

OCTOBER 2024



SOCIAL MEDIA TRANSPARENCY REPORTING:

A PERFORMANCE REVIEW

AUTHORED BY
RAKESH MAHESHWARI
ANNA LIZ THOMAS
SOUMYA AK



IGIP

Executive Summary	01
--------------------------	-----------

Introduction	06
---------------------	-----------

Part A- Effectiveness of reporting mechanism established under Part-II of the IT Rules	09
---	-----------

- Details on complaints received
- Action taken on complaints received under different heads
- Links disabled due to proactive monitoring using automated tools
- Disclosure of grievance redressal officer
- Presence of separate India grievance mechanism, or in-feed grievance mechanism, whether appropriate disclosures have been made for both such mechanisms
- Improvement in use of grievance redressal mechanism over time
- Consistency of monthly compliance reports
- Consistency in access and ease of access to monthly compliance reports
- Common content related grievances
- Publishing of Reports detailing compliance with GAC orders

Part B- Comparison of disclosures made by SSMLs in their monthly compliance reports. 25

- Categories of content-related grievances
 - Categories of non-content related grievances
 - Information on law-enforcement requests
 - Information on content taken down, distinguished by language
 - Comprehensive disclosures on actions taken for problematic content
 - Disclosure on actions taken for repeated violations
 - Disclosure on appeals made regarding content moderation decisions made by the platform
 - Disclosure of proactive monitoring using automated tools, and the number of content categories being proactively monitored
-

Part C- Disclosures required by SSMLs within other regions and jurisdictions. 32

Conclusions and Recommendations 34

Annexure 1 38

Executive Summary

Background

In February 2021, the Central Government published the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (IT Rules, 2021). Part-II of these Rules (now onwards referred to as Intermediary Rules) mandated the publication of monthly compliance reports by 'significant social media intermediaries' ("SSMIs").

Rule 4(1)(d) of the Intermediary Rules requires SSMIs to publish a monthly compliance report mentioning

- a) the details of complaints received;
- b) actions taken thereon;
- c) number of communication links or parts of information that the SSMI has removed, or disabled access to, in pursuance of any proactive monitoring conducted by using automated tools; and
- d) any other information as may be specified.

In November 2021, the FAQs released by Ministry of Electronics, Information and Technology (MeitY) further clarified the compliance obligation under Rule 4(1)(d) for SSMIs by requiring the following:

1. Numbers of communication links removed by the SSMI to fulfil the requirement to report voluntary actions taken by an SSMI; and
2. Summary details of complaints received (for example, the subject under which the complaint was received) and action taken under each of the different heads. This information could be disclosed in the aggregated form, without disclosing granular details of all cases.

The compliance reports are (broadly) required to discuss how SSMIs tackle complaints and proactively disable or remove content within their platforms. It is assumed that the intent behind these reporting obligations is to ensure, through regulation, that social media companies operating in India are transparent and accountable to their users in their content moderation practices. By requiring that SSMIs publish compliance reports on a monthly basis, both governments and users can develop insights into the effectiveness of grievance redressal mechanisms offered by social media companies, their compliance with Indian laws in their content removal practices, and their capacity to proactively remove unlawful content. Several SSMIs have been publishing monthly compliance reports as required under the Intermediary Rules. This includes Facebook, Instagram, WhatsApp, YouTube, X, Snap, ShareChat, Koo and LinkedIn ("**Reporting SSMIs**").

Approach

This report seeks to examine how SSMLs have chosen to offer transparency to their Indian users, based on the monthly compliance report released by the Reporting SSMLs between May 2021-December 2023. The reporting metrics provided in the Intermediary Rules are considered, within this report, to be the baseline transparency commitment to be met by SSMLs (with SSMLs having the discretion to make additional disclosures voluntarily). This report first looks at the extent to which Reporting SSMLs have complied with the compliance obligations set out under the Intermediary Rules. Second, this report examines the variation in disclosures by Reporting SSMLs, given the flexibility provided by the Intermediary Rules in meeting compliance obligations. Third, the report reviews transparency metrics from SSMLs in other jurisdictions to identify potential additional disclosures that could be valuable in the Indian context. Based on this analysis, the report offers recommendations for improving reporting practices by SSMLs in India.

Part A: SSML Compliance with the Intermediary Rules

On the basis of the expectations laid down by the Intermediary Rules, this section of the report examines SSML performance across the following metrics:

1. Disclosure of details of complaints received (types of complaints);
2. Disclosure of numbers of complaints received;
3. Disclosure of action taken on complaints;
4. Details of links disabled due to proactive monitoring using automated tools;
5. Disclosure of grievance redressal officer details;
6. Presence of separate India grievance mechanism, or in-feed grievance mechanism, and whether appropriate disclosures have been made for both such mechanisms;
7. Whether there has been an increase in grievance reporting over time; and
8. Whether reporting format has been consistent.

At the outset, it is observed that some Reporting SSMLs have made their compliance

reports more accessible than others. The analysis reveals that the SSMLs have interpreted their reporting obligations in various ways. While SSMLs have generally been consistent in publishing monthly reports, there are notable differences in the content of these disclosures. Most platforms, with exceptions like Koo and LinkedIn, offer categorized details of user complaints, such as copyright infringement or harassment. Koo shares only general numbers of 'content' and 'spam' reports, while LinkedIn provides total complaint numbers without specifying categories. Notably, WhatsApp, Instagram, and Facebook also include non-content-related grievances like hacked accounts, which other platforms omit.

Regarding proactive monitoring using automated tools, most platforms disclose the number of links or content removed, but the granularity varies. Twitter/X, Facebook, Instagram share specific categories like terrorism or child sexual exploitation, while Koo and WhatsApp report on content flagged through automated tools and detection systems.

However, platforms like ShareChat and Snap lack clear data on content removed solely due to automated monitoring, making it challenging to evaluate the effectiveness of these tools.

All platforms disclose details of their India-resident grievance redressal officer and offer mechanisms for contacting them. Some, like Facebook, Instagram, and Twitter/X, have separate India grievance mechanisms in addition to global in-app systems. This creates complexity, as the dual mechanisms can make user reporting less intuitive, and the data from the separate grievance channels is often reported differently.

Consistency in monthly compliance reports is generally high, with platforms like YouTube and LinkedIn showing the most uniformity. However, platforms like Twitter/X, Snap, ShareChat, and Koo have introduced new complaint categories and modified their reporting formats over time. Updating these categories is essential to address emerging issues like misinformation, deepfakes, and synthetic media, which have become more prominent in recent years. Platforms must ensure that such concerns are adequately reflected in their complaint systems and reporting mechanisms to remain relevant and effective in moderating content.

Part B: Comparison of disclosures made by SSMLs

This section of the report examines the differences in monthly compliance disclosures among Reporting SSMLs, going beyond the requirements of the Intermediary Rules. It highlights variations in how platforms categorize and address content-related grievances.

For instance, while WhatsApp offers only one content-related grievance category, Twitter/X provides 13, showing the flexibility allowed under the Intermediary Rules to tailor grievance mechanisms based on platform-specific needs. Some platforms disclose non-content grievances, such as account access issues. For example, Facebook offers 5 non-content categories, while others, like WhatsApp, disclose 4.

While not required, platforms like ShareChat voluntarily disclose law-enforcement requests, whereas global platforms such as Facebook and Google include this data in semi-annual reports with country-specific details. Koo's unique disclosure of content removal by language, offers insight into how platforms manage content in Indian languages—a significant factor in ensuring equitable moderation for non-English users. Additionally, while most platforms provide limited information on actions taken against flagged content, some, like Facebook and Instagram, share general consequences; ShareChat stands out by being the only platform that discloses actions against accounts with repeated violations, which highlights gaps in reporting by other platforms.

Furthermore, platforms like WhatsApp and Twitter/X disclose appeals made against content moderation decisions, a valuable metric for assessing the fairness and accuracy of these decisions. The section also addresses proactive content monitoring using automated tools, showing varying degrees of transparency across platforms regarding the categories of content they monitor. Clearer disclosures on these categories would allow regulators and users to better evaluate platform performance in content moderation.

Telegram is not included, as its reports have not been published or made publicly available. Koo has recently ceased operations and no longer qualifies as a Significant Social Media Intermediary (SSMI). However, since Koo had published its transparency reports and represents an Indian company, the analysis includes its reports for the relevant period.

