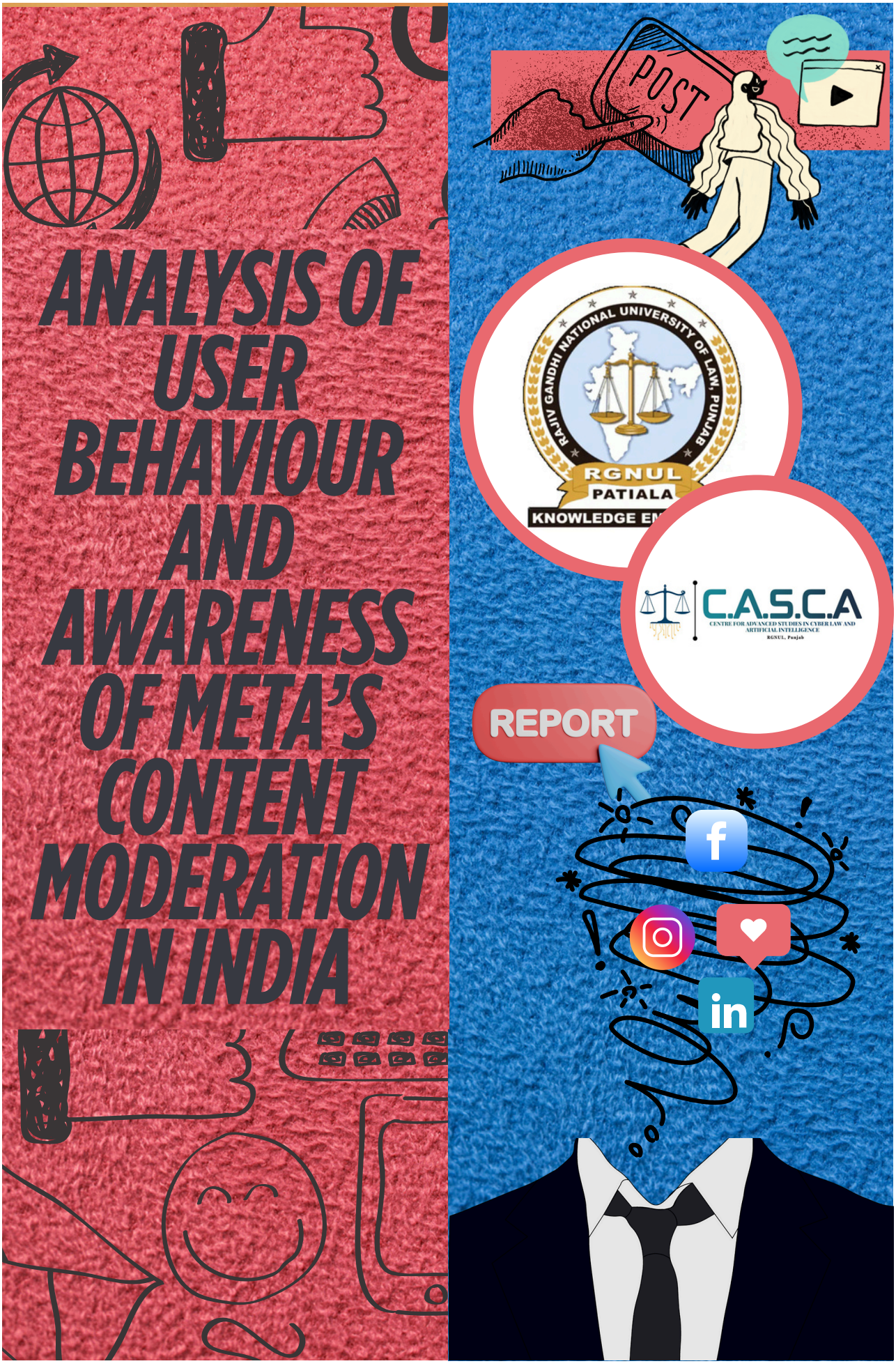
ANALYSIS OF USER BEHAVIOUR AND AWARENESS OF META'S CONTENT MODERATION IN INDIA



ANALYSIS OF USER BEHAVIOUR AND AWARENESS OF META'S CONTENT MODERATION IN INDIA

Authors:
Tanmay Durani, Sanskriti Koirala, Amishi Jain, Uday Gupta, R. Dayasakthi

Research Consultant:
Dr. Ivneet Walia, Associate Professor of Law and Officiating Registrar, RGNUL

CENTRE FOR ADVANCED STUDIES IN CYBER LAW AND ARTIFICIAL INTELLIGENCE

CASCA is a research-driven centre at RGNUL dedicated to advancing scholarly research and discourse in the field of Technology Law and Regulation. As a research centre of a leading institution in India, we are committed to promoting interdisciplinary research, fostering collaboration, and driving innovation in the fields of cyber law, artificial intelligence, and other allied areas.

