Commentary

Burden of Proof: Lessons Learned for Regulators from the Oversight Board's Implementation Work

Naomi Shiffman, Carly Miller, Manuel Parra Yagnam, and Claudia Flores-Saviaga

1 Introduction

This paper examines the Oversight Board's work in tracking and independently verifying Meta's implementation of nonbinding recommendations, with the aim of supporting the development of best practices for social media auditing under emerging regulation. While the Board only considers decisions made by Meta, the lessons learned may be useful for regulators and other social media platforms.

The Oversight Board is an experimental governance body created by Meta in 2020 to make independent policy decisions and recommendations that address the most significant and difficult challenges on its platforms. Its goal is to ensure users' rights and interests by bringing greater transparency, consistency, and accountability to Meta's approach to content moderation (OSB 2022a). Facebook and Instagram users can submit an appeal to the Board if they disagree with Meta's decision to leave content up or take content down. The Board then deliberates on the appeal and determines whether or not Meta's decision adheres to Meta's content policies, values, and human rights standards. Meta also refers cases and requests policy advisory opinions to the Board to consider. Policy advisory opinions review a selection of Meta's policies and enforcement mechanisms, such as on health misinformation or privacy, and how they can be improved. Since October 2020, the Board has issued decisions in 74 cases and three policy advisory opinions.[1]

In addition to its binding decisions on case content, the Board can also issue nonbinding recommendations to Meta. To date, the Board has issued 242 recommendations.[2] Unlike its response to binding decisions, Meta is not obligated to implement recommendations. However, Meta does need to respond to recommendations publicly within 60 days, creating a level of transparency unique to the Board's work.

When the Board was initially conceived, the focus lay mostly on who the Board members would be and the legal framework they would use to make decisions. In the first year of the Board's operation, tracking Board recommendations consisted of noting whether Meta said they would agree to implement the recommendation or not—there was no mechanism for independent evaluation of the extent of Meta's implementation, whether Meta was misinterpreting the recommendation, or what sufficient proof of implementation would mean. The Board came to understand it was struggling to get visibility into these questions, and so created an Implementation Committee in July

1. The case decision number includes summary decisions. All numbers in this paper are as of December 11, 2023, except for those specified otherwise.

2. Meta counts 243 issued recommendations as of September 2023. This is due to discrepancies in counting recommendations in the early days of the Board when recommendations were not numbered.

10.54501/jots.v2i2.168

2021 and hired a team to build an analytic and data-driven infrastructure to support it, significantly changing the Board's approach to recommendations. The knowledge gained over the two-plus years since this work began are explained in detail in this paper.

Key learnings from the Board's work include:

- When making recommendations to social media platforms, both regulators and platforms can benefit from clear expectations, evaluation criteria, and opportunities for clarification to successfully implement recommendations.
- Creating a methodology to evaluate both the comprehensiveness of a platform's response to recommendations as well as the extent of implementation incentivizes additional information-sharing, sets a high bar for proof of implementation, and puts the burden on platforms to demonstrate compliance.
- Determining the correct data to gauge the size and impact of a policy or operational change is challenging. Starting the negotiation with what data platforms currently have can help regulators obtain data that is both available and speaks to their concerns.

Section 2 briefly outlines the Board's best practices for making policy, operations, and product recommendations. Section 3 describes the Board's monitoring and evaluation methodologies. Section 4 details and justifies the types of data the Board has requested in order to evaluate implementation effectively. Section 5 demonstrates the complexities of applied evaluation through case studies. The paper concludes with findings that may be helpful to regulators, platforms, civil society, and industry groups as they navigate emerging regulation and work to set standards and best practices.

2 Best practices for writing implementable recommendations

Emerging regulation such as the European Union's Digital Services Act, Singapore's Online Safety (Miscellaneous Amendments) Bill, and others will require platforms to undergo audits or systemic risk assessments to ensure regulation compliance (EC 2022; PARL 2022). However, unlike auditors who work with preexisting standards, the Oversight Board both makes policy and product recommendations, and evaluates adherence to them.

While this paper will not address the creation of recommendations, there are learnings that could be useful for regulators about the way Board recommendations are implemented and how the Board assesses that implementation. In some of the Board's early decisions, Meta misinterpreted the goals of the recommendations, and responded to an adjacent point or entirely reframed the recommendation. For example, in the Claimed COVID-19 cure case heard in 2021, the Board recommended that Meta publish a transparency report on how the company's Community Standards have been enforced during the COVID-19 pandemic (OSB 2021c). In its initial response to the recommendation, Meta stated that it will "continue to look for ways to communicate the efficacy of our efforts to combat COVID-19 misinformation" (Meta 2021a). Meta's response reframes the recommendation to be about its enforcement on COVID-19-related content, whereas the Board's initial recommendation asked Meta to disclose information about its enforcement generally during the COVID-19 pandemic. While this misalignment was later corrected via a conversation with Meta to clarify the recommendation, the original misunderstanding happened for several reasons: (1) the Board and Meta only communicated about the recommendation through written channels, (2) the recommendation did not include benchmarks for what the Board considered successful implementation, and

(3) the overall objective was unclear because the recommendation made multiple requests.

To ensure accurate interpretation moving forward, the Board incorporated the following best practices:

- Create opportunities to clarify recommendation intent in writing or in conversation with Meta following the publication of a case decision.
- Include an expected measure of implementation alongside each recommendation to serve as a benchmark of criteria would need to be fulfilled for a recommendation to be considered implemented.
- Ensure any given recommendation only asks for one specific system change, rather than compounding multiple requests into one recommendation.

Over time, as a product of this process, Meta's responses have increasingly correctly addressed the intent of recommendations (see more on this in the following section on evaluation).