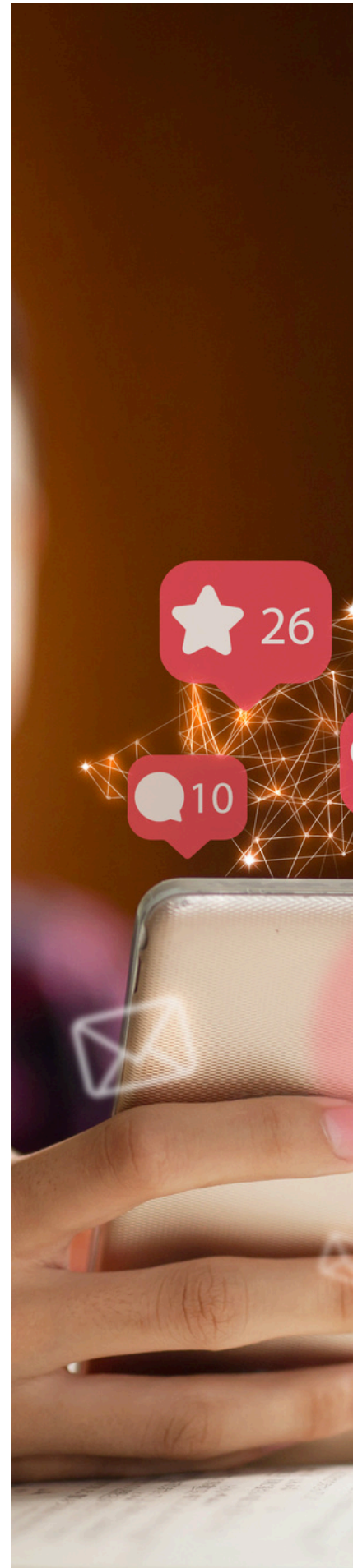
Part C : Comparison of disclosures made by SSIMs in other jurisdictions

This section of the report highlights how various jurisdictions outside India have imposed transparency and disclosure requirements on social media platforms. While India is unique in mandating monthly compliance reports under the Intermediary Rules, several other countries, such as those in the European Union, Austria, Germany, and Turkey, also require platforms to make similar disclosures. Although they require less frequent reporting, they often mandate more detailed and granular data, offering regulators deeper insights into the platforms' content moderation practices. Indian regulators could benefit from adopting similar granular reporting requirements. Key metrics that could be included are details on average monthly active users, the number of human moderators by language, and the speed and outcomes of content moderation processes. Expanding the scope of disclosures in India to include these metrics would be feasible, as global platforms already provide such detailed data in other regions.

Conclusion and recommendations

The report evaluates how SSIMs are addressing their transparency obligations under the Intermediary Rules. While most platforms generally comply, the inherent flexibility and ambiguity of the rules leave certain aspects lacking. Even though some platforms voluntarily exceed the mandated disclosures, transparency remains inconsistent overall. To address this, modifications to the Intermediary Rules could help strengthen the compliance process, ensuring that the original intent of the rules is met more comprehensively. Alternatively, the government can invoke its powers under Rule 4(9) to request additional information, ensuring greater consistency and uniform disclosures across platforms. The report also notes some of the additional disclosures (not mandated by the Intermediary Rules) that SSIMs have been making within India, and the disclosures that some of the reporting SSIMs are obligated to make in other jurisdictions. Based on a cumulative assessment of these disclosures, the set of recommendations are offered as follows:

- ▶ Mandatory disclosure requirements for all social media intermediaries, requiring them to report the number of registered and active users. Based on these disclosures, the government can publicly identify platforms classified as SSIMs, ensuring clarity on which platforms are obligated to provide monthly compliance reports and fulfill their regulatory responsibilities
- ▶ Periodic review of SSIM disclosures by regulators and publication of reports, with directions to SSIMs requesting additional information when required
- ▶ Requirement for more granularity in disclosures made by SSIMs under the Intermediary Rules;
- ▶ Need for disclosures in relation to how content moderation activities by SSIMs accounts for the diversity of languages within India;
- ▶ Need for disclosures on SSIM's efforts towards user protection in the content moderation activities being undertaken.



Introduction

Background

Since June 2021, the social media intermediaries that operate within India, and are categorized as SSMLs have released monthly compliance reports for India. These reports are published in compliance with SSML obligations under Rule 4(1)(d) of the Intermediary Rules. While several jurisdictions had begun to mandate comprehensive transparency reporting from social media companies, India is the first to require monthly reports of this nature. In comparison with transparency mandates in other jurisdictions, India requires relatively fewer disclosures, and are largely concerned with SSML grievance redressal practices, and their efforts towards proactive content removal. Moreover, the Intermediary Rules offer considerable flexibility to SSMLs in terms of how they wish to fulfil this reporting obligation, while preserving the government's discretion to specify additional disclosures. This compliance obligation intends to ensure the accountability and transparency in the content moderation practices engaged by SSMLs, in accordance with the government's expressed interest to ensure an open, safe, trusted and accountable internet for India.

This report examines the outcome of two and a half years (June 2021- December 2023) of compliance reporting by nine significant social media intermediaries based on the requirements laid out under the Intermediary Rules. The report may be examined in three parts:

- a) effectiveness of reporting mechanism established under the intermediary Rules;
- b) comparison of the various categories of information disclosed by the different SSMLs; and
- c) compliance obligations being fulfilled by SSMLs within other jurisdictions.

Based on a cumulative assessment of efforts undertaken by SSMLs thus far, as well as an analysis of reporting obligations, a few recommendations are laid down for improving the implementation of the reporting framework established under the Intermediary Rules for SSMLs.



Regulatory provisions: Reporting obligation under Rule 4(1)(d) of the Intermediary Rules

Under Rule 4(1)(d), of the Intermediary Rules, SSIMs are required to publish monthly compliance reports with the following information:

1. Details of complaints received;
2. Action taken thereon;
3. The number of specific communication links or parts of information that the intermediary has removed or disabled access to, based on proactive monitoring conducted by using automated tools;
4. Any other information as may be specified.

In November 2021, the FAQs¹ released by Ministry of Electronics, Information and Technology (MeitY) further clarified the compliance obligation under Rule 4(1)(d) for SSIMs by requiring the following:

1. Numbers of communication links removed by the SSIM to fulfil the requirement to report voluntary actions taken by an SSIM; and
2. Summary details of complaints received (for example, the subject under which the complaint was received) and action taken under each of the different heads. This information could be disclosed in the aggregated form, without disclosing granular details of all cases.

1. https://www.meity.gov.in/writereaddata/files/FAQ_Intermediary_Rules_2021.pdf

In addition to this, intermediaries are also obligated to upload reports detailing compliance with orders passed by the Grievance Appellate Committee instituted under the Intermediary Rules, in accordance with Rule 3A(7) of the Intermediary Rules. The compliance reporting obligation does not exist within a vacuum, and can be read in conjunction with several related obligations set out under the Intermediary Rules. For example, Rule 3(1)(b) provides several categories of non-permissible content that social media platforms must make efforts towards preventing on the platform. These specified categories could in turn inform the categories of complaints that users could make within the grievance redressal mechanism of an intermediary (as specified in Rule 3(2)(a)), and is disclosed in the SSML compliance reports).

Similarly, Rule 4(4) requires SSMLs to work towards deploying automated tools or other mechanisms to proactively identify content that depicts rape, child sexual abuse or conduct, as well as content that is identical to content that was previously removed. The same provision could be used by SSMLs to understand where there is an expectation for proactive monitoring, and its obligation to report content removed on the basis of such proactive monitoring. In essence, the reporting framework created by the Intermediary Rules can work to provide insight into the content moderation activities undertaken by platforms, the efficacy of content moderation measures through automated means, the number of user complaints received against problematic content, and how SSMLs choose to take action against such content on their platforms.

The flexibility offered by the Intermediary Rules in how SSMLs makes its disclosures, and the absence of any prescribed format for disclosures also means that there is some room for interpretation of the reporting requirement. For instance, it is unclear whether the compliance report needs to consider complaints received solely from Indian complainants, though some of the SSMLs with a global presence have chosen this approach. Similarly, it is unclear whether the monthly compliance report requires disclosure about the kinds of action taken against different categories of complaints, or whether it simply requires a disclosure on whether a complaint was actioned or not. In the absence of clarity, many SSMLs have chosen the latter route, or have simply accounted for content pieces that have been taken down. This flexibility also means that there is no strict disclosure template for SSMLs for monthly reporting. SSMLs have used this flexibility to not only disclose what has been understood by them to be the extent of their reporting obligations, but also disclose information beyond what is explicitly required under the Intermediary Rules.

Notwithstanding the flexibility offered by the Intermediary Rules, it is understood that the intention of the government in instituting the monthly reporting obligation was to ensure that grievance redressal and proactive content moderation is effectively undertaken.

By mandating that SSMLs ‘publish’ these compliance reports on their platforms, both the government and the general public can assess the SSMLs performance in providing a safe and accountable environment for users.

In the mid-to-long run, the regular monitoring of aggregate information obtained from monthly compliance reporting can also work to shed light on how effectively SSMLs perform with respect to each other when it comes to effective content moderation, and give insight into the areas where platforms have room to improve their own practices.



Part A - Effectiveness of reporting mechanism established under the Intermediary Rules

Since July 2021, social media platforms have been publishing monthly compliance reports as required under Indian law. This includes Facebook, Instagram, WhatsApp, YouTube, X, Snap, ShareChat, Koo and LinkedIn. These platforms have not formally disclosed the number of registered Indian users, as currently, there is no reporting requirement to this effect under the Intermediary Rules. Therefore, it is understood that these platforms meet the notified threshold of having five million registered users in India.²

There are more social media intermediaries that meet this threshold but whose compliance with their reporting obligations may not be as easy to determine.³ Telegram for example, has over five million users in India⁴, but does not appear to be fulfilling the ‘publishing’ obligation for their compliance reports, in accordance with Rule 4(1)(d) of the Intermediary Rules. The short video app Moj (owned by ShareChat) also does not seem to have shared their compliance reports, though it is said to have over 12 million monthly active content creators.⁵

Of the platforms that have consistently released their monthly compliance reports since June 2021, the below table analyzes their performance in meeting the requirements envisaged by the Intermediary Rules.

2. This is the threshold for a social media intermediary to be considered a significant social media intermediary, as specified in the gazetted notification dated 25th February, 2021.

3. It may be noted that even among the Reporting Intermediaries, LinkedIn has not made their monthly compliance reports easily accessible. While LinkedIn’s Transparency Section provides links to access Transparency Reports released under the EU’s Digital Services Act, and also provides links to access information on how LinkedIn responds to government and law enforcement requests for information, there is no mention of compliance reports issued under India’s Intermediary Rules. Instead, it is only upon accessing LinkedIn’s Help Centre, and using the search function that it is possible to find their monthly compliance reports in tabular form.