FOR MORE INFORMATION:

Visit cascargnul.com

Introduction

This report presents the findings of a survey undertaken by the Centre for Advanced Studies in Cyber Law and Artificial Intelligence (CASCA) to examine user interaction with Meta's content moderation processes in India. The survey sought to assess three interlinked aspects of platform governance:

(i) user awareness and accessibility of the reporting function across Facebook, Instagram, and Threads;

(ii) user experience and perceptions of Meta's internal review system; and

(iii) the extent of knowledge among the general populace regarding the role and functioning of the Meta Oversight Board.

A total of **299 responses** were collected over a two-week period. This provides a valuable indication of trends in user perception and behaviour. The findings therefore highlight important structural concerns and shed light on issues of trust, transparency, and accessibility within Meta's content moderation framework.

The purpose of this report is fourfold:

1. To identify the challenges faced by users at the initial reporting stage and evaluate the transparency of feedback mechanisms;
2. To analyse levels of awareness regarding Meta's internal review and appeal mechanisms;
3. To assess user knowledge and understanding of the Oversight Board, particularly within the Indian context, and
4. To propose recommendations aimed at enhancing accountability, user trust, and accessibility in Meta's reporting and review systems.

What is Content Moderation?

Content moderation may be understood as the systematic exercise of regulating user-generated content in the digital space to ensure that such content conforms to certain prescribed standards, norms, or legal requirements. It is essentially a mechanism by which harmful, unlawful, or otherwise undesirable material is monitored, restricted, or removed. This process operates on two broad levels:

a. Micro Level: Platform-based Moderation

At this level, moderation is undertaken directly by social media platforms themselves. Each platform sets out its community guidelines or terms of service which all users are expected

to follow. To enforce these rules, platforms employ a combination of automated systems, artificial intelligence tools, and human moderators. For instance, a platform like Facebook may use automated algorithms to detect and remove hate speech or graphic violence that contravenes its content policy. Similarly, platforms may suspend or permanently ban accounts that engage in persistent misinformation or targeted harassment. This form of moderation is primarily private in nature, arising out of the contractual relationship between the platform and its users.

b. Macro Level: State-based Moderation

Beyond self-regulation by platforms, governments also exercise authority over online content through laws and regulations. This constitutes the macro-level framework, where the State, pursuant to its legislative powers, prescribes standards that platforms must legally comply with. For example, under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 in India, social media intermediaries are required to remove unlawful content upon receiving directions from a competent authority or court order. Similarly, in the European Union, the Digital Services Act imposes strict obligations on very large online platforms to swiftly act against illegal content and ensure algorithmic transparency.

The present survey is scoped towards examining the accessibility and efficacy of self-regulation by platforms, particularly their in-built reporting mechanisms, and whether such mechanisms are being realised to their full potential or remain underutilized in practice.

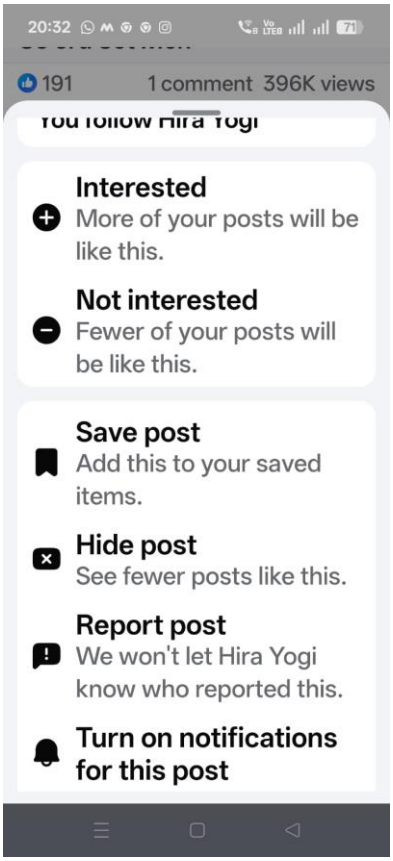
2. The Initial Stage: Hurdles in Reporting & Implications

In an effort to moderate user content, Meta social apps (Facebook, Instagram and Threads) allow users to report content that they see on their timeline (called "Feed" on Facebook[1] and Instagram[2], and "For you"/"Following" on Threads[3]), as well as in the search pages. This report button can be found in the options for each post, popularly called the "hamburger menu" in the apps. (see *Figures 1, 2, 3)*

[1] "How feed works" (*Facebook Help Centre,* 30 June 2018) <https://www.facebook.com/help/1155510281178725/> accessed 7 September 2025

[2] "How Instagram Feed Works" (*Instagram Help Centre,* 30 June 2018) <https://help.instagram.com/1986234648360433> accessed 7 September 2025

[3] "About your feeds on Threads" (*Instagram Help Centre,* 30 June 2018) <https://help.instagram.com/2249022798820145/?helpref=related_articles> accessed 7 September 2025



