our 2022 "Knin Cartoon" case, which concerned the posting of a cartoon that characterized ethnic Serbs in a city in Croatia as rats. (1) *Context*: A region that recently experienced ethnic conflict and discrimination against ethnic minorities. (2) *Speaker:* A Croatian news portal known for anti-Serb sentiments. (3) *Intent:* To incite ethnic hatred. (4) *Content:* The cartoon video form can be particularly harmful because it is engaging. (5) *Extent of dissemination:* The content was viewed over 380,000 times, shared over 540 times, received over 2,400 reactions and had over 1,200 comments. (6) *Likelihood and Imminence:* The Board did not believe the post was likely to result in imminent harm. Nevertheless, we noted the decision that Meta "can legitimately remove posts from Facebook that encourage violence in a less immediate way."

Engagement and Public Comments

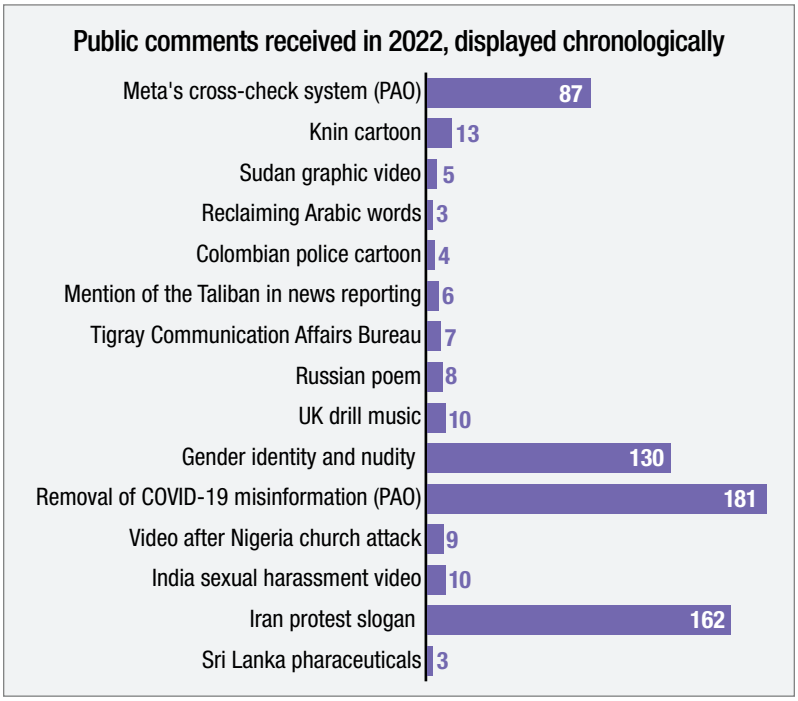
As a Board, we know we can only find lasting solutions to the challenges of content moderation if we listen to, and work with, other organizations. With this aim, our Board Members and staff take part in public events and debates, explaining what the Board does, and discussing why a principled, global approach to content moderation matters to users. As part of our decisions process, individuals and organizations can also submit public comments. These have given people a voice in our decisions, and provided expertise on language, culture, politics, and human rights. On many occasions, public comments have also shaped our decisions and recommendations to Meta.

OUTREACH AND ENGAGEMENT

In 2022, more than 1,000 stakeholders attended roundtables we hosted about issues raised in specific cases. Topics included our policy advisory opinion on cross-check and Meta's policies on COVID misinformation. We also circulated a newsletter, "Across the Board," to better inform the public about the Board's activities. In 2023, we will continue this crucial work, convening at least one roundtable a month in locations across the globe focused on our seven strategic priorities. By focusing our engagement work around our seven priorities, we also hope to benefit from our networks and the broader academic and advocacy communities that specialize in these areas.

PUBLIC COMMENTS

To enrich our decisions and policy advisory opinions, we carefully consider public comments submitted by individuals and organizations. In 2022, we received over 600 public comments, with an increase in the number of public comments submitted in the latter half of the year.



Our **"Tigray Communication Affairs Bureau"** decision was informed by valuable insights from experts on moderating content in conflict zones. One conclusion, that "Meta's current approach to content moderation in conflict zones [may] lead to an appearance of inconsistency" drew on comments submitted by several experts, including Dr. Samson Esayas, an associate professor at BI Norwegian Business School, who noted an apparent disparity between Meta's "swift measures" moderating content in the European context of the Russia-Ukraine conflict, and its apparently "differential treatment between this conflict and conflicts in other regions, particularly Ethiopia and Myanmar."