3 Evaluation: Assessing Meta's responses to and implementation of Board recommendations

Once the Board makes a recommendation, it evaluates Meta's adherence to the recommendation along two axes: (1) comprehensiveness of the response, and (2) implementation of the recommendation. This is analogous to the way auditors will evaluate against standards under emerging regulation such as the Digital Services Act.

3.1 Assessing Meta's responses

Meta must publicly respond to the Board's recommendations within 60 days of publication. After this initial response, the company provides updates on each recommendation on a regular basis. Once these responses are published, the Board evaluates them to be either "Comprehensive," "Somewhat Comprehensive," or "Not Comprehensive," based on the factors below:

- Acknowledged and addressed all components of the recommendation
- Provided a concrete timeline
- Committed to concrete action

If Meta's response includes all three factors, the Board will rate it "Comprehensive"; if the response includes two out of three factors, it will rate it "Somewhat Comprehensive"; and if the response includes one or no factors, the Board will rate the response "Not Comprehensive." Meta's responses are evaluated each time it provides an update to the Board, and the comprehensiveness is measured cumulatively, incentivizing Meta to fill information gaps over time.

Learnings from this process have included:

- Fifty-three percent of Meta's initial responses to recommendations are rated "Somewhat Comprehensive" because they do not include a timeline for implementation or update. The Board's recommendations often address intersecting product systems that implicate many different teams and technologies within the company, such

that it is challenging for Meta to quickly align on a timeline for implementation. Subsequent responses from Meta do tend to include timelines more often as product roadmaps are locked in.[3]

- In part due to the discourse between Meta and the Board on the comprehensiveness of Meta's responses, as well as the previously mentioned opportunities to clarify the intent of recommendations, Meta's responses have increasingly been "Somewhat Comprehensive" or "Comprehensive." The Board has seen a decrease in "Not Comprehensive" responses over time.

3.2 Assessing Meta's implementation of Board recommendations

In addition to tracking Meta's responses, the Board also evaluates Meta's implementation of Board recommendations. While Meta has its own method for tracking recommendations, the Board places critical importance on independently verified implementation assessments, and tracks implementation using an evidence-based framework, detailed in Table 1 on the following page.

Table 1 also details the breakdown of issued recommendations vs. closed recommendations to date. Of the closed recommendations (meaning Meta has either finished implementing or declined to implement the recommendation, and therefore is no longer providing updates on it), over one-third have been fully or partially implemented, with evidence for implementation.[4]

Of the recommendations categorized as either fully or partially implemented as verified through published information, over half are Content Policy recommendations (see definition in Table 2 on page 6). Additionally, nearly half of the recommendations in progress are Enforcement recommendations that will ultimately impact nearly all users. Enforcement recommendations almost always take more time to implement due to the challenges of integrating changes into roadmaps for products and enforcement systems that have multiple, complex dependencies.

3. For example, the Board has recommended in several cases that Meta align its Instagram and Facebook Community Standards and indicate where they differ. The recommendation was first issued in January 2021 and has not yet been implemented. In its Q4 2022 response to the recommendation, Meta said that it is working with its "legal, regulatory, and product teams to scope and implement this plan, adjusted to reflect our new corporate brand and mission, while still fully implementing the spirit of the board's recommendations," emphasizing the cross-team coordination that is needed for the recommendation. See Meta's full response here: https://transparency.fb.com/sr/meta-quarterly-update-q4-2022.

4. Note that the status is subject to change—as the Board gets additional implementation evidence from Meta, more recommendations will move into the "implementation demonstrated through published information" category.

Table 1: Percentage of issued and closed recommendations across each implementation status category, as of December 11, 2023. Of the total number of closed recommendations, 41% are considered either fully or partially implemented, and 30% are considered declined, omitted, or reframed, including those declined after a feasibility assessment.

Implementation Status	% of Total Recs Issued	% of Total Recs Closed
Implementation demonstrated through published information. *Meta has published information or data that allows the Board to confirm the recommendation has been completed.*	17 %	26%
Partial implementation demonstrated through published information. *Meta has implemented a central component of the recommendation and has provided sufficient data verifying this to the Board. Meta's implementation may miss an important part of the recommendation, or implement it slightly differently than the Board intended, therefore receiving the "partial" status.*	10 %	15%
Progress reported. *Meta has committed to implementing this recommendation but has not published information or data that allows the Board to confirm progress or completion.*	33%	N/A
Meta reported implementation or described as work Meta already does but did not publish information to demonstrate implementation. *Meta has stated it implemented the recommendation or that the recommendation refers to work Meta already does, but has not provided information or data to confirm this.*	19%	29%
Recommendation declined after feasibility assessment. *After assessing the feasibility of the recommendation, Meta decided not to implement it, and provided information to contextualize its decision.*	5%	7%
Recommendation omitted, declined, or reframed. *Meta failed to respond to the recommendation, stated it would not take further action, or significantly reframed the recommendation to the extent that its work would no longer result in implementation of the Board's recommendation.*	15%	23%
Awaiting first response. *Meta has not yet responded to the recommendation.*	1%	N/A
Total	100%	100%

Table 2: Breakdown of implemented recommendations for which the Board has proof of either full or partial implementation, or which are "in progress," by type as of December 11, 2023.

Category	Definition	% Full or partial implementation	% In Progress
Content Policy	Recommendations asking for a policy development, modification, or clarification in the Community Standards.	51%	28%
Enforcement	Recommendations asking for changes to how Meta applies its policies, including with automated or human identification and review, and information provided to users in the content removal and appeals process.	29%	47%
Transparency	Recommendations asking Meta to make changes in the information it provides the public about its policies and practices.	20%	25%
Total		100%	100%

Learnings from this implementation evaluation process have included:

- For recommendations Meta says it has implemented, the number of recommendations for which the Board can independently verify full or partial implementation is only slightly higher than those for which there is no evidence ("Meta reported implemented or described as work Meta already does but did not publish information to demonstrate implementation"). The Board has a high threshold for considering a recommendation fully implemented, and the challenges of data access (as outlined in the next section) make hitting that threshold difficult.
- Transparency and Content Policy recommendations most often meet the threshold for verified implementation, and often most closely adhere to the original recommendation. Perhaps obviously, this is because, in most cases, their implementation requires a public-facing change (e.g., a change to the public Community Standards or to user notifications) that directly adopts the language of the recommendation.[5] Public-facing changes inherently lend themselves to easily visible proof of implementation. In contrast, enforcement recommendations are usually more resource-intensive and require Meta to share internal data to demonstrate implementation, as described in the next section.