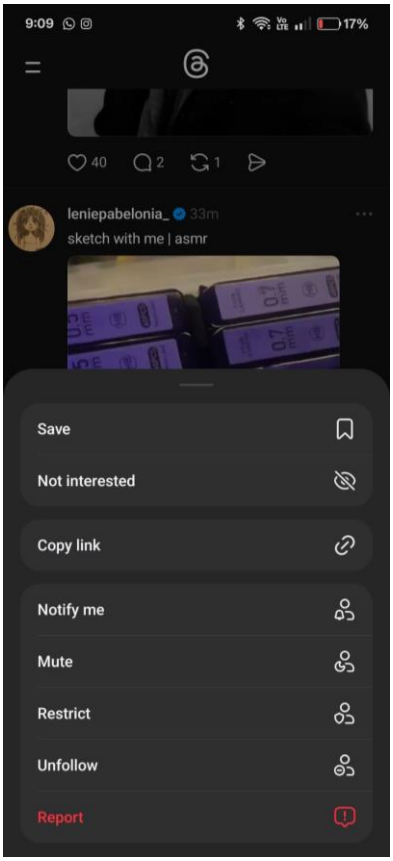


Figure 1 *Figure 2* *Figure 3*

The report button is highlighted in red on Instagram (see *Figure 2*) and Threads (see *Figure 3*), which helps with visibility and stands out from the other options. **Currently on Facebook, the report button blends with the other options** (see *Figure 1*)

Our survey shows that 93% of respondents are aware of the reporting function (see *Figure 4)*. However, only 73.6% of respondents have ever reported content on Meta platforms (see *Figure 5)*, with the reasons for not reporting every piece of content they deem clearly offensive or hateful mainly being a **lack of trust in the platform** and the belief that their action would not make a difference (see *figure 8)*.

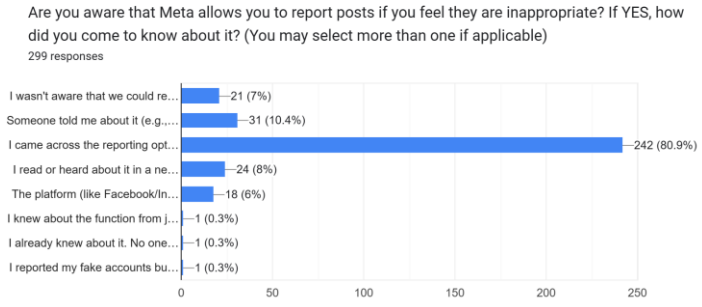


Figure 4

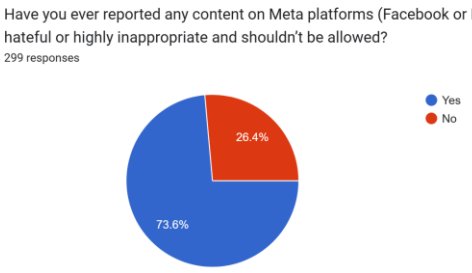


Figure 5

Additionally, our survey shows that 91.6% (see *Figure 6)* of respondents encounter posts on Meta platforms that they consider hateful, harmful, or otherwise offensive at least a few times each month. Yet, only 34.8% reported that their action of flagging such posts actually led to the removal or restriction of someone else's content (see *Figure 7)*. Many respondents also noted that they were not clearly informed about the outcome of their reports. This **gap between the frequency of harmful content and the limited effectiveness and transparency of the reporting process** highlights a structural problem that the next section explores in greater detail.

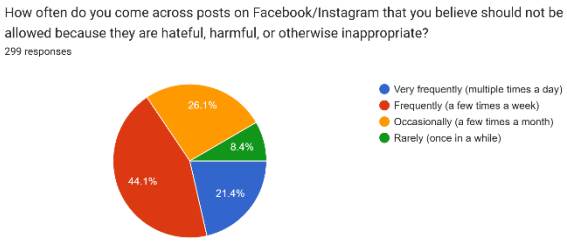


Figure 6

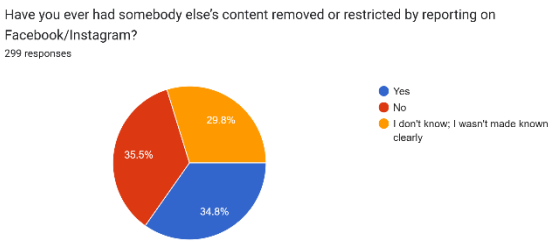


Figure 7

Currently, Meta's content policies involve a technology - first approach[4], and remove a large amount of violating content before anyone sees it.[5] If the technology misses something or requires more input, users also act as reviewers to enforce their Community Standards and Community Guidelines. For a report to be reviewed by a human, there are 3 prioritising factors[6]: Severity, Virality, and Likelihood of Violating. The decisions taken by these reviewers are used to train Meta's technology, allowing for new trends of violating content to be removed in an adversarial manner, so either the violating content stays on the platform, or it is proactively removed by Meta's evolving technology.

