Oversight Board

FROM BOLD EXPERIMENT TO ESSENTIAL INSTITUTION

How five years of independent oversight made Meta more accountable and protected the rights of users.

THE OVERSIGHT BOARD, DECEMBER 2025

MARKING 5 YEARS OF IMPACT 2020-2025

Table of Contents

Executive Summary

In 2018, Facebook was a company coming up against widespread condemnation. After launching on a wave of optimism followed by more than a decade of exponential growth, there was mounting fury coming from the company's more than 2 billion users worldwide. The Cambridge Analytica data-sharing scandal, in which the personal details of millions of accounts were shared without consent, brought attention to the critical issue of user privacy. Failures to deal with hate speech on the platform contributed to real-world harms and human rights violations during the Rohingya genocide in Myanmar.

At the same time, users saw their freedom of expression repeatedly compromised, as legitimate political speech, artistic voice, awareness-raising content and satire were all being removed from Facebook and Instagram in the name of rules that appeared vague and arbitrary.

During critical elections, disinformation and misinformation were making it more difficult for users to make informed choices and even distorting results. Content depicting and promoting self-harm and dangerous practices became more prevalent. Users outside the United States experienced limited content moderation tailored to their own regions, with decisions failing to account for local languages, culture or political context.

To make matters worse, users were often in the dark about how to appeal if their content was removed or what rules they had supposedly broken. There was a lack of transparency about the algorithms and processes that lead to the significant choices about what content would remain on or be removed from the platform.

Inside Facebook, there was an acknowledgement that groups of corporate executives should not make such highly consequential content moderation decisions on their own. This is what drove Facebook, now called Meta, to announce the creation of the Oversight Board in 2018. The Board selected its first cases in October 2020. It is still the only oversight body of its kind.

Transparency for Users and a Human Rights Perspective

Five years on, the Board has made important strides for Meta's global users, bringing transparency, reasoning and a human rights perspective to decisions that were long made behind closed doors, and with little or no public-facing rationale. The model we have built brings experts from around the globe to independently review sensitive content decisions on Meta platforms with input from the public and civil society. Our Board Members are politically, ideologically, geographically, culturally and professionally diverse. This means we can take better account of the varied contexts affecting speech and other human rights in different parts of the world. The Board makes binding decisions on whether content is left up or removed. We also make crucial recommendations for systemic improvements.

Meta committed to honoring our up-or-down decisions on particular content and is required to respond publicly to our recommendations, both promises the company has kept. Although Meta is not legally required to implement every recommendation, it has implemented 75% of the more than 300 we have issued, bringing more transparency and accountability. All recommendations helped inform the conversation and debates over content moderation policy and enforcement.

The result for the billions of users on Facebook, Instagram and Threads globally has been significant, from ensuring Meta adapts algorithms to better protect awareness-raising content and satire, to provision of better due process and second chances for users who break the rules and face penalties.

Meta is a more accountable, more transparent and better-informed company because of the Board's work, though there is still far to go. The Board has proved that independent oversight that is empowered to make binding decisions on content and non-binding recommendations that require a public response does work in practice, and, crucially, can serve as a framework for other platforms to follow.

Meta's decision to establish the Board, support it financially, impose safeguards to ensure its independent decision-making and empower it to make decisions and recommendations was unprecedented. At the time, the hope was that other social media platforms would take similar steps to hold themselves more accountable to users through independent oversight, perhaps even by engaging with the Board itself. Thus far, none have done so. The Board's interaction with Meta involves continuous advocacy and negotiation, and the Board never gets as much access or influence as it would like. Nonetheless, the Board is deeply respectful of the risks and obligations assumed voluntarily by Meta in the name of doing more to uphold the company's human rights responsibilities.

The Board began as an experiment, and we have approached our work in that spirit, eager to learn and refine. It has taken time to hone our approaches to setting priorities, choosing cases, gathering information for our decisions, crafting consequential recommendations and monitoring their implementation. We are proud of what we have achieved on behalf of Meta's users. We are also clear-eyed about the global influence of the platform, the limits of our scope and the pervasive challenges that persist. Looking ahead, we are committed to building on our singular role as a force for upholding respect for human rights throughout the technology industry.

Our Accomplishments for Users

The Oversight Board's work has led to greater public understanding of how Meta's systems function. When we identify risks or shortcomings in the company's content moderation, we recommend solutions that directly shape Meta's policies, processes and enforcement technology and practices. Since we started issuing decisions in January 2021, we have secured the following advances on behalf of users:

- **Expanding Freedom of Expression:** When we issue a decision, we defend not just one post but a principle: that the online public square must remain open and safe for a diversity of voices, even when speech is controversial, offensive or critical of those in power. Our recommendations have changed Meta's policies and enforcement approaches to support more speech remaining on Facebook, Instagram and Threads. This report explains how the Board took on these issues, including: in Iran, where "down with Khamenei" statements were restored to protect political speech; in Afghanistan, where more discussion and reporting on the Taliban was permitted if in the public interest; and in North America, where the Board restored posts from an indigenous artist that raised awareness over historic crimes against Indigenous People.

- **Transparency to Empower Users and Drive Accountability:** To be fair to users, we have insisted on Meta being more transparent about its rules and the actions it takes against content. In the UK, the Board uncovered enforcement mistakes in a case about drill music that unduly restricted artistic expression for marginalised groups. In multiple cases we pushed Meta to tell users which specific policy their content allegedly violates when it takes an enforcement action, a practice the company implemented from 2024.

- **Better Handling of High-Stakes Issues and Protecting Vulnerable Communities:** Our recommendations have prompted more attention to and responsive processes for conflicts and crises. Specifically, Meta has implemented a crisis protocol policy and a framework for handling content during elections. The company has also addressed bias in its cross-check system. Cross-check is intended to give high-profile users and pages with large audiences extra reviews on potentially violating content, but it was being implemented unfairly, disadvantaging regular users. Additionally, the Board has consistently defended vulnerable communities, such as human rights defenders and opposition leaders in repressive regimes, and our decisions have resulted in the removal of posts promoting homophobic violence.