Our **"UK drill music"** decision was informed by the specialist knowledge of several organizations. The Digital Rights Foundation argued that employing "a line-by-line lyrical analysis of the removed song to determine evidence of past wrongdoing or risk of future harm is notoriously inaccurate and that verifying supposedly factual statements within drill lyrics is challenging." The Electronic Frontier Foundation also criticized law enforcement's policing of lawful drill music. A public comment was also received from the Metropolitan Police Service, our first from a government agency. While the Metropolitan Police did not consent to publish the comment, it indicated it may provide such consent at a later point in time.

Our policy advisory opinion on **Meta's cross-check program** also attracted a wide range of comments. Organizations including the Center for Democracy and Technology, the Institute for Technology and Society of Rio de Janeiro, PEN America, and Mnemonic all shared their expertise with the Board. Democrat and Republican politicians also shared their concerns about cross-check by submitting public comments.

> "We've looked to precious information in public comments to help us make better decisions."
>
> Ronaldo Lemos
> OVERSIGHT BOARD MEMBER

The cases which received the largest number of comments in 2022 were published in early 2023. Our **"Iran protest slogan"** case attracted 162 public comments. Public comments confirmed that "marg bar Khamenei" (literally translated as "death to [Iran's Supreme Leader Ali] Khamenei") was widely used during recent protests in Iran. They also raised concerns that Meta had been incorrectly removing Farsi language content during these protests.

In our **"gender identity and nudity"** cases, many of the more than 100 comments we received came from people who identify as trans, non-binary or cis-gender women. Many commenters noted that, like the users in these cases, they too had experienced having their posts removed incorrectly. Comments expressed confusion about why Meta removed posts that included links to fundraising websites. In another comment, ACON, an HIV education NGO in Australia, wrote that content promoting HIV prevention messaging had been removed for sexual solicitation. This was echoed by Joanna Williams, a researcher who found that nine out of 12 sexual health organizations that she interviewed reported being negatively affected by Meta's moderation in this area.

A comment from the InternetLab research center expressed concern about the presumptive sexualization of women's, trans and non-binary bodies, when no comparable assumption is applied to cis-gender men. The disproportionate impact of content removals on women's bodies was also noted by a comment from Dr. Zahra Stardust. Finally, comments from the Gay & Lesbian Alliance Against Defamation (GLAAD) and The Human Rights Campaign Foundation raised concerns that content from users in marginalized groups is at greater risk of repeated or malicious reporting, where users report non-violating content to harass the person who posted it.

Our **policy advisory opinion on the removal of COVID-19 misinformation** received 181 public comments. This was the largest number of comments we received for a single decision or opinion in 2022. A submission from American Civil Liberties Union raised the concern that the difficulty in distinguishing, at scale, between fact and fiction, and between opinion, experience, and assertion of fact, means Meta will stifle speech that should be permitted.

Several submissions noted Meta's responsibility to address the risks to public safety given its reach and the role of its systems in amplifying misinformation. Concerns were raised about the adequacy of labels and demotions in addressing the risk of harm. For example, the submission from the senior vice president of the Center of Internet and International Studies highlighted concerns that labels are insufficient to address misinformation disseminated by politicians and prominent influencers.

Timeline of engagement activities in 2022

In 2022, Board Members, Trustees and Administration staff participated in nearly 100 events around the world. The timeline of engagement activities below provides a few examples of these events.

MARCH

A Board representative speaks about the value of independent digital content oversight at the **Mobile World Congress** in Barcelona.

MAY

A Board Member gives a speech at the **Oslo Freedom Forum** about social media's potential role strengthening democracy and resisting tyranny.

JUNE

Board Members and a Trustee participated in **RightsCon**, the world's leading summit on "human rights in the digital age." Panels covered topics such as online freedom of expression in Iran and multistakeholder governance.

Board Members meet in person for the first time in California, for meetings with **Meta senior leadership, Board Trustees,** and a range of stakeholders.

SEPTEMBER

A Board Member speaks at the **Concordia Annual Summit** held in tandem with the **United Nations General Assembly** in New York.

OCTOBER

Board Members join industry experts in New York to discuss the future of content moderation at the **Columbia Global Freedom of Expression** conference

NOVEMBER

A Board Member speaks at the annual **Web Summit** in Lisbon.

At the **Bread and Net** conference in Beirut, a Board Member participates in a roundtable attended by stakeholders from the **Middle East.**

What's Next
2023 AND BEYOND

Evolving our work with Meta

In 2022, we refined our work with Meta, publishing our first policy advisory opinions, and continuing to make an impact through our decisions and recommendations. In 2023, we will build on these strong foundations. In response to stakeholder feedback, we are continuing to iterate our approach and have set six goals for our work with Meta.