3. The Black Box: What Happens After Your Report?

Meta's content moderation system currently works in a three-step manner:

Step 1: Report the Content
Step 2: Request a Review of Meta's Decision ("Review")
Step 3: Appeal to the Oversight Board ("Appeal")

The first difficulty that many users encounter is simply locating the "Report" feature across Meta's platforms. This challenge is compounded by the lack of clarity in how users are updated on the status of their reports. By "status of reports," we refer to the information communicated to a user after submitting a complaint: *whether the report has been reviewed, what action (if any) has been taken, and the rationale behind that decision.*

In this regard, providing timely and transparent updates on report status is essential. It fosters user trust by demonstrating accountability, offers clarity on the actions taken, and promotes a safer environment by enabling users to take further steps if needed. In the absence of such updates, users often feel ignored, leading to frustration and a perception of opacity in Meta's moderation system.

---

[4] Jerry King and Kate Gotimer, "How We Review Content" (*Meta Newsroom,* August 11, 2020) <https://about.fb.com/news/2020/08/how-we-review-content/#:~:text=we%E2%80%99re%20going%20to%20use%20our%20automated%20systems%20first%20to%20review%20more%20content%20across%20all%20types%20of%20violations.> accessed 7 September 2025

[5] "How review teams work" *Meta Transparency Center* (November 12, 2024) <https://transparency.meta.com/en-gb/enforcement/detecting-violations/how-review-teams-work/#:~:text=we%20use%20artificial%20intelligence%20and%20machine%20learning%20tools%20to%20help%20us%20identify%20and%20remove%20a%20large%20amount%20of%20violating%20content%20%E2%80%93%20often%2C%20before%20anyone%20sees%20it.> accessed 7 September 2025

[6] "How Meta prioritises content for review" (*Meta Transparency Center,* 12 November 2024) <https://transparency.meta.com/en-gb/policies/improving/prioritizing-content-review/#:~:text=Prioritising%20factors%20for%20human%20review> accessed 7 September 2025

Survey Results:

The survey conducted by CASCA clearly indicates that, even when Meta users come across harmful content on the platform, they do not always submit a report against such posts. 64.2% Respondents have indicated their lack of belief in Meta acting upon the report submitted by them. Similarly, another 30.8% that they lacked trust that Meta would process their review requests and initial reporting in a fair manner. (see *Figure 8*)

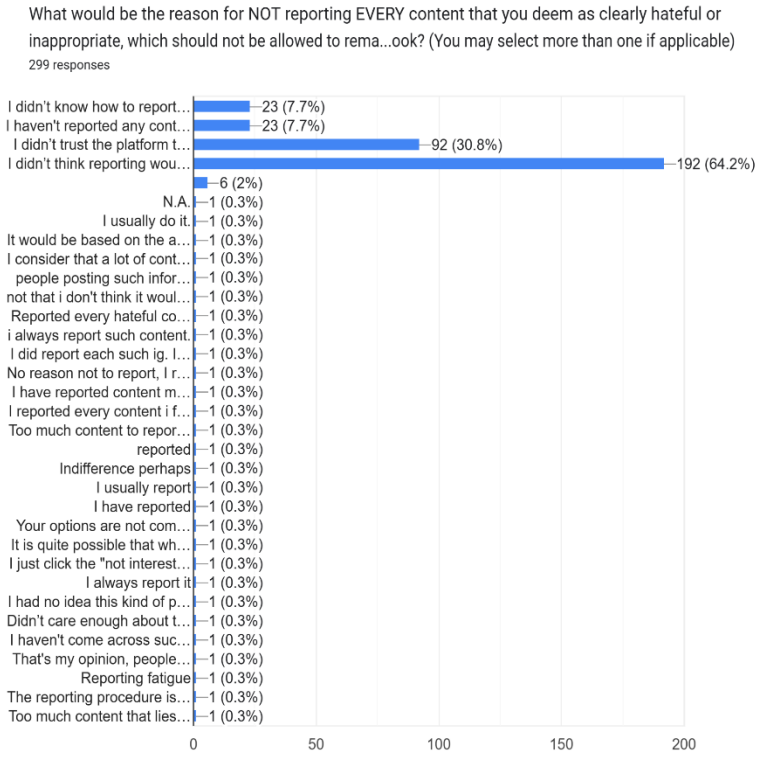


Figure 8

Further, the findings indicate that 50.2% of respondents did not perceive any notification regarding the status of their filed report *(see Figure 9)*. This suggests that while such notifications may exist, they are not sufficiently noticeable to users, thereby reinforcing a link between inadequate visibility of information and diminished trust. This reflects a significant disconnect between user expectations and the platform's moderation practices.