- **Embedding Human Rights Principles into AI and Automation:** We are accelerating recommendations that highlight from a human rights perspective areas in which Meta's policies and systems are not keeping pace with AI-generated content. The Board's recommendations have pushed Meta for AI labels, leading to them being placed on billions of pieces of manipulated content, so users are empowered to assess trustworthiness. In 2021, we underlined that Meta's automated systems were mistakenly removing breast cancer awareness-raising posts, bringing about changes that give more protection to such content.

Our Learnings

Our ability to make strong recommendations is born from the design of our unique model. The diversity of views our global experts bring to issues, the invitation of public comments into cases and our human rights-based approach have enabled our work over the past five years.

The independent judgment of Board Members is protected through a series of mechanisms, including: the Board having its own leadership; fixed terms for Board Members; fixed funding through an irrevocable trust; and, Meta cannot withdraw funding already committed if it disagrees with our decisions.

As we go, we set out our learnings for users, technology companies and policymakers, outlining the key factors that have allowed us to develop a content governance oversight model that works. We also talk about the limitations we have encountered along the way.

ADVANCING MORE TRANSPARENCY AND ACCOUNTABILITY

By the Numbers

200+
DECISIONS ISSUED

4
POLICY ADVISORY OPINIONS

320+
RECOMMENDATIONS MADE

75%
RECOMMENDATIONS IMPLEMENTED
IN PROGRESS OR REPORTED AS ALREADY DONE

11,000+
PUBLIC COMMENTS ON CASES

What the Future Holds

Our model of independent oversight continues to advance more transparency, accountability, open exchange and respect for free expression and other human rights on Meta's platforms, strengthening users' rights globally. Still, the Board is under no illusions about the scale of the challenges facing the industry, as rapid AI advancements, social media regulations and geopolitical tussles over user rights to express themselves complicate how best to govern content at scale.

We look forward to engaging with these challenges by formalizing our learnings beyond individual cases and into white papers on key priorities, including the new era of AI and automation. We have already taken a step into broader conversations on content governance and platform accountability. The Board is also well positioned to partner with a range of global companies as they navigate issues arising from free speech debates worldwide. Even as the world changes, we remain committed to our founding mission – to embed free expression and other human rights into evolving technologies and advance accountability for users.

Expanding Free Expression

In support of our founding purpose to protect freedom of expression, we have identified situations in which Meta's moderation has unduly restricted speech, hindering political commentary, self-expression and public discourse online. Our recommendations are designed to prevent the wrongful removal of speech, including political protest, news reporting and awareness-raising on subjects of public interest and personal expression, while also protecting other human rights.

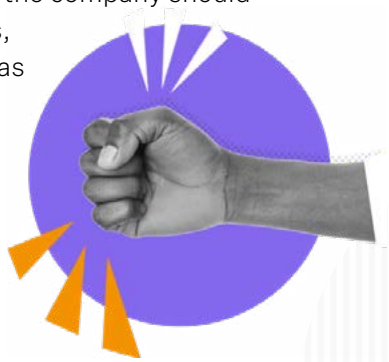
With billions of pieces of content posted daily, our recommendations are also directed at lessening the prevalence of moderation errors made by Meta's systems. We do so by seeking to pinpoint aspects of Meta's automated systems and human review processes that are prone to removing users' content due to misinterpretation, a failure to account for nuance or context, or the failure to implement applicable policy exceptions.

Removing Barriers to Speech Containing Non-credible Threats of Violence

We have repeatedly overturned Meta's mistaken removal of political content that contains figurative threats of violence, be it related to an impending election, political scandal or a protest movement (Russian Poem, Statements About the Japanese Prime Minister, Reporting on Pakistani Parliament Speech). Elections and politics are given to heated rhetoric and the use of metaphor and hyperbole. Users should be able to engage in robust, freewheeling political expression, including criticism of high-profile individuals like presidents or prime ministers, without platforms creating unnecessary barriers to this speech. We have consistently pushed Meta to bring a more refined approach to distinguish between threats of violence that may pose a real-world risk, and the language that, while harsh, is being used figuratively and thus fits within the bounds of political debate.

In a case involving the 2022 Iranian "Woman, Life, Freedom" protests, we argued that "marg bar Khamenei" statements should not be considered to violate Meta's rules against threats (Iran Protest Slogan). The phrase can be translated in more than one way (e.g., "death to Khamenei" - Ayatollah Ali Khamenei, the Supreme Leader of Iran), but in this context, it was being used as political rhetoric meaning "down with Khamenei". Meta agreed with our analysis in the context of ongoing protests, reversing itself to allow such statements. This led to a 29% increase in Instagram posts containing the statement, measured across the same pages, groups and accounts, after implementation.

And, in response to our recommendations, Meta is now conducting a policy development process on expression related to "calls for death," to consider how the company should approach such speech when it is used in non-threatening contexts, "such as banter, music and figurative calls for death." The Board has participated in the process and has provided feedback to Meta on potential policy and enforcement changes.



PREVENTING THE SILENCING OF OPPOSITION VOICES

Preventing the Silencing of Opposition Voices

Credible online threats of violence, while they constitute expression, are inimical to open discourse and political debate, especially when they come from those in authority. Particularly in politically repressive settings, social media, including Meta platforms, are often crucial settings for public discourse. If a range of voices, including those of political dissidents, are effectively silenced as a result of intimidation, the quality of debate deteriorates. The board has stressed that Meta's obligation to uphold principles of free expression requires it to ensure that such voices are not de facto excluded because of a climate of fear on its platforms (Cambodian Prime Minister, Human Rights Defender in Peru).

Sometimes threats of severe violence against individuals are veiled or disguised in nature and therefore require more intensive human review, with context taken into account. The Board has sounded the alarm about the risk of underenforcement (mistakenly leaving up content) of such threats and has called on Meta to bring additional resources to identifying veiled threats across cultural and political contexts (Human Rights Defender in Peru). In response, the company reports it is examining how to more clearly incorporate written, visual and verbal signals of coded threats into its assessments, with a commitment to augment this framework.