4 Gathering data to evaluate implementation: the Board's approach to metrics

Like regulators around the world, the Board faces the challenge of overseeing a company that from the outside often appears to be a black box. The Board seeks not only to independently verify whether a recommendation has been implemented, but also to understand the ultimate impact of its recommendations on users. Therefore, when evaluating the implementation of a recommendation, the Board looks at two types of information: (1) implementation data and (2) impact data.

Implementation data serves to address whether Meta actually did what it said it would do when committing to implement a recommendation. This type of request is often for internal process or product documents, screenshots of a user notification flow, or internal product or policy research. It is the type of data an auditor would request to understand whether a platform complied with its regulatory obligations.

Impact data serves to assess the effectiveness of a recommendation. This often includes requests for user recidivism data (the amount of repeat policy violations), content moderation error rates, and user behavior or perception research. This is similar to the type of data a regulator might request to understand the impact of systemic risk mitigations under, for example, the Digital Services Act.

While obtaining the data necessary to conduct these assessments can be a significant hurdle, the bigger challenge lies in determining precisely what data to request in the first place. Often, Meta is hesitant to share the data requested by the Board due to privacy and security concerns. Other times, the data the Board requests is not actively tracked, or Meta is unable to provide sufficiently accurate data to meet the Boards' request. An example is described in more detail in the "Wampum Belt" case study below.

The Board has approached the challenge of gathering data by leveraging a combination

5. An example of a Transparency Recommendation considered to be fully implemented is the Board's recommendation in the Reclaiming Arabic words case, where the Board recommended that Meta "publish a clear explanation of how it enforces its market-specific slur lists" (OSB 2022b). In response to the recommendation, Meta created a page in its Transparency Center on how the company creates and uses its market-specific slur list.

of using public data and negotiating for data with Meta, all through a framework of data types likely to exist within the platform.

4.1 Gathering data: Publicly accessible data

There is some widely accessible data that can help with independent verification of implementation and assessment of impact. While the Board has access to CrowdTangle, a public insights tool owned and operated by Meta, and has been able to leverage it for limited impact assessments (see case studies section below), most cases the Board has heard relate to emblematic examples of complex content moderation issues that appear on user accounts, as opposed to public pages or groups. This means that content identical to the subject of a Board case rarely appears in the CrowdTangle database unless the content is viral, limiting its usefulness. Recently, Meta rolled out two new research tools—Meta Content Library and API—that provide researchers access to more publicly available content across Facebook and Instagram. The Board received a preview of these tools in late 2023 and is in the process of obtaining access.

Meta's public Community Standards Enforcement Report (CSER) is another source that theoretically could help the Board understand very general trends of violations and content prevalence across several Community Standards, but the report is far too high-level to provide the depth and specificity necessary to evaluate the impact of individual recommendations (Meta 2023a). The Board's recommendations target specific subsections within policies, and CSER does not provide that level of granularity. Even if it did, attributing a change in such metrics to a single recommendation would require controlling for confounding variables into which the Board does not have visibility.

4.2 Gathering data: Requests for internal data

Due to the limitations of public data, the Board requests additional internal data from Meta to independently validate the implementation of recommendations and demonstrate impact. The examples below are all for recommendations that Meta has reported to have implemented, but for which the Board has no evidence. The Board has requested the data below from Meta, and any fulfilled data requests would be considered sufficient evidence for implementation.[6] Some example requests are listed in Table 3 on the following page.

In addition to data for the purpose of demonstrating the impact of a particular recommendation, the Board has also requested the datasets below from Meta as a means of understanding the reasons and trade-offs underlying Meta's product decisions.

4.3 Gathering data: Requests for internal data

The Board uses a version of the framework in Table 5 on page 11, extrapolating from what is publicly visible on Meta's platforms, to understand what data is realistic to request. One can suppose that social media platforms collect similar types of data and/or metadata as those outlined below, in order of proximity to what gets shared in transparency reports.

Over the past two years, the Board has begun to receive data from Meta demonstrating implementation and the impact of recommendations. The verification and negotiation process for several recommendations is outlined in the next section.

6. Note that many of the recommendations in Table 3 were written before the Board incorporated the practice of explicitly stating measures of implementation as part of the recommendation. The Board requested data from Meta for these recommendations separately.

Table 3: Data the Board has requested from Meta in relation to specific recommendations. This table includes summaries or abbreviations of the recommendations, as some are quite lengthy.

