Compliance reports by social media platforms are unhelpful

GUEST AUTHOR *on* APRIL 18, 2022

8 MINUTE READ

By Vasudev Devadasan

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 ("Intermediary Guidelines") represents India's first attempt at regulating large social media platforms, with the Guidelines creating distinct obligations for 'Significant Social Media Intermediaries' ("**SSMIs**"). While certain provisions of the Guidelines concerning SSMIs (like the traceability requirement) are currently under legal challenge, the Guidelines also introduced a less controversial requirement that SSMIs publish monthly transparency reports regarding their content moderation activities. While this reporting requirement is arguably a step in the right direction, scrutinising the actual documents published by SSMIs reveals a patchwork of inconsistent and incomplete information – suggesting that Indian regulators need to adopt a more comprehensive approach to platform transparency.

This post briefly sets out the reporting requirement under the Intermediary Guidelines before analysing the transparency reports released by SSMIs. It highlights how a focus on figures coupled with the wide discretion granted to platforms to frame their reports undermines the goal of meaningful transparency. The figures referred to when analysing SSMI reports pertain to the February-March 2022 reporting period, but the distinct methodologies used by each SSMI to arrive at these figures (more relevant for the present discussion) have remained broadly unchanged since reporting began in mid-2021. The post concludes by making suggestions on how the Ministry of Electronics and Information Technology ("**MeitY**") can strengthen the reporting requirements under the Intermediary Guidelines.

Transparency reporting under the Intermediary Guidelines

Social media companies structure speech on their platforms through their content moderation policies and practices, which determine when content stays online and when content is taken down. Even if the content is not illegal or taken down pursuant to a court or government order, platforms may still take it down for

violating their terms of service (or Community Guidelines) (let us call this content 'violative content' for now i.e., content that violates terms of service). However, ineffective content moderation can result in violative and even harmful content remaining online or non-violative content mistakenly being taken down. Given the centrality of content moderation to online speech, the Intermediary Guidelines seek to bring some transparency to the content moderation practices of SSMIs by requiring them to publish monthly reports on their content moderation activities. Transparency reporting helps users and the government understand the decisions made by platforms with respect to online speech. Given the opacity with which social media platforms often operate, transparency reporting requirements can be an essential tool to hold platforms accountable for ineffective or discriminatory content moderation practices.

Rule 4(1)(d) of the Intermediary Guidelines requires SSMIs to publish monthly transparency reports specifying: (i) the details of complaints received, and actions taken in response, (ii) the number of "*parts of information*" proactively taken down using automated tools; and (iii) any other relevant information specified by the government. The Rule therefore covers both 'reactive moderation', where a platform responds to a user's complaints against content, and 'proactive moderation', where the platform itself seeks out unwanted content even before a user reports it.

Transparency around reactive moderation helps us understand trends in user reporting and how responsive an SSMI is to user complaints, while disclosures on proactive moderation shed light on the scale and accuracy of an SSMI's independent moderation activities. A key goal of both reporting datasets is to understand whether the platform is taking down as much harmful content as possible without accidentally also taking down non-violative content. Unfortunately, Rule 4(1)(d) merely requires SSMIs to report the number of links taken down during their content moderation (this is re-iterated by the MeitY's FAQs on the Intermediary Guidelines). The problems with an overt, simplistic approach come to the fore upon an examination of the actual reports published by SSMIs.

Contents of SSMI reports – proactive moderation

Based on its latest monthly transparency reports, Twitter proactively suspended 39,588 accounts while Google used automated tools to remove 338,938 pieces of content. However, these figures only document the scale of proactive monitoring and do not provide any insight into the accuracy of the platforms' moderation – how accurate is the moderation in distinguishing between violative and non-violative content. The reporting also does not specify whether this content was taken down using solely automated tools, or some mix of automated tools and human review or oversight. Meta (reporting for Facebook and Instagram) reports the volume of content proactively taken down, but also provides a "Proactivity Rate". The Proactivity Rate is defined as the percentage of content flagged proactively (before a user reported it) as a subset of all flagged content. Proactivity Rate = [proactively flagged content ÷ (proactively flagged content + user reported content)]. However, this metric is also of little use in understanding the accuracy of Meta's automated tools. Take the following example:

Assume a platform has 100 pieces of content, of which 50 pieces violate the platforms' terms of service and 50 do not. The platform relies on both proactive monitoring through automated tools and user reporting to identify violative content. Now, if the automated tools detect 49 pieces of violative content, and a user reports 1, the platform states that: '49 pieces of content were taken down pursuant to proactive monitoring at a Proactivity Rate of 98%'. However, this reporting does not inform citizens or regulators: (i) if the 49 pieces of content identified by the automated tools are in fact the 49 pieces that violate the platform's terms of service (or whether the tools mistakenly took down some legitimate, non-violative content); (ii) how many users saw but did not report the content that was eventually flagged by automated tools and taken down; and (iii) what level and extent of human oversight was exercised in removing content. A high proactivity rate merely indicates that automated tools flagged more content than users, which is to be expected. Simply put, numbers aren't everything, they only disclose the scale of content moderation and not its quality.