4. <https://timesofindia.indiatimes.com/education/news/will-telegram-be-banned-in-india-five-exam-controversies-linked-with-this-app/articleshow/112854952.cms>










5. <https://inc42.com/features/how-sharechats-short-video-app-moj-lost-its-moj/>










This table examines the outcome of two and a half years (June 2021-December 2023) of compliance reports by nine Reporting SSIMs based on the requirements laid out under the Intermediary Rules. On the basis of the expectations laid down by the Intermediary Rules, the following table examines SSIM performance across the following metrics:










- a) disclosure of details of complaints received (types of complaints);
- b) disclosure of numbers of complaints received;
- c) disclosure of action taken on complaints;
- d) details of links disabled due to proactive monitoring using automated tools;
- e) disclosure of grievance redressal officer details;
- f) presence of separate India grievance mechanism, or in-feed grievance mechanism, and whether appropriate disclosures have been made for both such mechanisms;
- g) whether there has been an increase in grievance reporting over time; and
- h) whether reporting format has been consistent.




Finally, the table also provides some insight into the two most common grievances each platform has seen in the year 2023. A more detailed description of platform performance on the various metrics is provided below the table.



Platform									
Details of complaints received (types of complaints)	✓	✓	✓	✓	✓	✓	✗	✓	✗
Numbers of complaints received	⚡	⚡	✓	✓	✓	✓	✓	⚡	✓
Actions taken on complaints	⚡	⚡	✓	⚡	⚡	⚡	✓	✓	⚡
Links disabled due to proactive monitoring (in numbers)	⚡	⚡	✓	✓	✓	✗	✓	✗	✓
Disclosure of grievance redressal officer	✓	✓	✓	✓	✓	✓	✓	✓	✓
Presence of separate grievance mechanism for Indian grievances (where platform has a presence outside India)	✓	✓	✓	✓	✓	✗	✗	✗	✗

Platform									
Presence of grievance redressal mechanism (in-app/in-feed)	✓	✓	✓	✓	✓	✓	✓	✓	✓
If presence of more than one grievance mechanism, then whether report discloses grievances/reports made within all such mechanisms	⌘	⌘	⌘	✗	⌘	N/A	N/A	N/A	N/A
Whether Indian grievance reporting (where available) or other grievance mechanism has seen an increase in usage? (Jun'21- Dec'23)	✓	✓	✓	✓	✗	⌘	⌘	✗	✗
Consistency of reporting (i.e., whether there are changes in reporting format over time)	✓	✓	⌘	⌘	✓	⌘	⌘	⌘	✓
Consistency and ease of public access to compliance reports	✓	✓	✓	✓	✓	✗	✗	✗	✗
Most common content grievance (where such data is available) (average-2023)	Bullying or harassment	Inappropriate or abusive content	N/A	Abuse/ Harassment	Copyright violation	Bullying or harassment	Graphic/ Obscene/ Sexual content	Abusive	N/A

Platform									
Second most common content grievance (average-2023)	Inappropriate or abusive content	Bullying or harassment	N/A	Hate Speech	Trademark infringement	Sexual Content	Abusive	Sexually explicit content	N/A

 Yes	 No	 Somewhat/Inconsistent/Room for improvement
---	--	--

Analysis of data provided by platforms in their compliance reports in meeting requirements under the Intermediary Rules

1. Details on complaints received

Almost all platforms provide details of user complaints that have been received, categorised by types of complaints (categories include Copyright, Bullying, Abuse, Harassment, etc). An exception to this is Koo and LinkedIn. Koo fails to provide clear categories of complaints, only sharing numbers of 'content' reported, and the numbers of 'spam' reported.⁶ LinkedIn only shares the total number of complaints received without sharing any additional information on the kinds of complaints received.⁷ While other platforms share information only in relation to content related complaints, WhatsApp, Instagram and Facebook also disclose details of non-content related grievances they receive⁸, such as reports on hacked accounts, requesting access to personal data stored on the platform etc.

2. Providing numbers of complaints received

Six of the nine platforms provide the monthly numbers of complaints that they have received. Until May 2022, ShareChat disclosed percentages of total complaints received within each category.⁹ After May 2022, ShareChat also began to share the numbers of complaints received within each category.¹⁰ Facebook and Instagram disclose the numbers of complaints within various categories received through the Indian grievance mechanism alone.¹¹ Data is cumulatively provided for the number of content proactively removed and reported by the community through any alternate mechanisms. While data is also provided on the percentage of content that was proactively removed, the individual numbers for Facebook and Instagram cannot be ascertained.

3. Action taken on complaints received under different heads

This disclosure requirement is one that most platforms seem to be inconsistent in providing. ShareChat and Koo are notable exceptions. ShareChat provides granular information on the various kinds of actions that have been taken, including numbers in relation to content takedowns, the various kinds of temporary bans, and the number of permanent bans for users on their platform.¹²

6. Koo's reports are no longer accessible as the platform announced it was shutting down in July 2024

7. <https://www.linkedin.com/help/linkedin/answer/a1335719?hccppcid=search>

8. Instagram and Facebook: <https://transparency.meta.com/sr/india-monthly-report-May31-2022/>; Whatsapp: https://scontent-atl3-2.xx.fbcdn.net/v/t39.8562-6/317076809_811080299954149_2261501632158642438_n.pdf?nc_cat=102&ccb=1-7&nc_sid=b8d81d&nc_ohc=Dt7PN-xmxuwQ7kNvgHiFKCF&nc_ht=scontent-atl3-2.xx&nc_gid=AVQoAt1lc8sw2CgsvB_74Wx&oh=00_AYAoVoBg-F00ONp5mlUXVzYWSvzzD9JM46pjp9ixKCILQ&oe=66EF2645

9. <https://help.sharechat.com/transparency-report/february-2022>

10. <https://help.sharechat.com/transparency-report/august-2022>

11. Supra Note 8

12. <https://help.sharechat.com/transparency-report/march-2022/>

Koo also shares details of numbers of content pieces removed, as well as the numbers for which a different action was taken (blur, warn, etc).¹³ WhatsApp also shares information on action taken, which, in their case, is limited to banning, or overturning bans on accounts.¹⁴ Disclosures on actions taken are less transparent in the case of Facebook, Instagram, Twitter/X, Google, Snap and LinkedIn.¹⁵

In the case of Facebook and Instagram, numbers are offered on content that has been actioned, and some information is provided on what the consequences of actioning content can be. In the context of reports through the India grievance reporting mechanism, this can mean removing content, covering photos or videos with a warning, and restricting content availability in the country. While Facebook and Instagram provide details in numbers of instances where tools were provided for complainants (i.e., mechanisms by way of which complainants themselves can remedy their grievance or concern) through the India grievance mechanism, they do not share the various kinds of actions that the platforms have taken (segregated by numbers, category of actions and complaints). In the context of actioning content reported through any alternate grievance reporting mechanism, the report only shares that this can include removing content, or covering photos or videos with a warning. Meta's Transparency Centre discloses that they take additional

measures including measures to curb the reduction of spread of borderline content.¹⁶

These kinds of 'actions' are not discussed within the monthly India compliance reports. A more robust and transparent way of reporting in the case of would be to also provide details (in numbers) for different 'actions' undertaken for the various categories of problematic content.

In the context of Twitter/X, the reports share the numbers of URLs that were 'actioned' for each category of complaint received. However, they do not share details of what actioning content can mean.¹⁷ X's Help Centre discloses a range of measures they take when actioning violative content.¹⁸ However, these are not disclosures that are made within the monthly India compliance reports. They also share the number of accounts proactively monitored and suspended for the promotion of terrorism and child sexual exploitation using a combination of technology and other purpose-built internal proprietary tools. YouTube and LinkedIn share details on instances where content was removed. However, there may be alternate actions taken for flagged content, including age-based content restrictions, or measures to curb the reduction of spread of borderline content in the case of YouTube,¹⁹ and limiting the visibility of content or labelling content, in the case of LinkedIn.²⁰ These kinds of 'actions' are not discussed within the monthly India compliance reports.