Preserving Speech in the Public Interest on Meta's Platforms

In many cases, the Board has decided that Meta has wrongly removed content that, even though it may violate the company's Community Standards, nonetheless has public interest value and should be exempted from removal under the company's "newsworthiness" allowance. In a case from Mexico, the Board restored a post showing the assassination of Mexican mayoral candidate José Alfredo Cabrera Barrientos under the newsworthiness allowance (Candidate for Mayor Assassinated in Mexico). And in a case from Haiti, the Board

overturned Meta's decision to take down a Facebook video showing people entering a police station and threatening violence, under the same exemption (Haitian Police Station Video). These decisions have extended the free expression rights of users on Meta's platforms by underscoring the importance of allowing debate and expression on timely matters of public

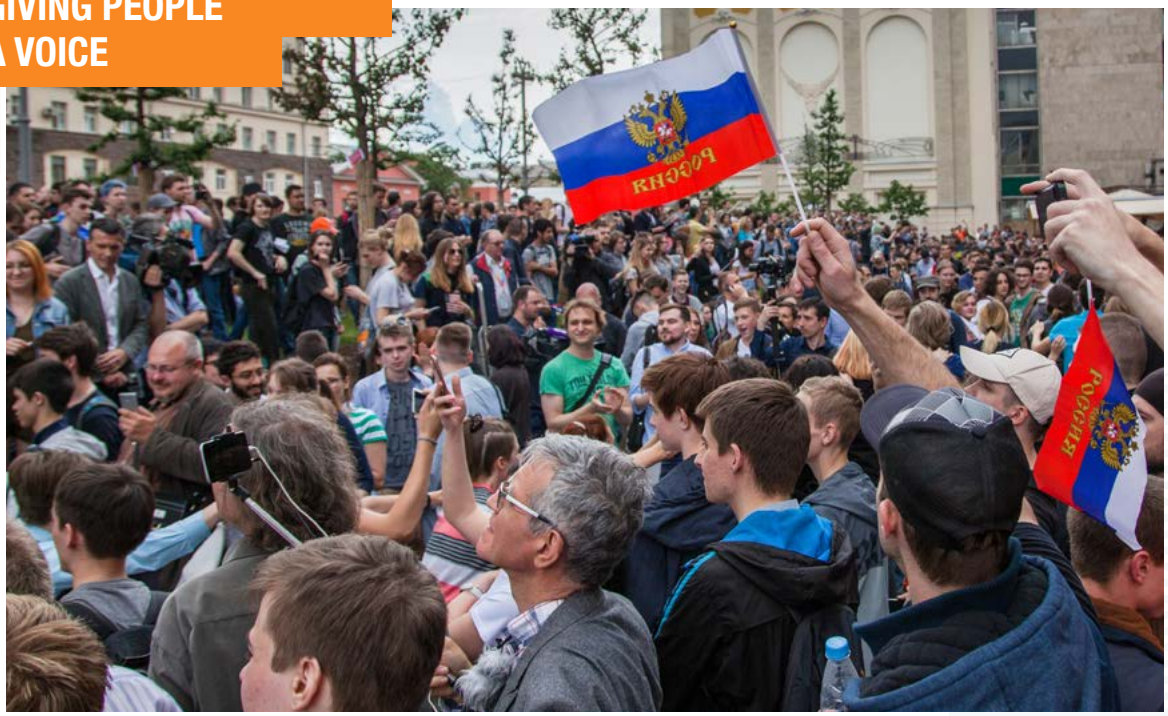
Preventing Content That has Been Wrongly Identified as Violating from Being Removed

Meta adds images and videos that human reviewers have identified as breaking the company's rules to its Media Matching Service banks. These banks can then automatically identify and remove the content each time it reappears on the platform, regardless of who posted it. But sometimes reviewers make mistakes, as in the 2022 case of a cartoon depicting police violence in Colombia (Colombian Police Cartoon). When content that is found not to have broken the rules nonetheless remains in these banks, the impact of incorrect moderation decisions is amplified across thousands or even millions of pieces of content. That's why we urged Meta to develop a way to reassess content with high rates of successful user appeals and remove it quickly from the bank as soon as it has been found to be non-violating. Meta has since established an internal team to improve management of this system, leading to the introduction of a "spike" detection mechanism to help it more quickly recognize a high volume of non-violating content and ensure that it is deleted from the bank and not wrongfully removed by Meta's automated systems when posted by users.

Allowing People to Share Context When Appealing Against Content Removal

Giving people a voice and listening to them can help platforms make better decisions when deciding whether to take down content. To avoid overenforcement (mistaken content takedowns) of posts calling attention to hate speech for reasons of condemnation, satire or awareness-raising, we asked Meta to create a convenient way for users to indicate in their appeal that their post fell into one of those categories ("Two Buttons" Meme, Sharing Private Residential Information). Between August and November 2023, the company introduced the facility for Facebook and Instagram users to add additional context to appeal submissions, explaining the nature of their post and making the case as to why it is non-violating. The company reported that it received more than 7 million appeals during the month of February 2024 from people whose content had been removed under the Hateful Conduct rules. Of those appealing, 80% chose to give additional context. The opportunity to add context has now been made available for user appeals across all Community Standards.

GIVING PEOPLE A VOICE



Offering Educational Initiatives to Reduce Burdens on Users' Speech

Some of the biggest penalties on users' speech result from strikes and account restrictions applied as part of Meta's enforcement against content. From our first policy advisory opinion (Sharing of Private Residential Information), we have urged Meta to consider alternatives to avoid disproportionate impacts on expression.

Introduced in early 2025 in fulfillment of a Board recommendation, Meta now sends an "eligible violation notice" to users committing their first violation of what Meta considers a "non-severe"* violation of a Community Standard. The notice includes details about the policy the user allegedly breached, along with the option of either appealing the decision or completing an educational exercise to better understand the applicable policies and thereby avoid a strike being applied to their account. More than 7.1 million Facebook and 730,000 Instagram users opted to view the "eligible violation notice" during a three-month period from January 2025. Among these users, nearly 3 million then embarked on the educational exercise, with the majority (more than 80% on Facebook and more than 85% on Instagram) completing the steps to avoid a strike and resulting account restrictions.