Oversight Board 2023 Goals

1. Publish our first summary decisions
2. Issue our first expedited decisions
3. Reach our updated Board membership goal for maximum efficiency
4. Deepen engagement around our seven strategic priorities
5. Pursue long-term plans for scope expansion
6. Monitor how Meta is implementing our recommendations and push the company to provide evidence of implementation and impact

1. PUBLISH OUR FIRST SUMMARY DECISIONS.

Of the 50 cases we shortlisted for selection in 2022, Meta immediately reversed its original decision in 32 of them — restoring or removing the content as necessary. In total, Meta has removed or restored more than 80 posts since 2020 due to the Board pointing out these errors, providing users with redress and revealing crucial insights into how the company moderates content. In 2023, Board Members on our Case Selection Committee will select some of these cases to be reviewed as summary decisions. These will set out why we consider the case to be significant and note the recommendations we have made in similar situations in past decisions and policy advisory opinions. Summary decisions will be drafted by the Case Selection Committee. They will be approved and voted on by the Case Selection Committee, rather than the full Board, and will not consider public comments.

2. ISSUE OUR FIRST EXPEDITED DECISIONS.

In 2023, we expect to publish our first expedited decisions. These cases will be referred to us by Meta on an expedited basis. Drafting and publishing a decision within days will allow us to contribute on issues of public concern and situations with serious impacts for human rights as they happen. A panel will deliberate, draft, and approve a written decision, which will then be published on our website. Expedited decisions on whether to take down or leave up content will be binding on Meta. Due to time constraints, these cases will not consider public comments and will be decided based on the information available at the time of deliberation.

3. REACH OUR UPDATED BOARD MEMBERSHIP GOAL FOR MAXIMUM EFFICIENCY.

While we originally expected the Board to reach 40 Members, three years of operations has shown us that, in practice, the optimal number of members allowing for timely, regular, and effective deliberation and decision-making would be 26. After the renewal of most of our existing Board Members in April 2023, and once the ongoing processes for selecting two new Board Members are complete, Meta will withdraw from the selection process. After this point, Board Member selection will be undertaken by Board Members and Trustees without Meta's involvement. The replacement of any Board Member has been and will remain at the discretion of the Board Members and Trustees alone. We expect the Board to have a complete set of Members in place by the end of 2023.

4. DEEPEN ENGAGEMENT AROUND OUR SEVEN STRATEGIC PRIORITIES.

In 2022, we chose seven strategic priorities based on an analysis of cases submitted to the Board, and issues facing users globally. In 2023, these priorities will guide the cases Meta refers to us and those we ultimately select to review. For all of our priorities, we will continue to work with stakeholders to understand the policies and enforcement practices that Meta most urgently needs to improve, and what kinds of cases could provide the opportunity to address them. We encourage stakeholders who specialize in these areas to reach out to us through our public comments process, roundtables, or through individual conversations.

5. PURSUE LONG-TERM PLANS FOR SCOPE EXPANSION.

In 2022, we gained the ability to make binding decisions to apply a warning screen when leaving up or restoring qualifying content. In 2023, we will continue our dialogue with Meta on expanding our scope further to include groups and accounts. While we foresee this work starting in 2023, Meta has told us that, for technical and operational reasons, groups and accounts are unlikely to come into scope before 2024. As a Board, we are also interested in exploring scope expansion in other areas, including content amplification and demotion.

6. MONITOR HOW META IS IMPLEMENTING OUR RECOMMENDATIONS AND PUSH THE COMPANY TO PROVIDE EVIDENCE OF IMPLEMENTATION AND IMPACT.

We will continue to closely monitor how Meta is implementing our recommendations and provide updates in our quarterly transparency reports. We will also push the company to provide evidence proving that it has implemented recommendations across policies and products, and share metrics on how these recommendations are impacting the experience of people who use, and are affected by, its platforms.

Sharing the benefits of independent oversight

In 2022, a growing number of companies sought external expertise on content moderation decisions. Spotify set up a Safety Advisory Council and microblogging site Koo set up an advisory Board. Twitter's former head of trust and safety has also called for content moderation councils to be established for the Google and Apple app stores.

Fundamentally, independent oversight is about firms opening up their internal processes and inviting outsiders to review their decisions. This kind of challenge and scrutiny leads to better and more robust decisions, helping to build trust with users in the long-term: a 'win-win' for communities and companies.