Recommendation	Impact data request
Link the rule in the Hate Speech Community Standard prohibiting blackface to the company's reasoning for the rule, including harms it seeks to prevent. *(Depiction of Zwarte Piet no. 1)* (OSB 2021e)	The blackface policy violation rate across markets and languages before and after the recommendation was implemented.
Restore human review and access to a human appeals process to pre-pandemic levels as soon as possible while fully protecting the health of Facebook's staff and contractors. *(Punjabi concern over the RSS in India no. 2)* (OSB 2021g)	The average percentage of content eligible for a human review on appeal that was actually addressed by a human reviewer during and after the end of the workforce restrictions of the COVID-19 pandemic.
Ensure internal guidance and training is provided to content moderators on any new policy. Content moderators should be provided adequate resources to be able to understand the new policy, and adequate time to make decisions when enforcing the policy. *(Ocalan's isolation no. 8)* (OSB 2021f)	The change in false positive and false negative enforcement rates on posts actioned under the Dangerous Individuals and Organizations policy after implementation of the recommendation.
Develop and publicize clear criteria for content reviewers to escalate for additional review public interest content that potentially violates the Community Standards but may be eligible for the newsworthiness allowance. These criteria should cover content depicting large protests on political issues. *(Colombia protests no. 3)* (OSB 2021d)	The number of (estimated) views on content preserved by applying the newsworthiness allowance to content that would potentially violate the Community Standards within a 30-day period.
Study the impacts of modified approaches to secondary review on reviewer accuracy and throughput. . . . Meta should share the results of these accuracy assessments with the Board and summarize the results in its quarterly Board transparency report to demonstrate it has complied with this recommendation. *(Wampum belt no. 2)* (OSB 2021h)	The reviewer accuracy rate for moderators conducting a secondary review of a piece of content.
Ensure that users are always notified of the reasons for any enforcement of the Community Standards against them, including the specific rule Facebook is enforcing. Doing so would enable Facebook to encourage expression that complies with its Community Standards, rather than adopting an adversarial posture toward users. *(Armenians in Azerbaijan no. 1)* (OSB 2021a)	The rate of recidivism for users across Community Standards before and after the recommendation was implemented.

Table 4: Data the Board has requested from Meta to understand product decisions more generally.

Data Type	Justification
Classifier test sets for a specific classifier,[a] including (1) the ground truth label for each piece of content (i.e., Meta's anticipated policy action for the piece of content); (2) the classifier output for each piece of content; and (3) the action threshold for this classifier.	Requesting this type of data allows the Board to calculate the precision and accuracy of a given classifier and understand how the outcome of a classifier's decisions might be different if the action threshold were higher or lower. Since classifiers are used in virtually every aspect of content moderation, requesting this type of data can be revealing about the decisions Meta makes about where to draw the line across all its policies.
Internal user research	Both the results and the underlying data of experiments Meta has run on its own platforms to make product decisions, such as the level of detail in user notifications or how it weights different types of content in ranked feeds, provide clarity on why it made certain decisions and why it may decline to implement certain recommendations.

a. A classifier is a technical system trained to learn and categorize content based on ground truth data. "Classifier test sets" refer to a collection of data used to test the precision (how closely the content categorised by the classifier matches the intended category definition) and recall (how successfully the classifier identifies all content matching the intended category definition out of a set of content) of the trained classifier.

Table 5: Oversight Board Data Index

Oversight Board Data Index		
Data type	Details	Proximity
User interactions with the platform	• Interaction volume/count (likes, shares, views, etc.) • Metadata (timestamps, actors, location, etc.) on the lifecycle of content: Creation of content; Edits of content; Deletion of content • Country (estimated) • Language (detected) • Social graph links and owned assets/content for every profile	3 levels less aggregated than transparency report
The content moderation systems themselves	• Automation code • Automation training sets, results, and procedures • Policy development testing data • Internal audit / systemic risk investigations results • Content Moderation Quality / Accuracy data (datasets, results, sampling protocols, etc.) • Data about Content Moderation workforces (location, type, amount of work, etc.) • Data about overall operational goals	2 levels less aggregated than transparency report
Content interacting with content moderation systems	• Volume of reports • Detection type (automated vs. human report) • Decision taken (nonviolating, delete, any other action) • Policy violated and any other labels (policy, subpolicy, notes, etc.) • Time to resolution • Type of employee that took the decision • Tool through which the decision was taken • Type of review/decision (regular, appeal, quality, training, etc.) • Time/date of the review • Prevalence samples and labeling results	1 level less aggregated than transparency report
System data that makes it to transparency reports	• Prevalence • Detection • Actions volume per policy (high level) • Volume of appeals and restores per policy (high level)	Transparency report data (most aggregated)

5 Evaluating implementation: Case studies

The Board leverages the data it gathers to evaluate whether Meta has actually implemented its recommendations. The successes and challenges inherent in the Board's evaluation methodology and data requests are illustrated in the four case studies below. All four case studies explore the dynamics of verifying Meta's implementation of a recommendation, including running into issues of a small sample size, historical metrics, and working with the platform to obtain evidence of implementation and impact.

5.1 The Wampum Belt case: the challenge of tracking moderation impacts on small communities

Case description

In 2021, the Oversight Board took a case concerning indigenous art and hate speech (the "Wampum Belt" case). A user posted a picture referencing the May 2021 discovery of unmarked graves at a residential school for indigenous children in British Columbia, Canada (OSB 2021h). Meta's automated systems detected the content as potential hate speech, and a human moderator agreed the post violated Meta's Community Standards. The Board found that the content was a clear example of "counter speech" and was protected by Meta's Community Standards for speech used for empowerment and awareness-raising.

Recommendation and requested data

The Board recommended that Meta conduct an accuracy assessment focused on Hate Speech policy allowances that cover artistic expression about human rights violations, and requested that Meta share the results of the accuracy assessment with the Board:

> Conduct accuracy assessments focused on Hate Speech policy allowances that cover artistic expression and expression about human rights violations (e.g., condemnation, awareness raising, self-referential use, empowering use). This assessment should also specifically investigate how the location of a reviewer impacts the ability of moderators to accurately assess hate speech and counter speech from the same or different regions. The Board understands this analysis likely requires the development of appropriate and accurately labeled samples of relevant content. Meta should share the results of this assessment with the Board, including how these results will inform improvements to enforcement operations and policy development and whether it plans to run regular reviewer accuracy assessments on these allowances, and summarize the results in its quarterly Board transparency report to demonstrate it has complied with this recommendation. (Wampum Belt recommendation no. 3)

In this case, the data requested can be thought of as "implementation data": because the Board's recommendation centered on conducting an assessment, proof of the assessment as described in the recommendation would have been sufficient to demonstrate that it was implemented.