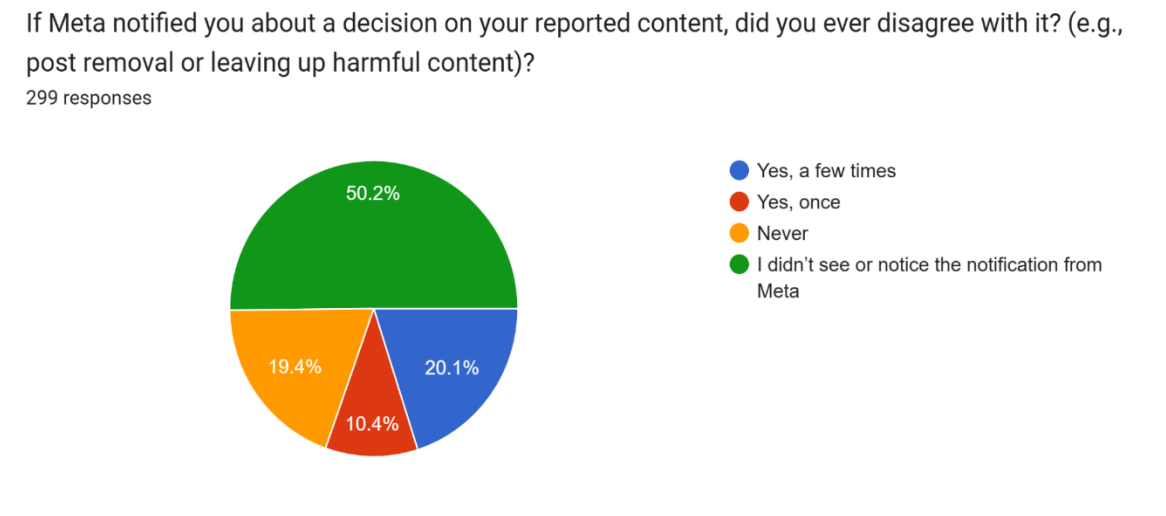
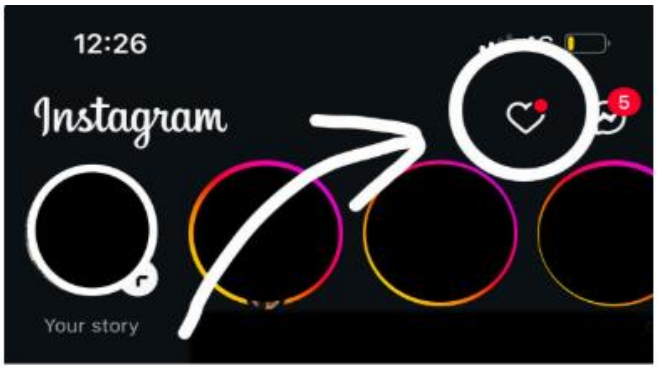


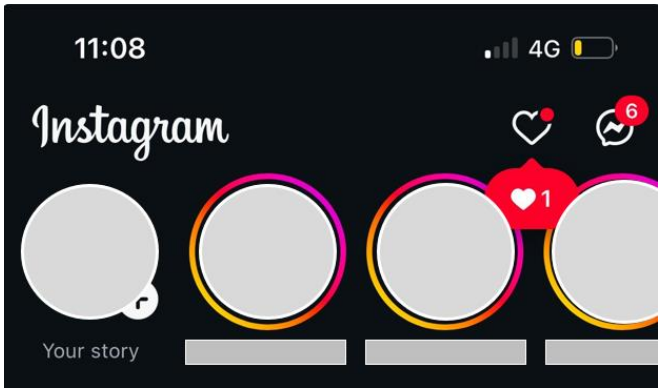
Figure 9

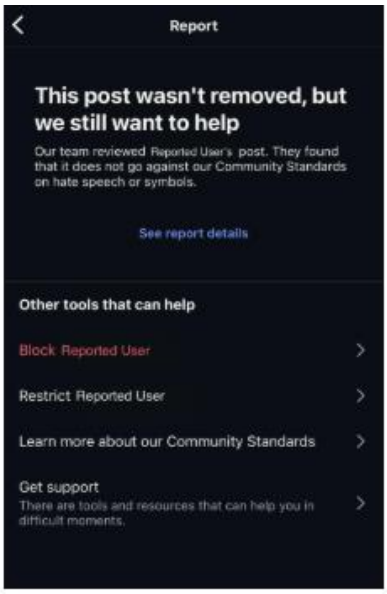
Limitations of the Meta's present mechanism:

In the status quo, Meta does fulfill its responsibility to inform its users of the status of the report filed by them against any content on the platform. Once Meta reaches its first decision, it is notified to the user; however, accessing this information becomes a tedious exercise. The images presented below indicate the manner in which Instagram's portal currently updates its users on the status of report(s) filed by them.

Any update on a user's report is initially indicated on the platform's notification pane (*Figure 10*). This is followed by a subsequent before the user finally reaches the update sought (*as shown in Figures 11, 12):*







(*Figure 10*) (*Figure 11*) (*Figure 12*)

The existing interface and placement of the content review mechanism on the platform make it difficult for users to discover this option, which is evidenced from the fact that **more than 60% of respondents stated they were unaware of their ability to request a review** of reported content (see *Figure 13)*. This lack of user awareness is a significant barrier to effective content moderation, as it prevents a large portion of the user base from engaging with the system and exercising their right to appeal.