From the outset the Oversight Board was not just about improving Facebook and Instagram. It was also a chance to experiment with a new independent approach to content moderation that could be adapted to other platforms and tech companies. As a Board, we have developed a wealth of experience in the last three years that could help other companies make better decisions and better serve their users.

A UNIQUE EXPERIENCE OF BUILDING AN OVERSIGHT BOARD

Creating and operating a content governance oversight board, something no one had tried before, turned out to be more complicated than meets the eye. From day one, we worked to overcome a huge number of challenges, from technical infrastructure to building a workplace culture that balances the fast pace of content moderation against judicious deliberation. For months, we worked with Meta to create 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 decision-making process. And we learned, with our different nationalities, backgrounds, and viewpoints, how to deliberate cases with no easy answers.

In our three years of operations, we have learned and applied many lessons about how an oversight board for content governance should function. In particular, we have identified five characteristics that could help other tech companies looking to establish such oversight.

> "The freedom of speech of billions of people on the planet is too important to be left to one company alone."
>
> Alan Rusbridger
> OVERSIGHT BOARD MEMBER

- **Independence** – Any oversight body must be structured to allow for independent judgment. To be successful in scrutinizing content decisions, it must be free from the commercial, reputational, or political constraints of platforms. As a Board, we have not hesitated to overturn Meta's decisions. This separation is crucial for building legitimacy with users and civil society groups across the world.

- **Transparency** – To trust a company's decisions, people need to understand how those decisions are made. In preparing our decisions, we include as much previously non-public information about policies and their enforcement as we can. To be seen as accountable and genuine in their commitment to upholding free speech, companies must tell users why their posts were removed or their accounts deactivated. They must also do more to explain why they take decisions and be more transparent when governments or state actors call on them to remove content.

- **Diversity** – Most users of social media platforms are based outside of the United States and Europe. Many of the concerns about social media's negative impacts – as well as its benefits – are felt most acutely in countries located outside the US and Europe. To be able to trust a company's decision-making, people in these countries need more than just 'engagement' with their part of the world. They need to feel heard and represented at the decision-making table. At the Oversight Board, our members have lived in 27 countries and speak 29 languages. This diversity enriches our final decisions.

- **Human rights** – Today, tech companies face a significant challenge when moderating content: What rules should apply to billions of people of different nationalities, languages, and cultures? We think international human rights standards are a crucial part of the answer. These apply equally to everyone and provide a consistent framework to consider a user's right to free expression alongside other human rights, like the right to life or privacy.

- **Partnership** – Finally, you need partnership. Independent oversight bodies will only have a lasting impact if companies give them access to their data and processes. While there are sometimes instances where sharing information could compromise user privacy or allow bad actors to game policies, our work depends on data and information that only Meta can provide. This kind of openness also helps to build legitimacy and trust with users and civil society. Companies also need to show a willingness to implement recommendations, as Meta has done with many of our proposals. Partnership between a company and an oversight board requires a willingness to learn from both sides.

As a Board, we hope that the approach outlined above provides a credible framework for other tech companies looking to reap the benefits of independent oversight. Through our work with Meta, we have already overcome many of the operational hurdles, and learned many of the lessons, associated with establishing such a body. While any approach would need to be adapted to the specifics of a company's work, our experience could help firms benefit from independent oversight more quickly and with less cost.

As we think about the best path to holding the tech industry accountable, we encourage all companies to consider building independent oversight into their platforms and services.

> "We have seen how our decision-making processes are now being looked at by other social media platforms who are considering the same issues that we are."
>
> Kristina Arriaga
> OVERSIGHT BOARD TRUSTEE

Helping companies adapt to emerging regulation

We are interested in working with companies that share our belief that transparent and accountable content governance, overseen by independent bodies, is an essential part of creating an online environment that respects freedom of expression and other human rights. Companies that are willing to make a meaningful commitment to such standards and structures to benefit their users and society more widely, as Meta has done, will earn users' trust and demonstrate their serious intent to regulators.