Meta's response

Meta responded by stating it needed to assess the feasibility of the recommendation (Meta 2021c). In its initial response, Meta noted several challenges with implementing the

recommendation. First, the company noted that it does not "have specific categories in our Community Standards or Community Guidelines for allowances for artistic expression or expression about human rights violations" (Meta 2021c). Meta noted that it allows for "condemnation, awareness raising, self-referential use, and empowering use," but that any experiment to assess accuracy would look for these allowances "generally, rather than at the granular level the board recommends." Second, the requested sample was not readily available. Meta stated, "we do not have an easily identifiable sample of content that falls under our hate speech allowances to test our automated and human review systems against." The company's automated enforcement systems do not identify nonviolating content, so Meta's subject matter experts would have needed to manually label each piece of content. Last, the sample would need to be large enough for subject matter experts to produce meaningful results (Meta 2021c).

Meta's provision of evidence

Thirteen months after the recommendation was initially issued, the Board received data from Meta on a broader Hate Speech accuracy assessment (Meta 2022b). Meta provided this because they "determined that a more system-level option for better understanding the accuracy rates of how [they] apply policy allowances would be more accurate" (Meta 2022a). Specifically, the Board received Meta's Hate Speech precision metric, which Meta stated is "consistently very high" (Meta 2022b). While the provision of this metric addresses the spirit of the Board's recommendation at a high level, it does not reveal anything about the original concern relating to allowances for artistic expression under the Hate Speech policy or assessing accuracy for the impacted group.

Lessons learned

Meta's responses highlight how resource-intensive it would have been to implement the Board's recommendation as originally written. This resource-intensiveness led to reframing the recommendation toward a less specific interpretation that Meta considered feasible. The Board's intention was to get Meta to pay attention to an impacted community despite it representing such a small percentage of Meta's user base—and Meta ultimately declined to do so because the high overhead of such a task would draw resources away from allegedly higher priority work. This is a somewhat straightforward lesson in the ways that content moderation at scale prevents attention to communities that do not make up a large or monetarily important enough percentage of the user base to warrant prioritization on product roadmaps. A secondary lesson is that requested data might not be readily accessible or be easy to construct.

5.2 The Breast Cancer Symptoms and Nudity case: the challenges of data retention and setting a baseline

Case description

In October 2020, a user in Brazil posted a picture to Instagram with a title in Portuguese indicating that the post was meant to raise awareness of signs of breast cancer. The post was removed by an automated system enforcing Facebook's Community Standard on Adult Nudity and Sexual Activity. The Board overturned Meta's decision and issued recommendations, in what it referred to as the "Breast Cancer Symptoms and Nudity" case (OSB 2021b).

Recommendation, Meta's response, and requested data

The Board recommended that Facebook change its enforcement systems related to the content:

> Improve automated detection of images with text-overlay so that posts raising awareness of breast cancer symptoms are not wrongly flagged for review. (Breast Cancer Symptoms and Nudity no. 1)

Meta agreed to implement the recommendation. The Board requested that Meta share impact data on the extent to which false positive identifications of content violating the nudity policy were reduced following implementation. The motivation behind this request was to demonstrate that with better accuracy in automated detection, overall false positives would likely be reduced.

The Board also requested as an alternative the nudity classifier precision metric, which would reflect on false positives (as precision increases, false positives decrease, but so does the amount of content detected).

Meta's provision of evidence

Ultimately, Meta's implementation team said they could share the number of pieces of content that were routed for human review rather than being automatically removed due to improvements to the classifier. They confidentially shared the types of improvements made to the classifier with the Board, and shared publicly that over the course of a 30-day period in 2023, 2,500 pieces of content were routed to human review rather than automatically removed because of the improvements (Meta 2023b). Meta also volunteered additional proof of impact—in addition to improving the classifier as requested, Meta deployed a new health content classifier to further enhance Instagram's techniques for identifying breast cancer content, which contributed to an additional 1,000 pieces of content sent for human review rather than automatically removed over a 30-day period (Meta 2023b).

Lessons learned

The Board took away two lessons from this exchange with Meta. First, data retention is a key blocker to understanding historical improvement over time for platforms. While Meta had both improved its text-overlay classifier and launched the new computer vision classifier in July of 2021 (Meta 2021b), it was unable to provide the Board with a sum of all pieces of content that had been preserved over the course of two years due to its data retention policies. These policies are often created to comply with legal obligations, highlighting the tension between transparency and accountability on the one hand, and data privacy on the other.

Second, Meta was unable to provide the Board with a denominator or any contextual information for either of these metrics. The Board requested additional contextual information such as the total number of pieces of breast cancer-related content removed from the platform over a set time period, even limited to content geolocated to Brazil or in Portuguese, as that was the context of the original case. The requests were ultimately unsuccessful, and while the discussion of sensitive healthcare topics are now better protected for potentially thousands of users per month, it is essentially impossible to understand the impact of the recommendation on policy enforcement in context.

5.3 The Iran Protest Slogan case: When public data can be useful

Case description

In July 2022, a Facebook user posted in a group that describes itself as supporting freedom for Iran. The post contains a cartoon of Iran's Supreme Leader, Ayatollah Khamenei, in which his beard forms a fist grasping a chained, blindfolded woman wearing a hijab. A caption below in Farsi states "marg bar" the "anti-women Islamic government" and "marg bar" its "filthy leader Khamenei." The literal translation of "marg bar" is "death to." However, it is also used rhetorically to mean "down with." The slogan "marg bar Khamenei" has been used frequently during protests in Iran over the past five years, including the 2022 protests. The Board overturned Meta's decision to remove the post, and issued recommendations in what it refers to as the "Iran Protest Slogan" case (OSB 2023a).