Meta received a positive response to a user feedback survey that followed the exercise, indicating that the educational segment was well-received and that users appreciate the approach of educational opportunities offered to avoid strikes on their account.

*This feature excludes the most severe Community Standards violations, such as sexual exploitation, high-risk drugs and representing, supporting or glorifying Meta-designated dangerous organizations.

Amending the Dangerous Individuals and Organizations Policy to Open Up Political Debate

To comply with international human rights law, rules restricting freedom of expression must, among other things, be clear, precise and publicly accessible. The Board applies those standards when evaluating Meta's policies. From the first decisions we issued, the Board has said that Meta's Dangerous Organizations and Individuals policy does not meet that requirement, with the result that it overly restricts political discussions online (Nazi Quote, Öcalan's Isolation, Shared Al Jazeera Post, Referring to Designated Dangerous Individuals as "Shaheed"). In 2022, the Board overturned Meta's decision to remove a post of a news article reporting on a Taliban spokesperson stating that schools for women and girls would be reopening. The post was removed on the basis that the Taliban had been designated by Meta (as well as many governments) as a dangerous organization and that the post included a form of "praise" for it. The Board noted that the post in question was reporting on the actions of the Taliban, and that such conveyance of information is an important right – both of those sharing information as well as of those receiving it – particularly in times of conflict and crisis.

In our 2024 policy advisory opinion, we noted Meta's disclosure that the Arabic term "shaheed" (loosely translated as "martyr") was the basis for more content takedowns globally under the Community Standards than any other single word. The takedowns resulted from a prohibition on the use of the word shaheed in proximity to the name of what the company had deemed a dangerous individual or organization. The prohibition presumed that all such posts would amount to the glorification of terrorists or other dangerous actors, when that was not true. The Board found that the term shaheed is used in many different cultural contexts and is not always a term of praise. Moreover, the proximity of the word shaheed to the name of a dangerous individual or organization did not always mean that a post was praising violence. To remedy this, we called for updates to the Dangerous Organizations and Individuals policy to avoid automatically removing references to dangerous individuals and organizations alongside the word shaheed. Following the implementation of most of our recommendations, the Board's data team identified a 19.5% increase in daily posts with over 50,000 views containing the word shaheed.

In a 2025 decision (Symbols Adopted by Dangerous Organizations), we again identified where the Dangerous Organizations and Individuals policy falls short, noting how Meta's internal criteria to identify symbols often used by hate groups is much broader than its public-facing explanation. It is important that the full definition is public so that users understand what they cannot post. The company has committed to a public-facing update.

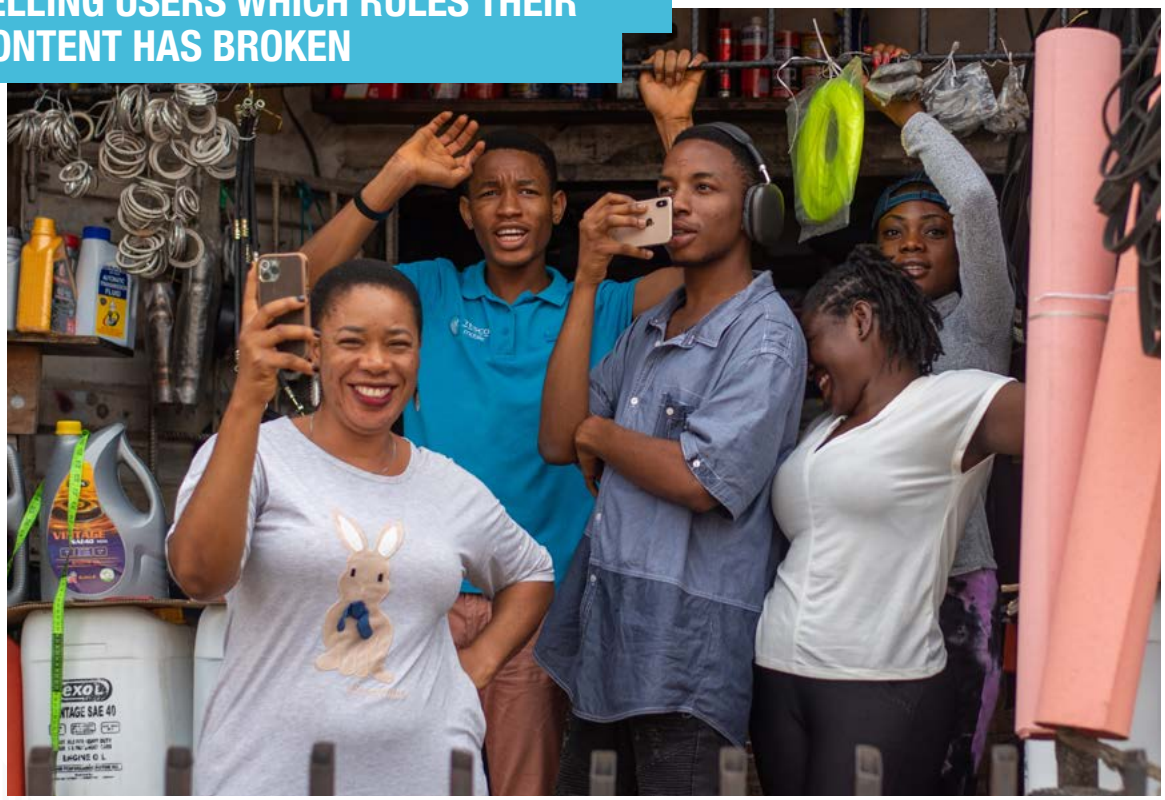
Transparency to Empower Users and Drive Accountability

The Board has pressed Meta to be more open about when governments influence content decisions or the policies governing its platforms. We have sought to empower users with knowledge and data about how enforcement algorithms work and how community standards are applied to their content. Meta has improved transparency in response to our recommendations to publicly report on enforcement errors and ensure that independent researchers have access to data.