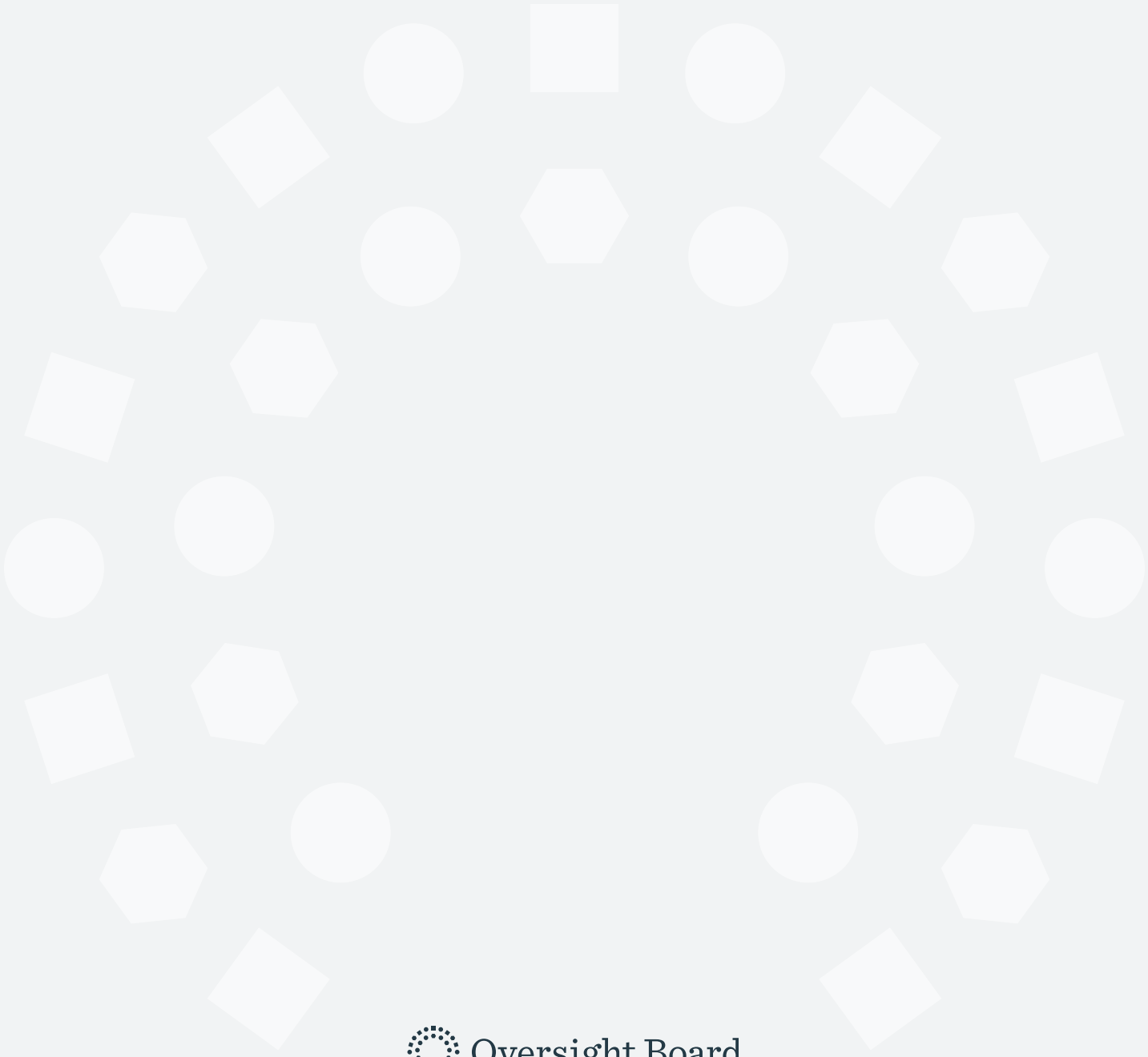
In the coming year, new regulation will bring new requirements for tech companies, leading many to change their approach. Past debates around social media regulation have presented government regulation and industry self-regulation as an 'either/or' choice, but we are now seeing movement towards 'co-regulation,' where elements of an independent regulatory mechanism are underpinned by legislation. While many actors will play their part in this emerging regulatory landscape, we believe that our independent approach, and our focus on transparency and treating users fairly, can be part of the solution.

> "We're not seeking to be the board for the whole industry. But we are seeking to share what we've learned, and work with companies interested in setting up different bodies to set standards and oversee content governance."

Thomas Hughes
OVERSIGHT BOARD DIRECTOR

Conclusion

At its best, social media can be an unparalleled catalyst for global connection and conversation. To reap these benefits, while containing its harms, is a daunting task. As a pioneering entity, the Oversight Board continues to apply lessons learned to improve itself. We believe that social media companies make content moderation decisions in a fairer, more principled way if they base those decisions on international human rights standards. We stand ready to share what we have learned so far with other companies and organizations that share our goals of increasing transparency and improving how people are treated online.



Oversight Board

www.oversightboard.com