13. Supra Note 6

14. Whatsapp- India Monthly Report, July 2023, https://scontent-atl3-1.xx.fbcdn.net/v/t39.8562-6/356864340_258864053422020_7727900727558791209_n.pdf?_nc_cat=108&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=FiEJMTof8WcQ7kNvgHxhWcS&_nc_ht=scontent-atl3-1.xx&_nc_gid=AVQoAt1lc8sw2CqsvB_74Wx&oh=00_AYBcb8RniQP9UAq8_QniFmB_2Txh2RfZ1BlizsSQisGSJQ&oe=66EF0F7B

15. <https://transparency.meta.com/sr/india-monthly-report-May31-2022/>

16. <https://transparency.fb.com/hi-in/enforcement/taking-action/>

17. <https://transparency.twitter.com/content/dam/transparency-twitter/country-reports/india/India-ITR-Mar-2023.pdf>

18. <https://help.twitter.com/en/rules-and-policies/enforcement-options>

19. <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/>

20. <https://www.linkedin.com/legal/professional-community-policies>

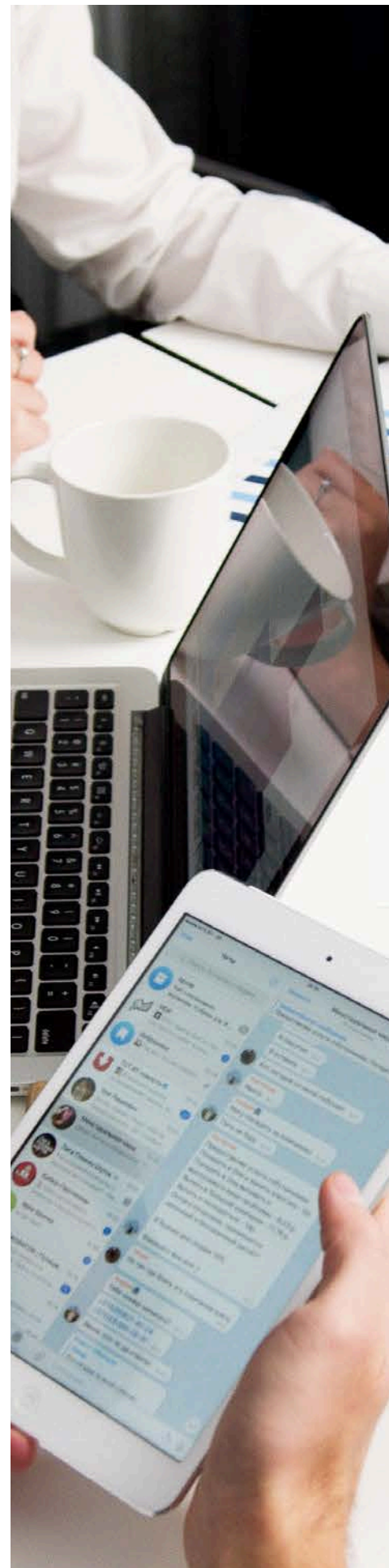
Snap's monthly compliance reports only provide information on content enforced and unique accounts enforced, as well as account deletions for CSEAI (Child Sexual Exploitation and Abuse Imagery) and terrorism content.²¹ However, Snap's Transparency section on their Community Guidelines also discloses a graded and risk-based approach to content moderation which includes a 'multiple-strikes' rule an account is disabled. For better insight into actions taken by Snap, the reports would also need to disclose the number of accounts disabled for multiple violations of community guidelines.

4. Links disabled due to proactive monitoring using automated tools

Seven²² of the nine platforms provide disclosures on the numbers of links or content pieces removed as a result of proactive monitoring using automated tools. YouTube and LinkedIn provide a monthly number for the total pieces of content detected using automated tools. Twitter/X discloses separate numbers of content pieces detected through automated means for terrorism promotion content, and content relating to child sexual exploitation, non-consensual nudity and similar content. Koo discloses separate numbers of content pieces detected through automated means for spam, as well as for content violating their community standards. WhatsApp also discloses the number of accounts removed as a result of their automated detection measures.

ShareChat does not explicitly disclose the number of content pieces removed via automated detection.²³ While it mentions proactive removals based on violations of Community Guidelines, Terms of Use and other policy standards, it is unclear if automated tools are involved, as the data combines content removed both proactively and through user complaints, preventing a clear estimate of removals solely due to proactive monitoring.

Snap provides information on account deletions for CSEAI (child sexual exploitation and abuse imagery), and terrorist content.



21. <https://snap.com/en-US/privacy/transparency/india-11-2021>

22. Meta provides a cumulative percentage of content proactively actioned on Facebook and Instagram.

23. <https://help.sharechat.com/transparency-report/?q=sharechat-october-2023>

However, it is not made explicit whether this was undertaken as an outcome of proactive monitoring. In their half-yearly reports, they give a cumulative percentage of CSEAI content detected and actioned.²⁴ For Facebook and Instagram, data is provided cumulatively for content removed proactively and through community reporting via alternate mechanisms. Additionally, the percentage of content removed proactively before user reports is shared as an indicator of the effectiveness of their monitoring tools in detecting violations.

5. Disclosure of grievance redressal officer

All platforms disclose details of their India-resident grievance redressal officer, as well as details on how to contact the officer. Many platforms also provide an address for contacting them within India via post. In all cases, this is a disclosure that is offered on the websites of the platforms.

6. Presence of separate India grievance mechanism, or in-feed grievance mechanism, whether appropriate disclosures have been made for both such mechanisms, and increase in reporting

Some of the SSMLs that have an international presence have chosen to include a separate 'India grievance mechanism' on their platform for users in India to channel their grievances. This is additional to, and separate from, any in-app or in-feed mechanism that they have for consumers to report content related grievances.

Platforms that have chosen to do this include Facebook, Instagram, WhatsApp, YouTube and Twitter/X. Platforms that have a strong Indian presence, and a largely or fully Indian consumer base such as Koo and ShareChat have chosen not to have a separate India grievance mechanism. Even though Snap has a global presence, it has chosen not to have a separate India grievance mechanism.

It may be noted that the presence of a separate India specific grievance redressal mechanism may not be a very user-friendly initiative. Most users will find it to be a more onerous process to visit a new webpage, find the grievance redressal form, insert the relevant information, including the link to the violating piece of content, and submitting the application. When the in-app or in-feed complaints mechanism works more effectively, and is easier to access, users may prefer that mechanism.

²⁴ <https://values.snap.com/privacy/transparency?lang=en-US>

Incomplete information provided by certain platforms

Multiple avenues for grievance redressal also mean that some of the disclosures made by platforms may be somewhat incomplete. In the case of Facebook and Instagram, data is cumulatively provided for the numbers of content proactively removed and reported by the community through in-app mechanisms. Further, details of complaints received through the India grievance redressal mechanism have been separately disclosed. It is difficult to come to a comprehensive understanding of all complaints received through both mechanisms, since the disclosure categories for content grievances through the in-app mechanism and the India grievance mechanism are different, and therefore cannot be compiled.

In the case of WhatsApp, details of complaints received through the India grievance redressal mechanism have been separately disclosed. With respect to in-app complaints, data can be estimated on the number of WhatsApp accounts that have been banned because of user reports. However, the total number of user reports shared with WhatsApp through the in-app mechanism has not been disclosed.

With Twitter/X, which also has two ways to report grievances, only grievances reported through their India mechanism has been disclosed.²⁵ As the reports indicate, Twitter/X has understood that Rule 4(1)(d) of the Intermediary Rules requires Twitter/X to only publish compliance reports detailing complaints received through its India grievance mechanism and actions taken thereon. This might also provide some context for why they have reported an average of 1800 complaints each month in 2023. It is more likely that there are more complaints being made by Indian users through the in-app or in-feed mechanism. However, information regarding such complaints is absent in the monthly reports. In the case of YouTube, the platform's monthly compliance reports provides details on complaints received through "designated complaint channels". It is unclear if this refers to both the separate India grievance redressal mechanism that YouTube has provided, as well as the in-app/in-feed reporting mechanism offered by the platform. It can be assumed that the monthly compliance reports contain data from all their grievance reporting options. However, the reports themselves do not clarify this.

25. <https://transparency.twitter.com/content/dam/transparency-twitter/country-reports/india/India-ITR-December-2023.pdf>

Obligations for grievance redressal under the Intermediary Rules

The Intermediary Rules do not explicitly require the carve out of a separate India grievance redressal mechanism. Under Rule 4(6), SSMLs are required to implement an appropriate mechanism for the receipt of complaints and reports of violations of Rule 4 of the Intermediary Rules. It is further stated that such a grievance mechanism must enable complainants to track the status of complaints by providing a unique ticket number for every complaint or grievance received.










It is likely that the carve-out of a separate grievance redressal mechanism by SSMLs that have a global presence was undertaken in order to comply with the specific requirements that have been laid down by under Rule 4(6). Rather than modify existing grievance redressal mechanism, to fit the requirements, Facebook, Instagram, Twitter/X and WhatsApp chose to introduce a new and separate mechanism for Indian grievance redressal, while also retaining their existing in-app grievance redressal mechanism. In line with the intent of the Intermediary Rules and the broader objective of transparency, it can be reasonably interpreted that SSMLs are expected to disclose data related to all complaints received from users from India , rather than limiting it to those filed through the India-specific grievance mechanism.

7. Improvement in use of grievance redressal mechanism over time

When there are multiple grievance redressal mechanisms operating in tandem with each other, it is difficult to gauge the efficacy of one mechanism vis-à-vis the other. The lack of clarity on whether platforms are disclosing data in relation to all the user complaints that they are receiving, and the absence of disaggregated data on the India grievance redressal mechanism and the general in-app grievance redressal mechanism, also makes it difficult to discern how effective the introduction of specific grievance redressal requirements under the Intermediary Rules has been. This therefore means that there is no conclusive data available through the monthly compliance reports to indicate whether changes in grievance-redressal obligations introduced under the Intermediary Rules has resulted in an improvement in user reporting of grievances.

It may also be noted that understanding the improvement of usage of the grievance redressal mechanisms over time is also nearly impossible without an accompanying disclosure on the number of active users within a platform. For instance, an increase in user complaints could be explained by an increase in user awareness of their grievance redressal options. It could also be explained by the fact that a platform has seen an increase in active users on the platform. Based on publicly available data (which may be inaccurate), the approximate number of Indian users of various SSMLs are detailed below.²⁶

26. This data has been aggregated from various databases and news reports including <https://napoleoncat.com/>, <https://www.statista.com/>, and [-]. This data could well prove to be very inaccurate. However, in the absence of clear disclosures by SSMLs themselves on their monthly active users within India, it will be difficult to conclude on any accurate statistic for the same.