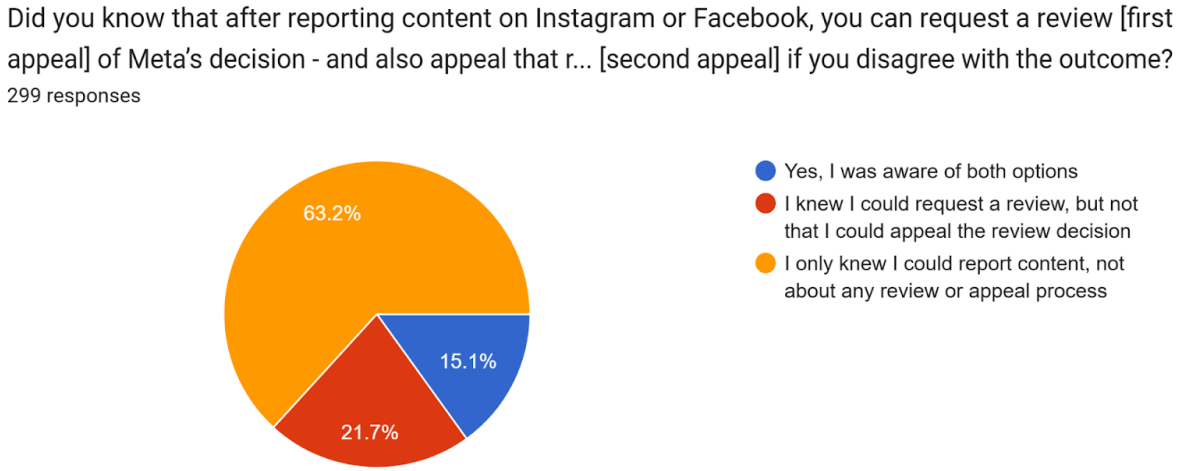


Figure 13

The difficulty in locating the various steps into filing a report and further with a review, **acts as a deterrent**, resulting in the platform users being unaware (due to difficulty in locating), or being discouraged (due to the amount of effort to file for an appeal, merely for content moderation).

The manner in which this information is displayed has a three-fold problem:

Firstly, initial notification through the small red dot on **the heart icon is not prominent and could be overlooked** by users. The current notification system, which utilizes a red dot, often fails to adequately signal new information, leading to users overlooking notifications on status updates. In contrast, Meta's notification for a "like" on a post is demonstrably more effective due to its distinct visual prominence (See *Figure 10*).

Secondly, the effectiveness of status update notifications may be compromised by their low visual saliency *(see Figure 11)*. The use of similar color schemes across multiple notifications causes the status update to become camouflaged, resulting in user oversight. Given that content moderation is a core priority for the platform, such a choice of interface undermines the communication of critical information and should not be continued in its present form.

Thirdly, the status update on a report filed is titled a "Support Request", which further leads to users overlooking the information as the vagueness of the words do not amply point towards any update on status of the report itself, which is important because the users may miss critical progress updates, assume their report has not been acted upon, or fail to follow up appropriately, thereby reducing overall trust and engagement with the system.

4.2. The Unknown Supreme Court: Appeal before the Meta Oversight Board

At the third stage of Meta's content moderation process, users may challenge Meta's internal review decision by filing an appeal before the Meta Oversight Board. Initially, the Oversight Board emerged as a product of backlash and unaccountability on its part. Its formation was shaped by the reputational crisis it faced post-2016 US presidential elections, which highlighted how susceptible social media platforms are to exploitation and spread disinformation and misinformation.[7] Meta, consequently, suffered criticism for failing to curb electoral manipulation. This is linked to the 2018 Cambridge Analytica scandal which showcased how millions of users' personal data was harvested without consent for political advertising. Thus, better oversight was needed and Meta needed to rebrand itself as a responsible and accountable platform. Consequently, Meta Oversight Board was created to project transparency and responsibility in digital platform governance.

Established as an independent body, the Oversight Board has often been described as Meta's "Supreme Court" for content moderation. Its mandate is twofold:

(i) to act as the final appellate forum for individual users contesting Meta's decisions on content takedowns or restorations, and;

(ii) to issue broader policy recommendations to guide Meta in refining its content regulation models across different contexts.

Despite this ambitious design, the Oversight Board has faced sustained criticism, particularly regarding its limited accessibility and representation in the Global South. This delayed engagement raises **concerns of under-representation and undermines the perception of fairness** in how content disputes are prioritized globally. The Board, in its 2021 Annual Report, itself acknowledged that lower numbers of user appeals from outside Europe, the USA, and Canada could indicate that many of those using Facebook and Instagram in the rest of the world are not aware they can appeal Meta's content moderation decisions to the Board,[8] while raising concerns about whether Meta has invested sufficient resources in moderating content in languages other than English.[9]

Our survey data illustrates this disconnect sharply. While 33.8% of respondents had heard of the Oversight Board, they did not understand how it operates, and 54.2% had never heard

---

[7] Evelyn Douek, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (2024) 37(2) Harvard Journal of Law & Technology https://jolt.law.harvard.edu/assets/articlePDFs/v372/3-The-Meta-Oversight-Board-and-the-Empty-Promise-of-Legitimacy.pdf accessed 18 July 2010