Surfacing Government Requests to Review or Remove Content

Government influence on content decisions can lead to censorship and a chilling effect on free speech. When the state makes content removal requests behind closed doors, users do not know if platforms' decisions about their content are politically motivated. Transparency is one of the foundations of trust in online platforms for users and civil society. It empowers open discussion about whether government action is legitimate regulation or censorship, and both platforms and governments are more likely to be held to account as a result of user visibility into the dynamics between governments and powerful online companies.

TELLING USERS WHICH RULES THEIR CONTENT HAS BROKEN



In more than one case, we have called out the haphazard and opaque ways that content removal requests from governments are made to Meta (Shared Al Jazeera Post, Öcalan's Isolation, UK Drill Music). Transparency around these requests is essential so that users are aware if governments have flagged their content directly to Meta for review. The UK Drill Music case uncovered enforcement mistakes prompted by takedown requests by the Metropolitan Police Service in London that led to undue restrictions on artistic expression for people in marginalised groups. In the Öcalan's Isolation case, users were being prevented from expressing their human rights concerns about a prisoner from the Kurdistan Workers' Party being held in solitary confinement. In response to our recommendations in this case, Meta now notifies all users whose content was removed due to a government request that the content was removed, and unless restricted from doing so by law, which government entity made the request.

We have highlighted the importance of platforms reporting publicly on these government requests, recommending that Meta create a standardized system and provide data on reasons for content removal in these cases. The company confirmed publicly that it has implemented a new Content Reporting System, with standard questions to be answered by state actors when they make such requests, and that 97% of eligible regulators have been included. For Meta to ask a consistent set of questions at the time of the request makes it more likely that governments will be held to consistent rules and reduces the potential for arbitrary political decisions.

Telling Users Which Rules Their Content has Broken

Under international human rights law, individuals must be informed about what the rules are and how to assert their rights. On Meta platforms, users must be able to understand what part of the Community Standards their content has allegedly violated and why enforcement action was taken against it (Armenians in Azerbaijan, Breast Cancer Symptoms and Nudity, Nazi Quote, Depiction of Zwarte Piet, South Africa Slurs, Ayahuasca Brew, Sharing Private Residential Information). This enables people to adjust their behavior next time they post or equips them to submit a persuasive appeal. In response to our recommendations pushing Meta to provide such notifications, the company has, since 2024, informed users which specific policy their content is alleged to have violated when it takes an enforcement action.

Explaining Publicly how Algorithms That Predict Policy Violations Work

To demystify how Meta's classifiers (a type of algorithm) work when it comes to predicting whether a specific piece of content is likely to break the rules, we have pushed the company to provide a public explanation in its Transparency Center. We asked for explanations on the classifier's predictive accuracy thresholds for content, when taking no action, sending for human review or removing without human review (Referring to Designated Dangerous Individuals as "Shaheed"). In other words, we have pushed Meta to make clear how certain an algorithm must be concerning an alleged violation for that automated system to have the power to remove content. The company has not fully met all of our recommendations in this area,

EXPLAINING PUBLICLY HOW ALGORITHMS WORK

although Meta has responded by adding information to relevant pages in its Transparency Center explaining how its enforcement technology works. There is a clear need for even more transparency about how automated systems function, how they are audited and how their success is measured.

Setting Out Policy Exceptions More Clearly in the Community Standards

Some Meta Community Standards are implemented with exceptions and allowances that permit users to share speech in certain contexts, even if it might appear to violate the policy as written. For example, such exceptions may allow for certain content when shared as news reporting or awareness-raising. However, users are at a disadvantage if they do not understand which exceptions apply to which specific Community Standards. Under international human rights law, restrictions on freedom of expression must be clearly articulated so that speakers know what the rules are. This requirement is not met when exceptions are kept from users. We asked Meta to rectify this by creating a new section within the Community Standards setting out the relevant exceptions (News Documentary on Child Abuse in Pakistan). In the first half of 2025, Meta reported that it is still assessing the feasibility of this recommendation and exploring ways to make the allowances clearer in its Transparency Center

IMPROVING POLICIES THAT APPLY TO BILLIONS OF USERS

Better Handling of High-Stakes Issues and Protecting Vulnerable Communities

Times of crisis often reveal regular moderation processes to be inadequate. Our recommendations have pushed Meta to be more responsive to high-stakes scenarios such as elections and conflict. Recognizing that there are users around the world for whom Meta platforms are a primary vehicle for finding and sharing information, upholding users' rights to free expression – including the rights both to impart and to receive information – is particularly important in crisis situations where such knowledge can be a matter of life and death.

For example, in considering several cases about the war in Gaza, the Board has upheld free expression on all sides of the conflict. In 2023, in an expedited case – in which decisions are issued within 30 days of accepting the case – the Board overturned Meta's decision to remove a Facebook video depicting a kidnapping during the October 7 Hamas-led terrorist attack on Israel, and supported reinstating it with a warning screen (Hostages Kidnapped from Israel). In a case from 2024, the Board reversed Meta's decisions to take down a Channel 4 News (UK) report on the killing of a Palestinian child (Reports on the War in Gaza).

Prompted by a recommendation from the Board, Meta has introduced a crisis protocol to ensure that users rights are respected even in high-pressure situations. Users benefit when more resources and focus are applied to crisis situations, so that content that could cause imminent harm is acted on more quickly, and that information that is in the public interest is not removed in error.

The company has also introduced a framework to evaluate the way it handles content during elections, and more effective guardrails around the cross-check system, a program that Meta created in order to prevent overenforcement for users with a high profile, or those likely to escalate takedown errors in the media or directly to high levels among company executives.