Recommendation, Meta's response, and requested data

The Board recommended that Meta allow users to post the slogan "marg bar Khamenei" in the context of the Iran protests:

> Pending changes to the Violence and Incitement policy, Meta should issue guidance to its reviewers that "marg bar Khamenei" statements in the context of protests in Iran do not violate the Violence and Incitement Community Standard. Meta should reverse any strikes and feature-limits for wrongfully removed content that used the "marg bar Khamenei" slogan. The Board will consider this recommendation implemented when Meta discloses data on the volume of content restored and number of accounts impacted. (Iran Protest Slogan no. 3)

Meta stated that they implemented the recommendation but did not provide any evidence to support their claim (Meta 2023b). In this case, the data the Board requested can be considered "implementation data," as it is about getting proof that content was restored.

Meta's provision of evidence

Meta did not provide any internal data to demonstrate a restoration of posts following the implementation of this recommendation. However, the Board's access to CrowdTangle enabled it to assess whether there was evidence for implementation.

Using public data obtained from CrowdTangle, the Board investigated whether there was a statistically significant difference in the number of posts present on Facebook and Instagram following implementation of the Board's recommendation compared to before the Board's decision. The Board did not find statistically significant differences in the number of posts on Facebook groups and pages that were restored because of implementing the recommendation. However, there were nearly 30 percent more posts on Instagram after the Board's decision, unattributable to random chance, which supports the notion that this increase was due to the implementation of the Board's recommendation. The full methodology of the study is in Appendix A of this paper.

Lessons learned

Because the content in question was prevalent in public conversations, the Board was able to use CrowdTangle to evaluate implementation of the recommendation. This underscores the stark differences in trying to evaluate the impact of changes to policy

on very visible trends, as opposed to impacts on content or users that do not enter the public sphere.

Additionally, the Board's ability to independently verify implementation of this recommendation demonstrates the importance of live data pipelines both for researchers and regulators—even when a platform is unable or unwilling to provide evidence of a specific action, evidence for it may be visible in pipelines like these.

5.4 The Pro-Navalny case: Growth in the Board's visibility on internal metrics

Case description

In 2021, the Oversight Board took a case concerning protests in support of imprisoned Russian opposition leader Alexei Navalny (OSB 2023b). The Board overturned Meta's decision to remove a comment in which a supporter of Navalny called another user a "cowardly bot." The Board found that while the removal was in line with the Bullying and Harassment Community Standard, the current Standard was an unnecessary and disproportionate restriction on free expression under international human rights standards.

Recommendation, Meta's response, and requested data

The Board recommended more transparency to users regarding potentially violating content:

> Whenever Facebook removes content because of a negative character claim that is only a single word or phrase in a larger post, it should promptly notify the user of that fact, so that the user can repost the material without the negative character claim. (Pro-Navalny no. 6)

Meta committed to fully implement the recommendation. Since this was an early case and there was no measure of implementation included in the recommendation, the Board later asked Meta to provide impact data in the form of the count and percentage of content preserved following user amendments to posts.

Meta's provision of evidence

In October 2023, Meta shared the following information with the Board:

> 'In response to a recommendation from the Oversight Board in 2021, Meta committed to explore ways of notifying users of potential violations to the Community Standards before we take an enforcement action. Since highlighting specific violating words could result in misleading notifications to the users, we focused on classifying the overall content to promptly notify users and give them an option to delete and repost before any enforcement actions are taken. Currently, when our automated systems detect with high confidence a potential violation in content that a user is about to post, we may inform the user that their post might violate the policy. This provides an opportunity for users to understand our policies, then delete and post again without the violating content. Over the 12-week period from July 10, 2023 to October 1, 2023, across all notification types, we notified users regarding more than 100M pieces of content, with over 17M notifications related to Bullying and Harassment. Across all notifications, users opted to delete their posts more than 20% of the time.

PLEASE NOTE: All information is aggregated and de-identified to protect user privacy. All metrics are estimates, based on best information currently available for a specific point in time.

Lessons learned

Meta was able to share 12 weeks of data at a time rather than only 30 days in this most recent data disclosure, as well as previously unshared information about user behavior in response to the product changes recommended by the Board. The disclosure advanced the Board and the public's understanding of the effectiveness of behavior-shaping moderation approaches beyond the leave-up / take-down binary. It also gave background on Meta's decision-making when implementing the recommendation.

6 Conclusion

The Oversight Board's work in making recommendations, determining evaluation methodology, and obtaining data to conduct the evaluation has led to significant learnings in a complicated terrain that may be helpful to regulators as they implement emerging regulation.

When making recommendations to social media platforms, both regulators and platforms can benefit from clear expectations, evaluation criteria, and opportunities for clarification to successfully implement recommendations. Additionally, creating a methodology to evaluate both the comprehensiveness of a platform's response to recommendations as well as the extent of implementation incentivizes additional information-sharing, sets a high bar for proof of implementation, and puts the burden on platforms to demonstrate compliance.

Determining the correct data to gauge the size and impact of a policy or operational change is challenging. The scale at which social media platforms operate, and the overhead involved in tracking and validating data for consumption by regulators, makes it very difficult for an external body to assess causal impact of new or amended policies, enforcement systems, and transparency initiatives. While regulation will have the benefit of binding authority on requesting metrics via audits, it also requests that auditors vouch for a certain level of assurance and mitigate potential auditing risks. Therefore, engaging in a dialogue with platforms over which are the correct metrics to assess may benefit the ultimate goals of the regulation and support the emergence of auditing norms.

Finally, the Board's work in case decisions and crafting recommendations, as well as assessing the effectiveness of those recommendations, is in some ways analogous to conducting systemic risk assessments and evaluating their effectiveness. The Board's learnings could be useful to leverage as regulators, civil society, and industry groups collaborate to set standards and establish this field of practice.