Platform									
Approximate number of users in India (2021) ²⁷	410 million	210 million	530 million	15 million	448 million	100 million	15 million	160 million	76 million
Approximate number of users in India (2024 or most recently available data) ²⁸	566.7 million	375.6 million	535.8 million	27.3 million	462 million	200 million	4.1 million	180 million	120 million

It would appear that all platforms except Koo have seen an increase in their user base over time, which could possibly be an explanation for the increase in user complaints in most platforms. To obtain a clearer picture on the efficacy of grievance redressal mechanisms, there needs to be more granularity in disclosures both in terms of the number of active users on a platform, as well as disaggregated data in relation to user grievances recorded through all avenues available on a platform for receiving complaints.

To the limited extent of the disclosures made by various platforms in relation to user grievance reporting, Facebook, Instagram, WhatsApp and Twitter/X have seen an increase in reporting over time (July 2021-December 2023). YouTube and ShareChat on the other hand has seen a reduction in user reports. In the case of Snap and Koo, there does not seem to be any correlated increase or decrease in the use of their respective grievance redressal mechanisms.



27. Note: These statistics have been compiled from the following sources:
For Facebook, Instagram, WhatsApp, Twitter/X and YouTube: <https://www.indiatoday.in/technology/news/story/government-reveals-stats-on-social-media-users-whatsapp-leads-while-youtube-beats-facebook-instagram-1773021-2021-02-25>
For Snap: <https://www.gadgets360.com/apps/news/snapchat-india-users-reach-100-million-2021-jiophone-next-flipkart-zomato-snap-partnership-2589429>
For Koo: <https://www.moneycontrol.com/news/business/koo-user-base-at-about-15-million-eyes-expansion-to-new-market-in-southeast-asia-in-h2-2022-7619231.html>
For ShareChat: <https://asiatechdaily.com/the-indian-unicorn-club-2021-entrants-sharechat/>
For LinkedIn: <https://napoleoncat.com/stats/linkedin-users-in-india/2021/05/>

28. Note: These statistics have been compiled from the following sources:
For Facebook: [https://napoleoncat.com/stats/facebook-users-in-india/2024/02/#:~:text=There%20were%20566%20751%20300,group%20\(207%20300%20000\).](https://napoleoncat.com/stats/facebook-users-in-india/2024/02/#:~:text=There%20were%20566%20751%20300,group%20(207%20300%20000).)
For Instagram: <https://napoleoncat.com/stats/instagram-users-in-india/2024/02/>
For WhatsApp: <https://www.demandsage.com/whatsapp-statistics/>
For Twitter/X: <https://worldpopulationreview.com/country-rankings/twitter-users-by-country>
For YouTube: <https://www.statista.com/statistics/280685/number-of-monthly-unique-youtube-users/#:~:text=As%20of%20January%202024%2C%20India,users%20watching%20content%20on%20YouTube.>
For Snap: <https://www.businesstoday.in/bt-tv/video/even-with-200-million-snapchat-users-in-india-no-still-looks-small-snap-inc-ceo-evan-spiegel-416362-2024-02-06>
For Koo: <https://www.financialexpress.com/business/industry-koo-loses-traction-on-active-user-count-3023954/>
For ShareChat: <https://sharechat.com/about>
For LinkedIn: <https://www.statista.com/statistics/272783/linkedin-membership-worldwide-by-country/>

8. Consistency of monthly compliance reports

All the platforms have been consistently sharing their monthly compliance reports. While the content of the reports has been largely consistent, there are some variations in formats and complaint categories that have been made by platforms. YouTube and LinkedIn have been most consistent in their reporting, while Facebook and Instagram have adopted a consistent format for reporting since October 2021. WhatsApp has also not introduced any changes in their reporting since July 2022. With Twitter/X, Snap, ShareChat and Koo, there have been introduction of new categories of complaints, as well as modifications and recategorizations, made over time. Twitter/X has even made changes to their complaint categories as recently as December 2023.²⁹ Updating complaint categories to reflect newer and more diverse types of complaints is essential. Platforms should also regularly update their disclosure reports to highlight initiatives that address emerging concerns related to the content they host.

For instance, misinformation may not have been a major issue in the early 2000s or 2010s but has become more prominent in the last 6-8 years. Similarly, deepfakes and synthetic media have recently become significant concerns, making it crucial for platforms to ensure these categories are reflected in their complaint systems. Compliance reports should, therefore, detail efforts not only to respond to user complaints about deepfakes but also to proactively address such content.

Regulatory oversight of compliance reports is equally important. Some platforms, despite meeting the criteria for SSMLs under Indian law, have not published monthly disclosures. Others have made minimal changes to their complaint categories over the past 2.5 years. Regular reviews of compliance reports by regulators can help create an iterative process for platforms to update their complaint categories as needed.

9. Consistency in access and ease of access to monthly compliance reports

Facebook, Instagram, WhatsApp, X and YouTube have offered their India compliance reports consistently and in an easily accessible manner. In most of these cases, the monthly compliance reports are available in the 'Transparency' sections of their platforms, alongside the disclosures they may have to make in other jurisdictions.

While SSMLs have been sharing compliance reports, and it was possible to collect the data required for the period of this report, in the case of some SSMLs it cannot be said that the monthly compliance reports are easily accessible, or consistently available. The data for the 2.5-year period covered in this research was initially collected between February and March 2024. However, by June 2024, the same data became harder to retrieve or was inaccessible. Koo's reports for the period between October 2022 to March 2023 are not available on their website.³⁰ Similarly, ShareChat's reports for the year 2023 were intermittently unavailable on their website.

29. A new category added, 'Others', which includes other reports with a majority related to ban evasion. <https://transparency.twitter.com/content/dam/transparency-twitter/country-reports/india/India-ITR-January-2024.pdf>

30. Koo's reports are no longer accessible as the platform announced it was shutting down in July 2024.

Attempting to access Snap’s compliance reports from November 2022 onwards resulted in the user being redirected to their aggregated 6-month reports for Jan-Jun 2022 rather than the required monthly compliance reports, though the monthly reports for the same period were previously available on their website at the time of data collection. While disclosures for other jurisdictions are available within their ‘Transparency’ section, in the case of LinkedIn, accessing the India monthly compliance reports requires an individual to use the ‘search’ function on their ‘Help’ section in order to access the same.

In the case of Snap and ShareChat, accessing monthly compliance reports is easy, but inconsistent. In July 2024, ShareChat’s monthly compliance reports are only available for the period between June 2021 and October 2022. Monthly reports until December 2023 (the timeline that this report considers) had been available earlier, but this no longer seems to be the case. Similarly, in the case of Snap, reports from November 2022 onwards are no longer available to access in July 2024, though monthly reports until December 2023 were available just a few months earlier.

10. Common content related grievances





As may be seen in the table above, the most common grievances among platforms in 2023 appear to be in the context of bullying and harassment. A second common complaint appears to be in relation to sexually explicit or graphic content. YouTube is an outlier in this context, with the most common grievance recorded on their platform being in relation to copyright and trademark infringements. Since LinkedIn does not make any disclosures of the kinds of complaints made in relation to content, the most common content related grievance cannot be determined in LinkedIn’s case.

11. Publishing of Reports detailing compliance with GAC orders

The Central Government established three Grievance Appellate Committees (“**GAC**”) on 27th January 2023 in accordance with the Intermediary Rules. Persons aggrieved by the decision of an intermediary’s Grievance Officer or whose grievance is not resolved within the period specified for resolution under the Intermediary Rules. Orders passed by the GAC are to be complied with by the concerned intermediary, and a report to that effect is required to be uploaded on the intermediary’s website.

The obligation to comply with GAC orders and upload reports detailing compliance is one that applies across intermediaries, and not just for SSMLs. The table below however, analyses how the SSMLs discussed in this report have chosen to comply with the reporting obligation for their compliance with GAC Orders.










Platform	Reports on GAC Compliance
	<p>Google uploads reports on a 6-monthly basis detailing their compliance with GAC Orders. These reports provide information on the number of appeals that were received in relation to Google’s various intermediaries (YouTube, Gmail, Google Play, Google Maps, and Google Search), the appeals where the GAC upheld Google’s original decision, and where the appeals were not admitted by the GAC.</p> <p>Thus far, it would appear that there are no instances of user appeals against Google’s decisions that were allowed by the GAC (thereby requiring Google to reverse their decision in relation to a user grievance).</p> <p>In the period between March 1 2023 to September 30, 2023, the GAC received 12 appeals in relation to Google’s services. For the period between October 1 2023 to March 31 2024, the GAC received 11 appeals in relation to Google’s services.</p>
 & 	<p>Facebook and Instagram upload reports on GAC Compliance within their monthly compliance reports.</p> <p>Each monthly compliance report from March 2023 onwards, provides information on the number of orders that have been received from the GAC for both Facebook and Instagram, and the number of orders that have been complied with. Facebook and Instagram comply with all orders received from the GAC.</p> <p>There has been an increase in the number of GAC orders that Facebook and Instagram receive over time.</p>
	<p>WhatsApp uploads reports on GAC Compliance within its monthly compliance reports.</p> <p>Each monthly compliance report from March 2023 onwards, provides information on the number of orders that have been received from the GAC for WhatsApp, and the number of orders that have been complied with. WhatsApp complies with all orders received from the GAC.</p>
	<p>No information available</p>

Platform	Reports on GAC Compliance
	No information available
	No information available
	No information available
	No information available

Part B- Comparison of disclosures made by various SSMTs in their monthly compliance reports.