Developing and Publishing a Policy for Crisis Response

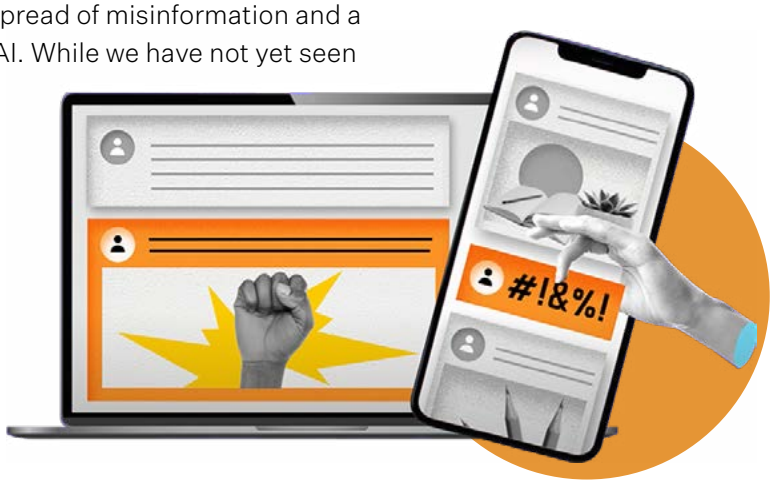
Social media platforms must be ready to respond to crises, whether a conflict, social unrest, contested elections or a natural disaster, to ensure that users' rights to free speech are protected even in fast-evolving, high-pressure circumstances. In these critical situations, additional processes are required to ensure rapid, equitable and consistent decision-making on content.

That's why in our first year of operations, we asked Meta to create a Crisis Policy Protocol (CPP) so that the company is prepared to act swiftly when needed (Former President Trump's Suspension). Following a comprehensive stakeholder policy forum, Meta created the CPP in August 2022, later adding a page about this mechanism on its Transparency Center. The CPP is a framework to intensify moderation when crises occur and make the application of policies consistent globally. The protocol is activated when pre-defined criteria identify high-risk situations, and it makes more tools available to make decisions quicker. This area of focus continues for the Board as we have recently recommended adjustments for more responsive activation of the CPP and called for greater leeway for content reviewers to make judgment calls based on context in high-risk situations (Posts Supporting UK Riots, Posts Sharing Speeches in Syrian Conflict).

Creating a Framework for Measuring Election Integrity Efforts

During elections, platforms must be ready to meet their responsibilities to both allow political speech and avoid serious risks to human rights, including the right to vote. However, it is challenging to understand whether the company's measures are sufficient for these events without a consistent framework by which the company approaches preparation for elections. In 2023, we noted that Meta should develop a framework to ensure election integrity and publicly report on those efforts (Brazilian General's Speech). We urged Meta to draw on local knowledge to identify coordinated online and offline campaigns aimed at disrupting democratic processes and set up permanent feedback channels to continuously improve the response around elections where political violence persists.

In June 2025, Meta confirmed that it had implemented the Board's recommendation and developed a framework consisting of eight core election integrity pillars. These will be applied across all its elections monitoring, to include factors such as risk management, cooperation with external stakeholders, reducing the spread of misinformation and a responsible approach to generative AI. While we have not yet seen this framework implemented nor been able to assess its efficacy, the announcement of its creation represents a positive step forward.



PROBING BIAS IN CONTENT MODERATION

Implementing More Effective Guardrails Around the Cross-Check System

Meta's cross-check system is a program designed to put extra checks in place before high-impact content – this means content from influential accounts or content that has gone viral – is removed. However, the cross-check system raised serious concerns for the Board about how the company was treating its most powerful users differently in ways that were non-transparent and that disadvantaged regular users. The existence of cross-check was revealed by a Facebook whistleblower. In response to the revelations, the Board undertook a thorough examination of the program, using the opportunity to seek and analyse information from the company that made the contours and implications of the program transparent to the public for the first time. In response to our policy advisory opinion on the program, Meta admitted to the Board that "a lack of governance" around these practices resulted in some users escaping enforcement actions.

We made 32 recommendations to put more effective guardrails around the cross-check system and ensure greater fairness to users, addressing: enforcement speed and quality of review; how eligibility for cross-check is determined; transparency on how the program works; and embedding human rights into the program, including the right to appeal.

For example, when a post from a user on cross-check lists was identified as violating Meta's polices, because of the cross-check system, that content remained up for days while additional reviews were carried out. It is during this time that violating content is at its most viral and, if it is something like revenge porn or a credible threat, therefore most harmful. We recommended clearer criteria for eligibility and transparency governing who is entitled to extra reviews and under what circumstances. In response, Meta changed the way the eligibility list is constructed, audited and monitored. Meta also took steps to eliminate delays in the review process, resulting in substantially fewer views on potentially violating content.

One of our key recommendations insisted that Meta deal with the backlogs in the review queue, which meant high-profile users' posts could remain on the platform for weeks or even months before being determined to be violating and removed. The company worked on this issue, reporting the following data back to the Board: For 90% of the tasks created in the company's cross-check review queues in the first half of 2023, there was a 96% decrease in resolution time (time taken for review and any subsequent enforcement), compared with the second half of 2022.

Probing Bias in Content Moderation

Perceptions of unfairness and bias in the moderation of political views can threaten the legitimacy of platform governance. Such bias can arise from several factors, including out-of-date or blunt policies, under-investment in language capabilities and cultural competence and repeated enforcement mistakes creating cyclical patterns of censorship whereby systems are learning and implementing the wrong lessons over time. In one case, we noted how the mistaken removal of content with opposing viewpoints on the issue of abortion was disrupting political debate (United States Posts Discussing Abortion). The Board recommended that Meta should regularly provide to the public the data that the company uses to evaluate the accuracy of its enforcement of the Violence and Incitement Policy. Meta has not provided the data, but such disclosure would allow for analysis of whether the errors in enforcement in the case – taking down content that did not actually violate the policy – were isolated or systemic.