The Board hopes that regulators will benefit from its work over the past two years and take its learnings about the trade-offs inherent in policy language, implementation, and auditing into account.

References

European Commission, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC, Article 37 (OJ L 277, p. 67–69 Oct. 27, 2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065.

Hayes, Adam. 2023. "Wilcoxon Test: Definition in Statistics, Types, and Calculation." *Investopedia,* Updated May 14, 2023. https://www.investopedia.com/terms/w/wilcoxon-test.asp.

Meta. 2021a. "Case on Hydroxychloroquine, Azithromycin and COVID-19," February 25, 2021. https://transparency.fb.com/oversight/oversight-board-cases/hydroxychloroquine-azithromycin-covid-19/.

———. 2021b. "Meta Q2 + Q3 2021 Quarterly Update on the Oversight Board," November 9, 2021. https://about.fb.com/wp-content/uploads/2021/11/Meta-Q2-and-Q3-2021-Quarterly-Update-on-the-Oversight-Board.pdf.

———. 2021c. "Oversight Board Selects Case Regarding a Post Depicting Indigenous Artwork and Discussing Residential Schools," December. Updated June 12, 2023. https://transparency.fb.com/oversight/oversight-board-cases/indigenous-artwork-residential-schools.

———. 2022a. "Meta Q1 2022 Quarterly Update on the Oversight Board," May 17, 2022. https://about.fb.com/wp-content/uploads/2022/05/Meta-Q1-2022-Quarterly-Update-on-the-Oversight-Board.pdf.

———. 2022b. "Meta Q2 2022 Quarterly Update on the Oversight Board," August 25, 2022. https://transparency.fb.com/sr/meta-quarterly-update-q2-2022.

———. 2023a. "Community Standards Enforcement Report." https://transparency.fb.com/reports/community-standards-enforcement/.

———. 2023b. "Meta Q1 2023 Quarterly Update on the Oversight Board," May 1, 2023. https://transparency.fb.com/sr/meta-quarterly-update-q1-2023.

Oversight Board. 2021a. "Armenians in Azerbaijan," January 28, 2021. https://www.oversightboard.com/decision/FB-QBJDASCV/.

———. 2021b. "Breast cancer symptoms and nudity," January 28, 2021. https://www.oversightboard.com/decision/IG-7THR3SI1/.

———. 2021c. "Claimed COVID Cure," January 28, 2021. https://www.oversightboard.com/decision/FB-XWJQBU9A/.

———. 2021d. "Colombia protests," September 27, 2021. https://www.oversightboard.com/decision/FB-E5M6QZGA/.

———. 2021e. "Depiction of Zwarte Piet," April 13, 2021. https://www.oversightboard.com/decision/FB-S6NRTDAJ/.

———. 2021f. "Öcalan's isolation," July 8, 2021. https://www.oversightboard.com/decision/IG-I9DP23IB/.

———. 2021g. "Punjabi concern over the RSS in India," April 29, 2021. https://www.oversightboard.com/decision/FB-H6OZKDS3/.

———. 2021h. "Wampum Belt," December 19, 2021. https://www.oversightboard.com/decision/FB-L1LANIA7/.

———. 2022a. “Oversight Board publishes first Annual Report,” June 22, 2022. https://www.oversightboard.com/news/322324590080612-oversight-board-publishes-first-annual-report/.

———. 2022b. “Reclaiming Arabic words,” June 13, 2022. https://www.oversightboard.com/decision/IG-2PJ00L4T/.

———. 2023a. “Iran protest slogan,” January 9, 2023. https://www.oversightboard.com/decision/FB-ZT6AJS4X/.

———. 2023b. “Pro-Navalny protests in Russia,” May 26, 2023. https://www.oversightboard.com/decision/FB-6YHRXHZR/.

Parliament of Singapore, Online Safety (Miscellaneous Amendments) Act, pt. 45L. 26 (Oct. 3, 2022), https://www.parliament.gov.sg/docs/default-source/default-document-library/online-safety-(miscellaneous-amendments)-bill-28-2022.pdf.

Authors

Naomi Shiffman (nshiffman@osbstaff.com) leads the Data and Implementation Team at the Oversight Board. She received her Masters in Public Policy from the University of California, Berkeley in 2019.

Carly Miller (cmiller@osbstaff.com) is a Data and Implementation Officer at the Oversight Board. She received her Bachelor's degree in Political Science from the University of California, Berkeley.

Manuel Parra Yagnam (mparra@osbstaff.com) is the Deputy Head of the Data and Implementation Team at the Oversight Board. He received an MBA from Universidad de Chile and an MA in International Security from SciencesPo Paris.

Claudia Flores-Saviaga (cflores-saviaga@osbstaff.com) is a Data and Implementation Officer at the Oversight Board. She received a PhD in Computer Science from Northeastern University in 2022 and an MSc in Information Technology from Carnegie Mellon University in 2012.

Acknowledgements

We thank our colleagues on the Data and Implementation Team, Jacob Silver and Kyle Orangio, and colleagues across the Oversight Board for making this work and submission possible—in particular, thanks to Matthew Sells, Tracy Manners, and Thomas Hughes for their review. We would also like to thank our colleagues on the Meta Governance Team for their work on the Board's recommendations. Lastly, we would like to thank Rachel Wolbers for her invaluable feedback on our commentary.

Keywords

Regulation; transparency; social media platforms; data access.