This criticism begs the question, **how do you understand the quality of proactive moderation?** The Santa Clara Principles represent high-level guidance on content moderation practices developed by international human rights organisations and academic experts to facilitate platform accountability with respect to users' speech. The Principles require that platforms report: (i) when and how automated tools are used; (ii) the key criteria used by automated tools in making decisions; (iii) the confidence, accuracy, or success rate of automated tools, including in different languages; (iv) the extent of human oversight over automated tools; and (v) the outcomes of appeals against moderation decisions made by automated tools. This last requirement of reporting the outcome of appeals (how many users successfully got content reinstated after it was taken down by proactive monitoring) is a particularly useful metric as it provides an indicator of when the platforms themselves acknowledge that its proactive moderation was inaccurate. Draft legislation in Europe and the United States requires platforms to report how often proactive monitoring decisions are reversed. Mandating the reporting of even some of these elements under the Intermediary Guidelines would provide a clearer picture of the accuracy of proactive moderation.

Finally, it is relevant to note that Rule 4(4) of the Intermediary Guidelines requires that the automated tools for proactive monitoring of certain classes of content must be 'reviewed for accuracy and fairness'. The desirability of such proactive monitoring aside, Rule 4(4) is not self-enforcing and does not specify who should undertake this review, how often it should be carried out, and to whom the results should be communicated.

Contents of SSMI reports – reactive moderation

Transparency reporting with respect to reactive moderation aims to understand trends in user reporting of content and a platform's responses to user flagging of content. Rule 4(1)(d) requires platforms to disclose the "details of complaints received and actions taken thereon". However, a perusal of SSMI reporting reveals

how the broad discretion granted to SSMIs to frame their reports is undermining the usefulness of the reporting.

Google's transparency report has the most straightforward understanding of "complaints received", with the platform disclosing the number of 'complaints that relate to third-party content that is believed to violate local laws or personal rights'. In other words, where users raise a complaint against a piece of content, Google reports it (30,065 complaints in February 2022). Meta on the other hand only reports complaints from (i) a specific contact form, a link for which is provided in its 'Help Centre'; and (ii) complaints addressed to the physical post-box mail address published on the 'Help Centre'. For February 2022, Facebook received a mere 478 complaints, of which only 43 pertained to content (inappropriate or sexual content), while 135 were from users whose accounts have been hacked, and 59 were from users who had lost access to a group or page. If 43 user reports a month against content on Facebook seems suspiciously low, it likely is – because the method of user reporting of content that involves the least amount of friction for users (simply clicking on the post and reporting it directly) bypasses the specific contact form that Facebook uses to collate India complaints and thus appears to be absent from Facebook's transparency reporting. Most of Facebook's 478 complaints in February have nothing to do with the content on Facebook and offer little insight into how Facebook responds to user complaints against content or what types of content users report.

In contrast, Twitter's transparency reporting expressly states that it does not include non-content related complaints (e.g., a user locked out of their account), instead of limiting its transparency reporting to content-related complaints – 795 complaints in March 2022: 606 of abuse or harassment, 97 of hateful conduct, and 33 of misinformation were the top categories. However, like Facebook, Twitter also has both a 'support form' and allows users to report content directly by clicking on it but fails to specify from what sources "complaints" are compiled for its India transparency reports. Twitter merely notes that 'users can report grievances by the grievance mechanism by using the contact details of the Indian Grievance Officer'.

These apparent discrepancies in the number of complaints reported bear even greater scrutiny when the number of users of these platforms is factored in. Twitter (795 complaints/month) has an estimated 23 million users in India while Facebook (406 complaints/month) has an estimated 329 million users. It is reasonable to expect user complaints to scale with the number of users, but this is evidently not happening to suggest that these platforms are using different sources and methodologies to determine what constitutes a "complaint" for the purposes of Rule 4(1)(d). This is perhaps a useful time to discuss another SSMI, ShareChat.