In another case, which responded to public concerns around bias in Facebook's content moderation of Palestinian and Israeli content, the Board recommended that a third party look into the issue of systemic bias in content moderation on the platform in relation to the conflict (Shared Al Jazeera Post). Meta then commissioned a report from the advisory body Business for Social Responsibility, which indicated that Facebook's content moderation during the May 2021 Israel and Palestine conflict appeared to have had an adverse human rights impact on Palestinian users' freedom of expression. Much of this bias is related to a lack of internal language resources and poor guidance given by the company to content moderators. Limited language capabilities had also contributed to the underenforcement of antisemitic content. In 2022, Meta released a response, saying it would hire more content moderators who could review in relevant dialects and that it had deployed a machine-learning classifier in Hebrew.

Protecting Vulnerable Communities

The Board has consistently defended vulnerable communities, such as human rights defenders and opposition leaders in repressive regimes, from online harms. Board decisions have resulted in the removal of posts harming transgender people and promoting homophobic violence.

Overturning Meta's decision to leave up a Facebook video in which the Hun Sen, then Prime Minister of Cambodia, threatened his political opponents with violence, the Board concluded in June 2023 that leaving this content up was at odds with the company's human rights responsibilities (Cambodian Prime Minister).

In 2024, when overturning Meta's decision to leave up a post violently advocating for transgender people to commit suicide, the Board asked the company to improve its enforcement against such speech in non-textual form. In this instance, the Board found the content combined partially coded references to suicide with a visual depiction (transgender flag) of a protected characteristic group, taking the form of "malign creativity." This refers to the intentional use of coded language or visual/text memes that need context to be understood, often employed by people trying to avoid detection on social media (Post in Polish Targeting Trans People).

Systemic failings around Meta's enforcement were identified in a case of a video that violated four different Community Standards. Showing men who had been beaten for allegedly being gay in Nigeria, a country that criminalizes same-sex relationships, the video was reviewed by three moderators but still left up on Meta's platforms, risking immediate harm by exposing the men's identities. The Board overturned Meta's original decision to leave up the video (Homophobic Violence in West Africa).

Across a number of cases, including a 2023 decision about a video on Facebook showing identifiable prisoners of war, the Board has recommended that Meta preserve content depicting grave human rights violations or atrocity crimes, as defined by international law, and where appropriate, share with authorities such as international courts (Armenian Prisoners of War Video). In responding to the Board, Meta reported that it has implemented the recommendation.

Embedding Human Rights Principles Into AI and Automation

The Board has begun to get visibility into some of Meta's automated and AI systems, noting when we have found systemic failings, and drawing attention to areas in which enforcement technology is not working properly.

To ensure the fair and effective use of AI tools and automation to enforce policies, and curate and create content, we believe it is essential to embed human rights principles into their design and deployment.

AI Labels Empower Users

Across several cases, the Board has recommended labelling AI-generated content. This is to protect users' free expression while empowering them to better assess the authenticity and underlying message of a piece of content (Altered Video of President Biden).

In response to the Board's recommendations, Meta reports that it has been adding "AI info" labels to a range of AI-created or altered videos, audio and images on Facebook, Instagram and Threads. For example, over 29 days in October 2024, users viewed more than 360 million pieces of content with AI labels on Facebook and 330 million on Instagram. Of these, users clicked on labels on 6 million posts on Facebook and 13 million pieces of content on Instagram, to learn more about how the content had been created.

New Classifier Better Considers Text Signals in Automated Enforcement of Images

Considering the scale of social media, automated content moderation is essential to enforce standards and protect users from harm. However, reliance on automation can lead to situations where legitimate and non-violating content is taken down because automated systems find it hard to understand context – demonstrating awareness-raising or satire, for example – that a capable human moderation would be able to discern. The Board has addressed this issue from its earliest cases.

In 2021, we noted that Meta's automated systems were failing to pick up breast cancer awareness-raising posts when they contained images with text overlay (Breast Cancer Symptoms and Nudity). We raised concerns about the mistaken removal of this type of content, which Meta responded to by deploying a new health content classifier to enhance Instagram's techniques for identifying breast cancer posts. Over a 28-day period in 2023, an additional 1,000 pieces of breast cancer-related content were sent for human review instead of simply being removed. Meta has since reported that this classifier has not only improved moderation of this type of content but also, more broadly, the enforcement of images with text. The Board continues to encourage Meta to improve in this area and its ability to accurately enforce its policies at scale (Breast Cancer Awareness Content).

Our recommendations to Meta are evolving to consider the changing nature of AI-created content and potential harms, including whether Meta's policies are adapting to keep pace.

Our Independence Serves Users

In the space of just five years, the Board has moved from bold experiment to established model. It has a growing body of reasoned, principled jurisprudence that has created a first-of-its-kind blueprint for the application of human rights principles to online content. Many platforms have advisory bodies. The Board is unique because we are legally independent from Meta, we issue binding decisions and recommendations to which the company is obliged to respond, and we are comprised of diverse experts from around the world. At the same time, any new experiment will face challenges in terms of navigating the imperatives of a profit-making global business that is used by billions of worldwide users each day.

Here we share several of our learnings:

- **Make international human rights the foundation for consistent and fair decision-making:** Embedding a global human rights framework into an independent oversight board's model supports principled, consistent decision-making. The Board analyzes Meta's approach to content moderation by drawing on international human rights law, openly sharing which key treaties and provisions have guided our decisions. The analytical prerequisites and tests defined by human rights law help to ensure that our jurisprudence is consistently reasoned and can be followed over time. Strong freedom of expression protections are an important cornerstone of this recognized global framework, which also allows for some restrictions on expression in order to address other important goals such as the protection of the rights of others, national security or public order. The Board's grounding in international human rights law allows us to make decisions and recommendations that can be applied globally, in all jurisdictions. The extrapolation of obligations designed for governments onto a private company involves a degree of interpretation and imagination. More so than anybody in any industry, the Board has drawn those links and elaborated a practical paradigm for making the human rights obligations of corporations a reality.