This section compares the data and metrics shared by nine SSMTs in their monthly compliance reports, and provides a comparison of what kinds of data are being shared by each SSMT. As can be seen below, disclosures made by platforms can often go beyond the scope of the required disclosures under the Intermediary Rules. There are variations in disclosures made across platforms, some of this may be due to the inherently different nature and structure of the platform itself (for eg. WhatsApp is inherently different in terms of how information is shared and disseminated). In the remaining contexts, the variations appear to be based on how platforms may have chosen to interpret their obligations under the Intermediary Rules, as well as based on how much information they are willing to share voluntarily.

Platform									
Number of categories for content-related complaints (Within India specific grievance mechanism)	5	5	1	13	9	NA	NA	NA	NA
Number of categories for content-related complaints (General grievance mechanism, and if disclosed separately in the compliance report)	13	12	NA	NA	NA	11	2	14	0
Number of categories for non-content related grievance	5	3	4	4	0	0	0	0	0

Platform									
Disclosure in monthly compliance report regarding data on requests by law enforcement, and action taken	No	No	No	No	No	No	No	Yes	No
Disclosure on content removed distinguished by language	No	No	No	No	No	No	Yes	No	No
Comprehensive disclosures on various actions undertaken for problematic content	No	No	No	No	No	Yes	Yes	Yes	No
Disclosure on actions taken for repeated violations	No	No	No	No	No	No	No	Yes	No
Disclosure on appeals of content moderation/account deletion/decisions made	No	No	Yes	Yes	No	No	No	No	No
Disclosure on tools provided for users to resolve their concerns	Yes	Yes	Yes	No	No	No	No	No	No

Disclosure of proactive monitoring using automated tools	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Number of content categories being proactively monitored	13	12	1	2	-	N/A	2	N/A	N/A



1. Categories of content-related grievances

Within the grievance redressal mechanism specifically carved out for India, as disclosed within monthly compliance reports, Facebook and Instagram offers 5 content related grievance categories. WhatsApp offers 1 content related grievance category, Twitter/X offers 13, and YouTube offers 9. In the general in-app or in-feed mechanism, and as disclosed within the monthly compliance reports, Facebook and Instagram offer 13 and 12 categories for complaints, Snap offers 11, Koo offers 2, and ShareChat offers 14. LinkedIn offers no disclosures on their categories of content-related grievances.

The variations across platforms in categories for content-related grievances might be a result of the flexibility offered by the Intermediary Rules in allowing platforms to determine the operation of the grievance redressal mechanism that they offer. This flexibility is also necessary, given that different platforms may operate differently, and cater to different user bases as well.

2. Categories of non-content related grievances

Though the Intermediary Rules largely intended disclosures on content related user grievances, some platforms have also chosen to disclose non-content related grievances made on the platform. This includes grievances, or requests in relation to users needing help accessing accounts, requesting access to personal data collected by the platform, or other kinds of support relating to the platform's product or safety features. Facebook offers 5 such categories of grievances, Instagram offers 3, and WhatsApp and Twitter/X offer 4 such categories.

3. Information on law-enforcement requests

Though there is no obligation under the Intermediary Rules to disclose the number of law-enforcement requests received by platforms, and the number of requests that were responded to, or actioned against, ShareChat still chooses to disclose this information. Platforms with a global presence including Facebook, Snap, Google, Twitter/X and LinkedIn have chosen to share details of global government requests for information in separately issued semi-annual Transparency reports which they have been publishing even before the Intermediary Rules came into effect. These global reports also include country specific information on law-enforcement requests received.



4. Information on content removed, distinguished by language

While not a mandated requirement, Koo also chooses to share data on the number of content pieces removed, on the basis of language. Koo separately shares the number of pieces removed in English, and in Indian languages. While no other platform discloses this information, it provides valuable insight into the ability of platforms to moderate content across various Indian languages. In a context where platforms have been known to neglect non-English language users,³¹ this metric can help gauge how well platforms are performing on vernacular content moderation within the Indian context as well.

5. Comprehensive disclosures on actions taken for problematic content

Most platforms do not offer any comprehensive disclosure on what takes place when they ‘action’ content. In the case of Facebook and Instagram, numbers are offered on content that has been actioned, and some information is provided on what the consequences of actioning content can be. In the context of reports through the India grievance reporting mechanism, this can mean removing content, covering photos or videos with a warning, and restricting content availability in the country. In the context of actioning content reported through any alternate grievance reporting mechanism, the report only shares that this can include removing content, or covering photos or videos with a warning. Meta’s Transparency Centre discloses that they take additional measures including measures to curb the reduction of spread of borderline content.³² These kinds of ‘actions’ are not discussed within the monthly compliance reports. In the case of WhatsApp, only disclosures in relation to banned accounts. It has been stated that safety grievances and responses to the same are not recorded as actions taken against the grievance report. With respect to YouTube and LinkedIn, only content removals have been discussed. However, there may be alternate actions taken for flagged content, including age-based content restrictions, measures to curb the reduction of spread of borderline content, limiting the visibility of content or labelling content. These kind of ‘actions’ are not discussed within the monthly compliance reports. In the context of Twitter/X, only disclosures in relation to URLs actioned, without disclosure on what sort of action may have been taken.

Taking ‘action’ against content extends beyond content removal, and can include actions such as blurring content, issuing warnings, age-restricting content, and limiting the access of borderline content. More comprehensive disclosures are necessary from almost all platforms when it comes to detailing how they act against the content that has been deemed to be violative or problematic within their platforms.

31. <https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/>

32. <https://transparency.fb.com/hi-in/enforcement/taking-action/>

6. Disclosure on actions taken for repeated violations

Only ShareChat discloses details on numbers of accounts that have been permanently banned, and the instances of graded time-based penalties for multiple violations of their policies. While many platforms have a multiple strikes rule for banning accounts with problematic content, this aspect is not usually disclosed in most compliance reports.

The Intermediary Rules only require disclosures in relation to the number of communication links or pieces of content that have been removed each month. However, disclosures on the occurrences of repeated violations, and the actions taken against such individuals can also help to shed more light on the number of bad actors operating within platforms, and the ways in which platforms act against such bad actors.

7. Disclosure on appeals made regarding content moderation decisions made by the platform

A valuable metric that assesses the effectiveness of content moderation decisions is also the number of instances when users have made appeals against content moderation decisions, and the number of times an earlier decision made by the platform was overturned. While this is not a mandated disclosure obligation under the Intermediary Rules, this metric can work to ensure platforms are holding themselves accountable and being careful in their decisions to take down content.

As it stands, only WhatsApp and Twitter/X disclose information on when account bans or content takedowns have been appealed and content moderation decisions have been overturned.

8. Disclosure of proactive monitoring using automated tools, and the number of content categories being proactively monitored

Platforms are obligated under the Intermediary Rules, to disclose the content that has been taken down on the basis of proactive monitoring using automated tools. Most platforms, barring Snap and ShareChat, explicitly provide data on the number of content pieces or communication links that have been removed on the basis of automated detection. However, among the platforms that have disclosed their use of automated tools for detecting problematic content, there are also variations in disclosures on the content categories for which automated tools are being deployed.

For example, Twitter/X only discloses automated detection of violative content in relation to Terrorism, and Child Sexual Abuse Exploitation, Non-Consensual nudity, and content of similar nature. WhatsApp discloses the number of accounts it has deleted proactively, to prevent harmful activity.³³

33. It may be noted that WhatsApp's capacity to proactively detect harmful activity using automated means may only be limited to its ability to detect spam or bulk messaging, as disclosed in its [White Paper](#).

YouTube discloses the number of removal actions it has undertaken as a result of automated detection, without specifically disclosing the categories of content for which they deploy automated detection tools.³⁴ LinkedIn only discloses the number of content takedowns based on 'content moderation' practices. Which could potentially include the use of both human and automated content moderation practices. Koo discloses the use of automated detection for monitoring violations of spam and community guidelines. Facebook and Instagram also disclose their 'proactive rate' across various content categories, indicating the percentage of violating content that the platforms were able to detect before any user complaints were made. This is also a valuable metric that is being disclosed that enables regulators and users to better understand the contexts within which platforms have the most capacity to proactively detect violative content. While the Intermediary Rules have required that platforms need only disclose the number of communication links that have been removed as a result of proactive monitoring using automated detection mechanisms, the content categories for which such automated tools are deployed is also an important metric to understand a platform's performance in moderating content. As visible in the case of most platforms, the largest number of user complaints seems to be in the context of bullying, harassment and sexual content. These may also be the categories of content that might be most difficult to monitor through automated means. Clearer disclosures need to be made by platforms on the categories of content they choose to proactively monitor, to understand whether the users' most prominent concerns are being adequately addressed.

34. YouTube has specified that they use "automated detection processes for some of our products to prevent the dissemination of harmful content such as child sexual abuse material and violent extremist content". However, it is unclear if they only use automated detection for these categories of content, or if they use it for categories of problematic content as well.



Part C- Disclosures required by SSMLs within other regions and jurisdictions.

Similar to the transparency and disclosure requirements laid down under the Intermediary Rules, there are requirements imposed on social media platforms by other jurisdictions as well. While India is the only jurisdiction that requires monthly compliance reporting from social media platforms, there are several other governments that require disclosures of a similar nature from social media platforms. Some of these countries and their respective regulations are detailed below.

1. The European Union’s Digital Services Act;
2. The European Union’s CSAM Derogation Law;
3. The European Union’s Regulation on addressing the dissemination of terrorist content online;
4. Austria’s Federal Act on measures to protect users on communication platforms (Communication Platforms Act, or Kommunikationsplattformen-Gesetz – KoPI-G);
5. Germany’s Network Enforcement Act (Netzwerkdurchsetzungsgesetz, or NetzDG); and
6. Turkey’s Law no. 5651 (The Law on the Regulation of Broadcasts via Internet and Prevention of Crimes Committed through Such Broadcasts).