ShareChat is reported to have an estimated 160 million users, and for February 2022, the platform reported 56,81,213 user complaints (substantially more than Twitter and Facebook). These complaints are content-related (e.g., hate speech, spam etc.) although, with 30% of complaints merely classified as 'Others', there is some uncertainty as to what these complaints pertain to. ShareChat's reports state that it collates complaints from 'reporting mechanism across the platform'. This would suggest that, unlike Facebook (and potentially Twitter), it compiles user complaint numbers from all methods a user can complain against content and not just a single form tucked away in its help centre documentation. While this may be a more holistic approach, ShareChat's reporting suffers from other crucial deficiencies. Sharechat's reports make no distinction between reactive and proactive moderation, merely giving a figure for content that has been taken down. This makes it hard to judge how ShareChat responded to these over 56,00,000 complaints.

Conclusion

Before concluding, it is relevant to note that no SSMI reporting discusses content that has been subjected to reduced visibility or algorithmically downranked. In the case of proactive moderation, Rule 4(1)(d) unfortunately limits itself to content that has been "removed", although, in the case of reactive moderation, reduced visibility would come within the ambit of 'actions taken in response to complaints' and should be reported on. Best practices would require platforms to disclose when and what content is subjected to reduced visibility to users. Rule 4(1)(d) did not form

part of the draft intermediary guidelines that were subjected to public consultation in 2018, rather appearing for the first time in its current form in 2021. Ensuring broader consultation at the time of drafting may have resulted in such regulatory lacunae being eliminated and a more robust framework for transparency reporting.

That said, getting meaningful transparency reporting is a hard task. Standardising reporting procedures is a detailed and fraught process that likely requires platforms and regulators to engage in a consultative process – see this document created by Daphne Keller, listing out potential problems in reporting procedures. Sample problem: "If ten users notify platforms about the same piece of content, and the platform takes it down after reviewing the first notice, is that ten successful notices, or one successful notice and nine rejected ones?" Given the scale of the regulatory and technical challenges, it is perhaps unsurprising that the transparency reporting under the Intermediary Guidelines has gotten off to a rocky start. However, Rule 4(1)(d) itself offers an avenue for improvement. The Rule allows the MeitY to specify any additional information that platforms should publish in their transparency reports. In the case of proactive monitoring, **requiring platforms to specify exactly how automated tools are deployed, and when content takedowns based on these tools are reversed would be a good place to start.** The MeitY must also engage with the functionality and internal procedures of SSMIs to ensure that reporting is harmonised to the extent possible. For example, reporting a "complaint" for Facebook and ShareChat should ideally have some equivalence. This requires, for a start, MeitY to consult with platforms, users, civil society, and academic experts when thinking about transparency.

*

Vasudev Devadasan is a researcher at the Centre for Communication Governance (CCG-NLU Delhi). This article has been cross-posted with permission and the original post can be found here.

Also Read:

- WhatsApp Says It Banned 3 Million Accounts In India But The Actual Number Could Be More
- How The IT Rules FAQs Add To The Arbitrariness And Confusion Around The Rules

Have something to add? Subscribe to MediaNama here and post your comment.

Support our journalism:

For You

- **Sign up for our Daily Newsletter to receive regular updates**
- **Stay informed about MediaNama events**
- **Have something to tell us? Leave an Anonymous Tip**
- **Ask us to File an RTI**
- **Sponsor a MediaNama Event**

DISCOVER MORE

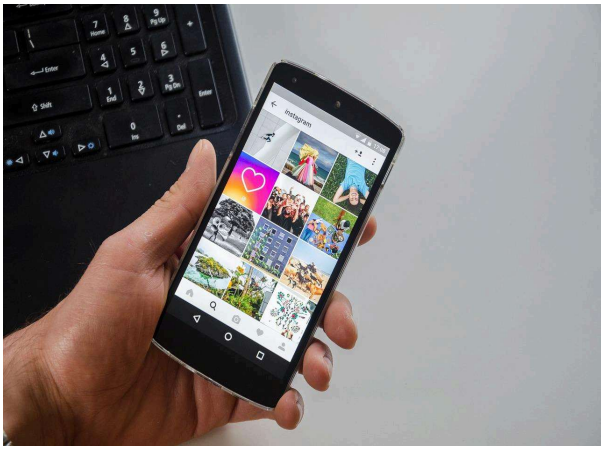
ccgnlu | compliance report | content moderation | free reads | it rules | online content moderation | platform regulation | significant social media intermediaries | transparency report | views

Related Posts:

Google sees slight increase in user complaints in December 2021, reveals compliance report

Google received 3,000 fewer user complaints in February 2022, compliance report shows

Summary: Tech companies adopt self-regulatory code to tackle harmful online content in New Zealand



Instagram beats own record on number of user complaints in a month

Google continues to witness a rise in number of user complaints in January 2022, compliance report shows

Instagram received a record number of user grievances in January, latest compliance report shows

MEDIANAMA

MediaNama is the premier source of information and analysis on Technology Policy in India. More about MediaNama, and contact information, here.