- **Maintain structural independence:** An oversight body must be legally, institutionally and functionally independent from companies and governments to ensure that its decision-making is not swayed by business or state interests. As with the Board, this should include a mechanism for sustainable funding, empowering the entity to take responsibility for its own budget, leadership, strategy and operations. Through the Board's irrevocable trust, our operations are fully funded up to 2027.
The Board's adjudication and decisions are independent, but we rely on Meta's cooperation, including by answering questions as we deliberate cases, providing information necessary to support informed decision-making, accepting our recommendations and informing us of the impact of our r ecommendations on company policies so we can measure our efficacy.

- **Encourage robust commitments to implementation:** Oversight bodies strengthen their impact when they create rigorous processes for monitoring implementation of their decision-making and have the information necessary to do so. The Board's decisions on whether individual pieces of content stay up or come down are binding on Meta, as written into our Charter. We also make non-binding recommendations to improve how Meta moderates content, with the company being given 60 days in which to provide its response. The Board's data team has developed its own research methodology to independently track Meta's implementation of our recommendations, which is shared publicly on our website. This is distinct from Meta's own reporting, which requires the company to provide us with proof of implementation and means we can share our own assessment of the company's responses. At times, Meta has declined to implement our recommendations. Even in those instances, we believe that, at least on some occasions, spotlighting problems as well as bringing more information about Meta's practices to the public domain may lead to awareness, which eventually prompts remediation steps. In turn, our work can enable civil society and other actors to engage with Meta on how it can further improve.

- **Empower decision-makers from around the world:** To reflect the global nature of social media platforms, oversight bodies can benefit when they invite input into decision-making from different parts of the world. The perspectives of our 21 Board Members have been useful in ensuring that critical context and regional factors are taken into account before crucial decisions impacting users everywhere are made. We also make concerted efforts to reach out to affected stakeholders and civil society groups from all over the world to offer written input into our cases and attend roundtable conversations with the Board to better inform our decision-making. This dialogue can also help when different regulatory realities related to user rights need to be considered through a global lens.

- **Reflect the rights of people most reliant on platforms:** Stakeholders, who are reliant on or affected by a platform's services, need opportunities to provide their input. Providing such avenues can add legitimacy to such bodies. Such channels should be accessible and inclusive, encouraging the viewpoints of organizations and individuals who may have faced obstacles to directly engaging with platforms themselves. The Board's analysis in its cases is bolstered by external stakeholder input provided through a valued system of public comments. In proactively seeking such engagement, we have received more than 11,000 public comments from researchers, organizations and individuals to help shape our decision-making and push Meta to address its human rights responsibilities.

- **Promote safety by design:** Requesting early input from an independent oversight body into product innovations, platform features and content moderation technology makes good business sense, helping to push up standards and encourage best practices. The Board believes it is essential for human rights principles to be embedded during the development phases of rapidly evolving technology. To date, we have engaged with design problems arising from the features of Meta's platforms through our recommendations and found opportunities to improve a limited number of Meta's products. But in a rapidly evolving industry, where opportunities for growth also bring with them risks to user rights, allowing an oversight body to expand its scope as responsively as possible would be necessary to tackle ethical and systemic challenges successfully.

The Board's Evolution: A Look to the Future

In the five years the Board has been accepting appeals, social media platforms and the way users interact with them have changed significantly. In that time, the Board's work has reflected the most contentious issues arising from content moderation, as we have attempted to reconcile our strong commitment to freedom of expression with concern for other human rights. We have issued decisions addressing global elections, conflicts and societal debates, considered the design of enforcement systems and shed light on systemic problems affecting Meta's platforms.

Our scope too has evolved, to include both content being restored and taken down on Facebook and Instagram, the application of warning screens, and a third platform, Threads. In 2026, our scope expands once more as we pilot the ability to review Meta's decisions removing and impacting accounts, something that has created ongoing frustration for platform users.

Since the Board was first announced, we have understood the magnitude of the task before us, and we have taken seriously the concerns of skeptics and critics who doubted the legitimacy and efficacy of our efforts. Over the last five years, we have had frustrations and moments when hoped-for impact did not materialize. We have learned to focus on using the powers and influence we have to achieve the maximum impact that we can. We are uniquely empowered to seek information from Meta, get our questions answered, receive responses to our recommendations, and engage with company leaders. No other entity has that access. We will continue to use those openings to try to pull back the curtain for users, allowing them an improved understanding of a company that shapes their lives. When our decisions and recommendations are implemented, they put greater power in the hands of billions of people worldwide to express themselves and protect themselves online.

The Road Ahead: Beyond Content Moderation

In 2024, we began to formalize some of our learnings beyond individual cases into white papers, publishing observations on elections and a new era of AI and automation. More recently, the Board has taken a step into broader conversations on content governance and platform accountability, advancing how freedom of expression is thought about in systemic risks assessments carried out under new European Union regulations.

A commitment to an examination of the human rights implications of evolving technologies, both relating to and reaching beyond content moderation, will inform the Board's future work. Building on the insights we have gained from enforcement technology, automation and AI, the Board will be widening its focus to consider in greater detail the responsible deployment of AI tools and products. For example, as tech platforms push ahead with the design and development of generative-AI products, such as large language model chatbots and agentic AI, how can proper consideration for human rights be effectively baked into their design and development? And as generative-AI transforms the very nature of content online, what guidance can help the industry tackle newly emerging harms associated with its creation, many of which mirror harms the Board has addressed in the context of social media. We hope our contribution to these essential industry discussions delineates a way forward with a global, user rights-based perspective.

The Board is also well-positioned to partner with a range of global tech companies as they navigate issues arising from free speech debates globally, including when government pressure on platforms has a censoring effect online. We have an interest too in the tremendous challenges around how social media's youngest users experience platforms. How do we protect their rights to freedom of expression and access to information online in a way that keeps them safe from abuse and exploitation?

Such fundamental challenges demand engagement from a broad range of global perspectives, to ensure respect across political divides for rights such as free expression and to deliver ongoing platform accountability for users. The Board looks forward to playing a role in bridging these perspectives, offering recommendations that will address some of the most difficult policy and ethical dilemmas facing the tech industry in the months and years to come.

Oversight Board

www.oversightboard.com