[8] "Oversight Board Annual Report 2021" (*Meta Oversight Board*, 2022) https://www.oversightboard.com/wp-content/uploads/2023/11/Annual-Report-English.pdf accessed 18 September 2025

[9] BNS Economic Bureau, 'Non-English content moderation: Meta Oversight Board flags lack of investment' The Indian Express (New Delhi, 24 June 2022)

of it at all (see *Figure 14).* Further, 64.5% were unaware that they could appeal to the Board after exhausting Meta's internal review system (see *Figure 15*). These findings suggest that a significant proportion of Indian users, arguably among Meta's **most important stakeholders, remain uninformed about the very mechanism** designed to determine whether disputed content stays online or is removed.

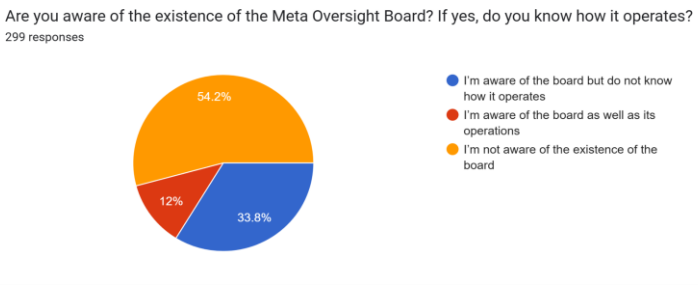


Figure 14

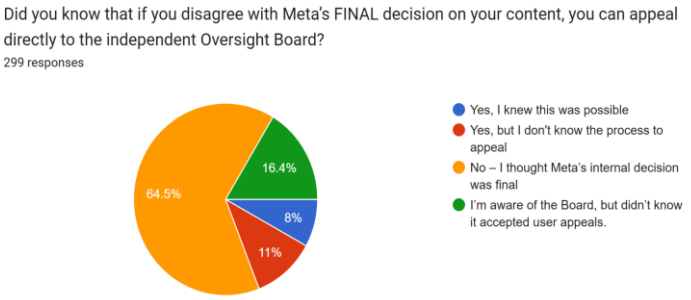


Figure 15

This lack of awareness and engagement raises a fundamental paradox: how can the Oversight Board operate as a "Supreme Court" for accountability and rights protection when its largest community of users neither knows of its existence nor understands how to access it? The legitimacy of the Board's decisions, especially their contextual relevance to India, is called into question if users remain structurally excluded from the process.

The challenge, therefore, is twofold: *first*, to examine why the Oversight Board remains an "unknown Supreme Court" for much of the Indian user base; and *second*, to propose concrete steps that Meta could adopt to close this gap and ensure that the Board's processes are accessible, transparent, and representative for users in India.

Subsequent to basic awareness of the Oversight Board's existence, a deeper problem emerges: **users are largely unaware of their right to escalate a rejected review decision by filing an appeal before the Oversight Board.**

The results show that only 15.1% of users were aware of both stages, the review of the first decision on our report (first appeal) and the appeal to the Oversight Board to challenge the appeal (second appeal) (see *Figure 13*). The fact that just 15% of respondents understood they could seek a review and, if rejected, escalate further, is not a minor communication gap but points to a systemic issue within Meta's processes. If users remain unaware of the possibility of appeal, the majority of moderation decisions are effectively insulated from independent scrutiny.

The low rate of appeals by Indian users therefore, stems from two interlinked factors:

1. Limited awareness of the Oversight Board's very existence.
2. Even lower awareness of the right to appeal after exhausting Meta's internal review system.

When the Oversight Board was established, it was widely hailed as a landmark in digital governance. It was designed to address longstanding criticisms of Meta's opaque, arbitrary content moderation by introducing an independent, rights-based appellate forum.

Yet, for the vast majority of Indian users, the Oversight Board is perceived, if at all, as a distant entity with little practical relevance. The promise of an independent appellate body has not been effectively communicated to the masses. As a result, the Oversight Board functions more as a symbolic innovation, respected in policy and academic discussions but disconnected from the daily realities of content moderation on Meta platforms.

The primary cause of this disconnect lies in Meta's communication and awareness strategy. The Oversight Board's launch was heavily publicized among policymakers, journalists, and international experts as evidence of Meta's commitment to platform governance. However, Meta has not matched this with direct, accessible communication to its user base, particularly in India. Users rarely receive a clear, simple explanation that certain contentious content decisions can be escalated beyond Meta's internal processes to an independent adjudicatory body. This failure **undermines both the purpose and legitimacy** of the Oversight Board.

Recommendations for Meta

Meta, despite having the largest content moderation operation, has procedural complexities in their current system that even the persistent users of such platforms cannot navigate. The fact that less than a quarter of users know only about the appeal procedure reflects that the present user interface does not effectively serve its objective, but not necessarily policy or strategy problems. In order for Meta to give meaningful checks regarding content being uploaded to social media platforms, it is foremostly required to bring clarity and simplicity into the process. For Meta users to have trust in the reporting process as well as awareness of the Oversight Board, both in Meta's ability to evaluate the report fairly, and for it to have a tangible impact on the social ecosystem of the platform.