The disclosure requirements across these laws vary in nature, with some legislations being limited in scope to specific categories of content such as child sexual exploitation and abuse imagery, or terrorist content. However, at the same time these legislations require disclosures from social media platforms that are similar in nature to those required within India under the Intermediary Rules.

The table provided in **Annexure 1** lays down the disclosure requirements under these various legislations and discusses the extent to which Indian laws require the same from social media platforms.

It may be noted that while these disclosures are required to be made less frequently within these various jurisdictions when compared to India, many of these laws require significantly greater granularity in the disclosures that are made, allowing regulators greater insight into the day-to-day content moderation being undertaken by various platforms.³⁵ Given the ability of platforms to share more granular data within other jurisdictions, instituting similar obligations within the Indian context would also prove to be feasible. In particular, more detailed information on the following metrics, which are disclosed in other jurisdictions may in fact be valuable within the Indian context as well:

- Average monthly active users;
- Number of human content moderators, broken down by applicable official language within the jurisdiction;
- Basis of taking action (violation of law, or violation of terms and conditions of the platform);
- Number of complaints that led to deletion or blocking of content including the stage at which the examination of the complaint led to such deletion or blocking;
- Median time taken to respond to complaints, or overview of the periods between receipt of a complaint start of review process by the platform and the deletion or blocking of illegal content (turnaround time);
- Number of appeals of content moderation decisions, details of decisions taken pursuant to such appeal, the average time taken to arrive at a decision, and the number of instances where an initial content moderation decision was reversed;
- The indicators used by the platform in relation to the accuracy of automated content moderation;
- Error rates or false positives for technologies used to detect violating content;
- Overview of the number and type of cases in which the platform has chosen not to take action.

35. For example, based on disclosures made under EU's Digital Services Act, the European Commission has set up a dashboard that aggregates disclosures made by various platforms and shares insights developed from such disclosures: <https://transparency.dsa.ec.europa.eu/>.

Conclusion and Recommendations

This report sought to examine the effectiveness of the monthly compliance reporting mechanism set up by the Intermediary Rules within India. This was undertaken through three ways a) analysing monthly compliance reports and checking how closely the disclosures in the reports meet the obligations set out under the Intermediary Rules; b) comparing how different platforms provide disclosures, and the instances where they choose to disclose more or less information than other platforms; and c) understanding the kinds of disclosures platforms are obligated to make within other jurisdictions.

Through an examination of the above, it can be seen that the disclosure requirement under Rule 4(1)(d) of the Intermediary Rules was a strong and forward-looking initiative towards ensuring a more transparent and accountable regime for SSMLs within India. At the same time, there is still some way to go before these monthly disclosure obligations translate into meaningful outcomes and contribute towards progressive improvement in the content moderation activities undertaken by SSMLs in India. While most SSMLs broadly adhere to the requirements, the flexibility and intent behind the rules create inconsistencies in how they are applied across different areas. It also cannot be said that any single SSML is performing better or worse than any other SSML in how they choose to meet their monthly disclosure obligations. As Tables 1 and 2 indicate, platforms perform differently across various metrics. The variations across platforms in their disclosure, and in some instances the inadequacy of compliance with disclosure obligations indicate that these disclosures need to be strictly monitored and reviewed to ensure completeness, and consistency in disclosures. In the context of the evolving nature of social media, and the growing use of automated means for content moderation and violation detections, the creation of transparency norms will have to be an iterative and adaptive process to prove successful.



Some key recommendations that might help improve the implementation of the monthly compliance reporting, and work towards creating greater transparency and accountability among significant social media intermediaries are as follows:

1. Mandatory disclosure by all social media intermediaries on number of users:

A key missing metric that is currently not required under the Intermediary Rules is a disclosure by social media intermediaries on their number of users. The notified threshold for being considered a significant social media intermediary under the Intermediary Rules is five million registered users. However, in the absence of a mandatory periodic disclosure by all social media intermediaries on their number of users (either their number of active users or registered users), it is difficult to gauge whether all intermediaries that are meeting the threshold under the Intermediary Rules are publishing their monthly compliance reports. The government may consider implementing a tiered (slab-based) disclosure framework for social media intermediaries, where platforms report user data based on predefined thresholds. This should focus primarily on monthly active users (MAUs) in India, as this metric provides a more accurate reflection of user engagement and activity on the platform, aligning with global best practices. Reporting based on MAUs will ensure transparency and allow for more meaningful regulatory oversight, compared to registered users, which may not accurately reflect current platform usage. Furthermore, such a disclosure will also allow regulators and users to gauge the effectiveness of grievance redressal mechanisms and improvements in proactive monitoring of content relative to the user-base of the platform.

2. Periodic review of SSMI disclosures by regulators and publication of reports with directions to SSMLs requesting additional information when required:

Mandating monthly disclosures from SSMLs must be understood as only the first step of an iterative process for creating an accountable and transparent framework for the operation of SSMLs. The next step also involves periodic review of the reports being released in order to determine where additional information is required from SSMLs where a disclosure is inadequate. For example, the Intermediary Rules and FAQs clearly require a disclosure of the number of communication links that have been disabled because of proactive monitoring, based on automated tools. A periodic review and a publications of the outcomes of such a review can also provide takeaways in terms of disclosures that may need improvement in the context of the evolving nature of online content, or disclosures that certain SSMLs may be making that might be valuable for all SSMLs to make. For instance, a periodic review might have also revealed that Twitter/X is the only SSML that is sharing information on complaints and actions taken in relation to synthetic or manipulated media content. In a context where deepfakes are an ongoing and grave concern, requesting additional disclosures from all SSMLs on their efforts to curb such content is within the scope of the Rule 4(1)(d) of the Intermediary Rules. It may well be the case that such a review is also being undertaken, and additional information is being requested from specific SSMLs.

However, for full compliance across SSMLs, in a manner where such compliance is also publicly visible, periodic review reports could be published by regulators recommending additional disclosures where required, in exercise of the regulator's discretion to require more information under 4(1)(d) of the Intermediary Rules.

3. Requiring more granularity in disclosures made by SSMLs under the Intermediary Rules:

The Intermediary Rules provide considerable flexibility to SSMLs in the contents of their disclosures. This is valuable in many ways, particularly since it offers SSMLs the possibility to disclose what may be feasible for them, and accounts for the fact that they may be diverse in nature, and might also have varying capacities. At the same time however, a degree of uniformity might prove to be valuable for developing insights on SSMLs' comparative performance in content moderation. For example, insights developed based on the SSMLs that provide information only in relation to the India-specific grievance redressal mechanism would not only be incomplete, but also cannot be compared to the insight developed on SSMLs that provide information on complaints received through all their channels. While uniformity in terms of a specific disclosure format for SSMLs may not be compared, there can be greater granularity in the specifications for disclosures that are required. For example, this can mean

a) requiring disclosure of all complaints received from Indian users across all grievance redressal channels;

b) requiring not only information on how many pieces of content was actioned, but also what kinds of action was taken for each complaint; c) requiring disclosures on the average timelines for grievance redressal;

d) requiring disclosures on the number of times a content moderation decision was appealed, and such decision was reversed. A part of the reason why mandating disclosures from platforms is important is because it helps to hold platforms accountable, even in the event that there may be limited follow-up on disclosures made. Ensuring granularity in disclosures would, at the very least, help platforms to mould their content moderation practices and mechanisms into one that is prima facie fair and trustworthy.

4. Need for disclosures in relation to how content moderation activities by SSMLs accounts for the diversity of languages within India:

In a country with 22 official languages, it is necessary that SSMLs operating within India accounts for the diversity of languages within India, and thereby the diversity of language-based content violations that could take place within the platforms offered by the intermediary.

Disclosures offered by SSMLs would therefore also need to account for the number of human content moderators deployed by SSMLs across the Indian languages offered in their platforms, as well as details on the use of automated tools for detecting violating content in vernacular languages, and indicators used to determine the accuracy of such automated tools.

5. Need for disclosures on SSMI efforts towards user protection in the content moderation activities being undertaken:

Given the larger goal of creating a safer, trusted and more accountable internet for India's Digital Nagriks, disclosures also need to focus on ensuring the protection of user rights within the SSMI platforms. This would mean more information would need to be provided through the disclosures on how users are informed of their right to seek redressal, how users may appeal content moderation decisions, measures that have been instituted to prevent bad actors from misusing the grievance redressal mechanism and data on appeals and reversals of content moderation decisions.