Appendices

Appendix A: Iran Protest Slogan

Iran Protest Slogan Implementation Evaluation Methodology

Data was obtained using CrowdTangle, a public insights tool owned and operated by Meta. Posts from Instagram and Facebook were collected for the period from July 17, 2022, to October 17, 2022, both by Memetica (a research consultancy that works with the Board) prior to the case decision, and by the Board's Implementation Team following the decision's implementation. We made sure that only the posts with the slogan "ای خامنه بر مرگ" ("down with Khamenei") were included in all analyzed datasets. Specifically, the datasets collected were:

Facebook Datasets:

- A dataset provided by Memetica in October 2022, spanning July 17–October 17, 2022, of Facebook posts that included the slogan "ای خامنه بر مرگ." The dataset contains a total of 504 Facebook posts from 114 groups or pages.
- A dataset pulled by the Data and Implementation Team in May 2022 of Facebook posts that mention the slogan "ای خامنه بر مرگ." The dataset contains a total of 504 Facebook posts from 114 groups or pages. The dataset spans the same time period as Memetica's original dataset and contains a total of 488 Facebook posts from 104 groups or pages.

Instagram Datasets:

- A dataset provided by Memetica in October 2022, spanning July 17–October 17, 2022, of Facebook posts that included the slogan "ای خامنه بر مرگ." The dataset contains a total of 94 posts from 41 accounts.
- A dataset pulled by the Data and Implementation Team in May 2022 of Facebook posts that include the slogan "ای خامنه بر مرگ." The dataset spans the same time period as Memetica's original dataset and contains a total of 102 posts from 43 accounts.

Assumptions

It is important to note that since CrowdTangle was the only source of information we had available to obtain the data, we filtered our datasets on the basis of the following assumptions:

- We discarded the groups, pages, and accounts that appeared in Memetica's datasets but did not appear in our datasets. This is because we are unable to account for what happened to the missing posts, as they could have been removed for different reasons (e.g., users could have taken them down). Overall, we found 14 groups with 22 posts on Facebook and 10 accounts with 10 posts on Instagram that met this assumption and eliminated them from the original Memetica datasets. This means that both our dataset and Memetica's original datasets now contained the same groups, pages, and accounts.
- We considered posts that appear in our dataset but do not appear in Memetica's datasets to have been restored by Meta as a result of implementing the recommendation. Even though this assumption has its limitations, it is the only way we have to infer that certain posts were restored by Meta.

Following our assumptions for filtering out the datasets, we uncovered 38 posts that appeared in our data and not Memetica's, suggesting they were restored in the time since Memetica's October 2022 analysis. Twenty of those posts were on Facebook, mostly stemming from one anti-Khamenei group with 55,000 members. The remaining 18 posts were from an array of Instagram pages and have earned a cumulative 111,000 interactions on the platform. These posts contain a mix of support for, coverage of, and commentary on protests against the Iranian government.

Furthermore, our dataset was missing 14 posts that were present in the Memetica dataset. We cannot ascertain whether these posts were removed in the time since Memetica's analysis and never reinstated by Meta, whether the users removed the posts themselves, or if there is another explanation for their absence. Nevertheless, we made the choice to incorporate them into our statistical analysis because they were initially part of the Memetica dataset. By including them, we ensure that our analysis captures the overall impact of the recommendation implementation, taking into account any potential changes in the post landscape since the Memetica analysis.

Facebook Data Analysis

This first analysis aims to investigate whether there was a statistically significant difference in the number of posts restored in the different Facebook groups and pages contained in the datasets, derived from implementing the Board's Iran Protest Slogan recommendation no. 3 (Facebook datasets).

To determine the statistical difference, we employed Wilcoxon Signed-Rank Test, a nonparametric test suitable for analyzing paired data (Hayes 2023). This choice was made considering the small sample size and the non-normal distribution of the data. A significance level (α) of 0.05 (5%) was chosen to assess the results.

Finding

The Wilcoxon test analysis revealed that there was no statistically significant difference at the 5% level in the number of posts in the groups and pages in the Memetica dataset and our dataset on Facebook ($p = 0.9322$, $W = 0$, $\alpha = 0.05$). Consequently, we cannot conclude that the observed changes in the number of posts are directly attributed to Meta implementing the recommendation. The lack of statistical significance suggests that factors other than the recommendation implementation may be responsible for the observed variations in the post numbers.

Instagram Data Analysis

This second analysis aims to investigate whether there was a statistical difference in the number of posts restored in the different Instagram accounts, derived from implementing the Board's Iran Protest Slogan recommendation no. 3 (Instagram datasets). To determine the statistical difference, we employed a Wilcoxon Signed-Rank Test, a nonparametric test suitable for analyzing paired data. This choice was made considering the small sample size and the non-normal distribution of the data. A significance level (α) of 0.05 (5%) was chosen to assess the results.

Finding

The Wilcoxon test analysis revealed that there was a statistically significant difference between the change in number of posts across the Instagram accounts in the Memetica dataset and our dataset ($p = 0.001977$, $W = 13$, $\alpha = 0.05$). This indicates that the

changes seen in the number of posts are unlikely to be solely attributed to random variation. Instead, they are likely to be a result of Meta implementing the Oversight Board's recommendation. However, it is crucial to acknowledge that assumptions were made that in turn led to these results.

Conclusion

In this study, we ran a data analysis to investigate whether there was a statistical difference in the number of posts restored on Facebook and Instagram, derived from implementing the Oversight Board's recommendation in the Iran Protest Slogan case. Our analysis showed that the effects of implementing the recommendation by Meta are statistically significant for Instagram but not for Facebook. However, it is crucial to interpret the results cautiously, considering the limitations arising from the inadequate availability of suitable data and the assumptions made to be able to perform this analysis. For instance, it is important to note that we cannot confidently attribute the observed increase in the number of posts in our dataset solely to restorations following the implementation of the recommendation. Furthermore, it should be noted that CrowdTangle only tracks public content (including pages and public groups and accounts). Consequently, it is not possible to evaluate the impact of Meta implementing the recommendations on private groups and users on Instagram and Facebook. This limitation restricts the extent to which our findings can be generalized to the overall user base affected by the recommendation. Additionally, as our only source of information was CrowdTangle, we could only make certain assumptions with the public data obtained, but we acknowledge that other factors may influence the observed differences.