I. Reporting Stage

At the stage of submitting reports, Meta will need to establish a transparent feedback mechanism to inform users about (1) the outcome of their reports, and (2) generally on why it matters for users to report harmful content.

This can be affected by Meta through,

1. **Inclusion of an advisory section:** Including a section in the settings that acts as an advisory page to make users aware of the benefits of reporting, the process of reporting, and the content review policies in a clear and concise manner. It is imperative that users understand these components and are aware of the process of reporting and why it is important.
2. **Public Service Announcements (PSAs)**: These could be done by Meta independently or through collaborations with social media influencers as intermediaries of communication. Here, it would be beneficial if Meta ensures diversity in the selection of influencers, particularly with respect to regional representation. This approach would help amplify the reach of awareness campaigns, which would subsequently increase the regional impact and overall effectiveness of the initiative.
3. **Communication through 'Stories' feature:** WhatsApp follows the practice of using the status feature to disseminate information about new features; Meta can adopt a similar communicative strategy on its social media platforms as well. Periodic stories could serve as a medium through which the platform informs users about its operational accountability, such as reporting progress on content moderation. (For example, the platform could occasionally publish updates in the form of stories with statements like "450 reports addressed in the past month.")

II. Tracking the Submitted Report: Review and Appeal :

The inherent problem with tracking Meta's content moderation is the complexity of the process. The complicated steps create a maze-like situation for users. The existing complexity in the review and appeal process can be addressed by Meta by shifting their primary focus from designing policy compliance to user clarity and ease.

The first step to this goal would be to have a pop-up whenever there is a decision to the reported content (like the pop-up for "sleep mode" [see *Figure 16*]), along with straightforward action buttons such as **"Accept", "Review"**, and **"Appeal this decision"**. Alternatively, there can be a full-screen dialog, similar to when there is a suspicious login attempt (see *Figure 17*), with the aforementioned action buttons.

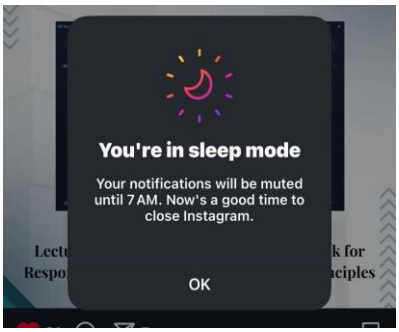


Figure 16 *Figure 17*

After these pop-ups are closed by the user, the decision of the report must remain on the notification pane as well, in a color distinct from other reports, such as likes and follows (or follow requests) from other uses. Reporting is a vital component of any social media platform, and must be displayed in a more unique manner.

The scheme of an internal review, further followed by an Oversight Board Review shall be explicitly informed by Meta. A simple pictorial/graphical pop-up stating **"Step 1: Report → Step 2 : Review → Step 3 : Appeal to the Oversight Board"**, the moment any user reports any content for the first few times, would clarify the journey. Similarly, the designed policy compliances can be linked for those who want to read it in detail, but the notification/inceptive communication should be written in user-friendly language clearly stating the reason behind the decision made and what the user can do about it. Further, in order to make the appeal system more user friendly, a tracking system could be made available to the users which would show the stage of appeal they have reached, and the expected timeline of decision, ensuring timely decisions and accountability.

III. Awareness of the Oversight Board

The primary shortcoming in incorporating the third step, i.e., appeal to the Oversight Board, effectively, the moot problem is the unknown-ness of the Meta Oversight Board. To undo the communication failure, it is imperative that the Oversight Board's purpose does not get limited to a public relations initiative but as an essential feature of its platform governance.

For the Oversight Board to succeed, Meta must take proactive steps to integrate it into the daily experience of users, especially in a country that hosts one of the largest users of any app of Meta.

Just as Meta invests in user onboarding for new features, such as beta experiences and occasional self-help check-ins, it should create accessible resources on how to file appeals. Infographics and easy-to-navigate support centres can ease the process.

Conclusion

The report illustrates the discrepancy that is present quite prominently and continues for a long time between the content moderation systems of Meta and the user experience in India. On the one hand, Meta has put together a moderation framework with several layers from initial reporting to an independent Oversight Board. On the other hand, our survey results signal that for a substantial portion of the Indian user base, these systems are not working as planned. The main reasons behind this are not technical ones but the lack of clarity, availability, and user confidence in the whole process. In order to cross this canyon, Meta needs to turn its attention away from just policy adherence towards a user-centric approach. The suggestions brought up in the report such as clear feedback mechanisms, intuitive user interfaces, and direct awareness campaigns using media like "Stories" and public service announcements are not only superficial changes. They are the prime steps to confirm that the whole moderation process does not become a "black box" but is a transparent, accessible system, which empowers the users as well as provides a safer online environment.