Annexure 1:

Details of Social Media Platforms' Disclosure within other Jurisdictions

SSMI Transparency Reporting Obligations within Other Jurisdictions					
Country/Region	Regulation	Relevant Provision	Reporting Period	Disclosures required	Required under Indian Law (Intermediary Rules)
European Union	Digital Services Act (Regulation (EU) 2022/2065)	Article 15: Transparency reporting obligations for providers of intermediary services (except MSEs)	Annual	Orders received from Member State Authorities categorised by type of illegal content including orders to act against illegal content, and orders to provide information (for intermediaries)	No
				Median time taken to take action based on orders received	No
				Number of notices submitted under notice and action mechanism for individuals or entities to report illegal content and trusted flaggers (for hosting services)	Yes
				Action taken pursuant to notices (differentiated by whether action was taken on the basis of the law, or the T&Cs of the provider)	Yes
				Median Time taken for action under such notices	No
				Number of notices processed by automated means	No
				Content moderation at provider's own initiative, including measures to provide training and assistance to content moderators, number and types of measures that affect availability, visibility and accessibility of information provided by recipients of the service,	No

				and the recipients ability to provide information, categorised by the type of illegal content or violation of T&Cs, detection method, and type of restriction applied (for providers of intermediary services)		
				Number of complaints received through internal complaints handling system (appeals of content moderation/account suspension/service suspension decisions), basis for complaints, and decision taken in respect of complaints, average time to arrive at decision, and number of instances where the decision was reversed	No	
				Use of automated means for content moderation, including a qualitative description, a specification of precise purpose, indicators of the accuracy and possible rate of error of automated means used in fulfilling those purposes and safeguards applied.	No	
			Article 24: Transparency reporting obligations for providers of online platforms	Annual	Number of disputes submitted to certified out-of-court dispute settlement bodies, outcomes of dispute settlement procedures, median time needed for completion of dispute settlement procedures, and share of disputes where the provider of the online platform implemented the decisions of the body	No
					Number of suspensions imposed, distinguishing between suspensions for providing manifestly illegal content, manifestly unfounded notices and complaints	No
				Semi-annual	Average monthly active recipients of the service	No
			Article 42: Transparency reporting obligations for very large online platforms or very large online search engines	Semi-annual	Reports under Article 15 (mentioned above)	NA

				Human resources for content moderation, broken down by each applicable official language for Member States, including for compliance with notices mechanism, trusted flaggers, and internal-complaints handling.	No
				Qualifications and linguistic experience of human content moderators, and training and support given	No
				Indicators of accuracy and related information for automated content moderation broken down by member state official language	No
				Average monthly recipients of the service	No
European Union	EU CSAM Derogation Law (Regulation(EU)2021/1232)	Article 3(1)(g)(vii)- Report on personal data processing providers of personal and other data in connection with the provision of number-independent interpersonal communications services	Annual	The type and volumes of data processed	No
				Specific grounds relied on for data processing	No
				Grounds relied on for transfers of personal data outside the EU	No
				Number of cases of online child sexual abuse identified, differentiating between online child sexual abuse material and solicitation of children	No
				Number of cases in which a user has lodged a complaint with the internal redress mechanism or with a judicial authority, and the outcome of such complaints	No
				The number and ratios of errors (false positives) of the different technologies used	No
				The measures applied to limit the error rate, and the error rate achieved	No
				The retention policies and data protection safeguards applied	No

				Names and organisations acting in the public interest against child sexual abuse with which the data has been shared	No
European Union	Regulation (EU) 2021/784 on addressing the dissemination of terrorist content online	Article 7- Transparency Obligations for Hosting Service Providers	Annual	Information about measures in relation to identification and removal of, or disabling of access to terrorist content	No
				Information about the hosting service provider's measures to address reappearance online of material which has previously been removed or to which access has been disabled because it was considered to be terrorist content, in particular where automated tools have been used	No
				The number of items of terrorist content removed or to which access has been disabled following removal orders or specific measures, and the number of removal orders where the content has not been removed or access to which has not been disabled, with grounds therefor.	No
				The number and outcome of complaints handled by the hosting service provider	No
				The number and the outcome of administrative or judicial review proceedings brought by the hosting service provider	No
				The number of cases in which the hosting service provider was required to reinstate content or access thereto as a result of administrative or judicial review proceedings	No
				The number of cases in which the hosting service provider reinstated content or access thereto following a complaint by the content provider.	No

Austria	Federal Act on measures to protect users on communication platforms (Communication Platforms Act)	Section 4: Reporting Obligation (applicable to communications platform service providers whose registered users are more than 100,000 people in the previous calendar year, and sales revenue from operation in Austria was more than EUR 500,000)	Annual (less than 1 million registered users), and semi-annual (more than 1 million registered users)	General information on efforts of service provider to prevent illegal content	
				Description of design and user-friendliness of reporting procedure, and decision making criteria for deletion or blocking of illegal content, including steps taken to determine whether content is illegal, and whether contractual provisions between service provider and user have been violated	No
				Description of number of reports of allegedly illegal content received during the reporting period	Yes
				Overview of the number of reports of allegedly illegal content that led to the deletion or blocking of the content reported during the reporting period, including information on which stage of the examination led to the deletion or blocking, as well as a summary description of the type of content	Somewhat
				Overview of the quantity, content and result of the review procedures	Yes
				Description of the organisation, number of staff and technical equipment available, and the technical competence of the staff responsible for processing reports and for the review procedures, as well as the training and supervision of the persons responsible for processing reports and reviews	No
				Overview of the periods between receipt of the report by the service provider, the start of the review and the deletion or blocking of illegal content, broken down into the periods “within 24 hours”, “within 72 hours”, “within seven days” and “at a later point in time”	No
				Overview of the number and type of cases in which the service provider has refrained from carrying out a reporting and review procedure	No

Germany	Network Enforcement Act (Netz-DG)	Section 2: Reporting obligations (for social network providers who receive more than 100 complaints about illegal content in a calendar year)	Semi-annual	General information about what efforts the social network provider makes to prevent illegal content on the platforms	No
				Type, basic principles of functionality and scope of any procedures used to automatically identify content that is to be removed or blocked, including general information about the training data used and the review of the results of these procedures by the provider, as well as information about the extent to which scientific circles and research are supported in the evaluation of these procedures and they have been granted access to the provider's information for this purpose.	No
				Presentation of the mechanisms for submitting complaints about illegal content, description of the decision-making criteria for the removal and blocking of illegal content and description of the review process including the order of checking whether there is illegal content or whether contractual provisions between the provider and the user are being violated.	No
				Number of complaints about illegal content received in the reporting period, broken down into complaints from complaint bodies and complaints from users and the reason for the complaint.	Somewhat
				Organization, staffing, technical and linguistic competence of the work units responsible for processing complaints and training and support of the people responsible for processing complaints	No
				Membership in industry associations with an indication of whether there is a complaints office in these industry associations	No

			Number of complaints where an external body was consulted to prepare the decision	No
			Number of complaints that led to the deletion or blocking of the disputed content in the reporting period, according to the total number and broken down into complaints from complaint bodies and from users, according to the reason for the complaint, which step in the examination sequence led to the removal or blocking and whether a transfer to a recognized body of regulated self-regulatory organisation took place	No
			The number of complaints about illegal content that, after receipt, led to the removal or blocking of the illegal content within 24 hours, within 48 hours, within a week or at a later date, additionally broken down into complaints from complaint offices and users as well as broken down according to the reason for the complaint	No
			Measures to inform the complainant and the user for whom the disputed content was stored about the decision on the complaint,	No
			Number of counters (appeals) received in the reporting period in accordance with Section 3b Paragraph 1 Sentence 2, according to the total number and broken down into counters (appeals) from complainants and from users for whom the disputed content was saved, along with information on how many cases the counter-presentation was remedied	No
			Information about whether and to what extent scientific and research circles were granted access to the provider's information during the reporting period in order to enable them to carry out an anonymous evaluation of content blocking, content spread, and content moderation practices	No

				Other measures taken by the provider to protect and support those affected by illegal content	No
				A summary with a tabular overview showing the total number of complaints received about illegal content, the percentage of content removed or blocked as a result of these complaints, the number of counters (appeals) and compares the percentage of decisions changed in response to appeals with the corresponding figures for the two previous reporting periods, combined with an explanation of significant differences and their possible reasons.	Somewhat
				Explanation of the provisions in the provider's general terms and conditions regarding the permissibility of distributing content on the social network used by the provider for contracts with consumers	No
				Description of the extent to which the provider's terms and conditions are in accordance with provisions of the German Civil Code and other laws.	No
Turkey	Law no. 5651	Additional Article 4, clause 4 (applicable to social network providers with over 1 million daily visits from Turkey)	Semi-annually	Statistical and categorical information regarding the implementation of decisions to remove and/or block content	Yes
				Information on their algorithms, advertising policies and transparency policies regarding title tags, featured or reduced access content	No
				Information on complaints received, responses to complaints and turnaround time	Somewhat
				Information on advertisements including information on content, advertiser, advertisement duration, target audience, number of people or groups reached.	No

Notes on data collection:

- The analysis of this report is based on publicly available compliance reports released by nine social media platforms (categorised as Significant Social Media Intermediaries under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021).
- The compliance reports taken for consideration are from the period between May/June 2021-December 2023.
- The data from these reports was collected between the months of January and April 2024.
- It may be noted that the reports were published by SSIMs in varied ways: five SSIMs offered downloadable PDFs of reports, one SSIM redirected to inconsistently available Google Drive links, and three SSIMs provided links connecting to the relevant webpage containing the compliance report for that particular month. In all cases downloadable pdfs, the reports are available for review. In cases where SSIMs offer hyperlinks to webpages containing the reports, it is noted that at the time of publication of this report, some of the hyperlinks were broken or redirected to the wrong webpage.
- While attempt was made to be thorough, any oversight may be communicated to relations@igap.in.



WHAT IS IGAP ?

The Indian Governance And Policy Project (IGAP) is an emerging think tank focused on driving growth, innovation, and development in India's digital landscape. Specializing in areas like AI, Data Protection, FinTech, and Sustainability, IGAP promotes evidence-based policymaking through interdisciplinary research. By working closely with industry bodies in the digital sector, IGAP provides valuable insights and supports informed decision-making. Core work streams include policy monitoring, knowledge dissemination, capacity development, dialogue and collaboration,

