INFOCOMM MEDIA DEVELOPMENT AUTHORITY

ONLINE SAFETY ASSESSMENT REPORT 2025

Designated Social Media Services

CONTENTS

Executive Summary

01 The Online Safety Assessment Report (OSAR) 2025 outlines IMDA's assessment of the online safety measures that Designated Social Media Services (DSMSs) have in place, as required by the Code of Practice for Online Safety – Social Media Services (the Code) to enhance user safety and mitigate risks from harmful content, for the period from 1 April 2024 to 31 March 2025. The six DSMSs are Facebook, HardwareZone, Instagram, TikTok, X and YouTube.

02 IMDA's main priority as Singapore's online safety regulator is to ensure a safe online environment for users in Singapore and to protect children, in particular, from harmful content. Under the Broadcasting Act (BA), IMDA has powers to direct social media services to block access to egregious content found on their services. The BA also empowers IMDA to issue Codes of Practice for Online Safety. Rather than focusing on individual pieces of harmful content, these Codes place obligations on platforms to put in place systems and processes to tackle harmful content. IMDA holds the DSMSs accountable for meeting their legally binding obligations under the Code. We assess their compliance with the Code using a standardised methodology, based on the DSMSs' annual online safety reports submitted to IMDA, as well as IMDA's own robust testing of their online safety measures.

03 In addition, IMDA has engaged the DSMSs on their online safety measures, detected risks, flagged harmful content, and raised concerns with the DSMSs throughout the year. While IMDA adopts a collaborative approach to engage the DSMSs, we will hold the DSMSs accountable when we find that their online safety measures do not adequately achieve the outcomes of the Code. IMDA will not hesitate to exercise its regulatory powers under the BA to remedy systemic breaches of the Code. Our overriding objective is to ensure the online safety of Singapore users, especially children.

Summary of Key Findings

04 The aim of the OSAR is to allow Singapore users to make informed decisions about the risks and available safety measures on each DSMS and to ensure that DSMSs are held accountable for providing a safe user experience. In OSAR 2025, while there has been improvement in some areas, IMDA has also identified areas of serious weakness that the DSMSs must take immediate action to rectify.

05 Consequently, Letters of Caution have been issued to X and TikTok for serious weaknesses in their measures to proactively detect and remove significant numbers of child sexual exploitation and abuse material (CSEM) content and terrorism content respectively. X and TikTok have accepted IMDA's findings and committed to putting in place rectification measures. While we welcome their commitment, both X and TikTok are placed under Enhanced Supervision by IMDA until they demonstrate improvement.

06 IMDA has also identified areas of weakness for all DSMSs which they will need to account for in their next annual online safety reports. We note that the DSMSs have improved in some of the areas of weakness highlighted in the OSAR 2024, and we urge the DSMSs to continue strengthening these measures.

Areas of Serious Weakness

07 X's measures to proactively detect and remove CSEM, as well as TikTok's measures to proactively detect and remove terrorism content, were assessed to have serious weaknesses. Paragraph 15 of the Code requires DSMSs to minimise Singapore users' exposure to CSEM and terrorism content through the use of technologies and processes that proactively detect and swiftly remove such content on their services. Both CSEM and terrorism content are very egregious harms that must be proactively detected and removed before users encounter them.

a. **X** – IMDA detected a 120% increase in CSEM cases on X originating from or targeting Singapore users from 2024 (33 cases) to 2025 (73 cases).

   i. In 2024, IMDA detected 33 cases of CSEM on X originating from or targeting Singapore users that was not proactively detected and removed by X. These cases involved content sharing or linking to CSEM, and self-generated CSEM. These cases had a Singapore nexus and featured well-established indicators that are commonly associated with CSEM, such as coded words and number patterns indicating that a user is under 18 years of age.

   ii. In 2025, IMDA detected 73 cases of CSEM on X with the same characteristics as those detected in 2024. This occurred despite IMDA sharing with X its analysis of those CSEM cases and their indicators in 2024. In addition, the measures X reported to address CSEM in its latest annual online safety report did not explain how it addressed the specific types of CSEM that IMDA flagged to X.

b. **TikTok** – IMDA detected terrorism content shared by Singapore-based accounts for the first time in 2025.

   i. In 2025, IMDA detected 17 cases of terrorism content shared by Singapore-based accounts on TikTok that was not proactively detected and removed. The content included videos with well-established indicators that the content was terrorism-related, in particular the use of edited footage or audio related to known transnational terrorist organisations that were in some cases blended with benign content. In some cases, the terrorism-related audio was also concealed under TikTok's "original sound" label, which adds the audio to TikTok's database for others to use in their posts as well. IMDA has shared with TikTok our analysis of the content and indicators associated with the terrorism content.

   ii. In addition, when some of the terrorism content was user reported to TikTok via its in-app user reporting mechanism, TikTok found the content non-violating. TikTok only removed the content after IMDA flagged them to TikTok. While these cases should have been proactively detected and removed by TikTok in the first place, as required by paragraph 15 of the Code, it also demonstrates the weakness of TikTok's user reporting system, which is also outlined in paragraph 12b below.

08 IMDA has issued Letters of Caution to X and TikTok regarding their measures to proactively detect and remove CSEM and terrorism content respectively. X and TikTok have committed to put in place specific measures to rectify these issues. Additionally, both services have been placed under Enhanced Supervision, in which they must meet with IMDA regularly to account for their progress in implementing the rectification measures they have committed to, until IMDA is satisfied that the issues are adequately resolved. X and TikTok are also required to provide supporting data and information to IMDA in their next annual online safety report due on 30 June 2026, to demonstrate the effectiveness of their implementation of the rectification measures. Should X or TikTok fail to satisfy IMDA that they have improved their measures to address the specific types of CSEM and terrorism content that IMDA has detected, IMDA will not hesitate to explore further options, including potential regulatory action under the BA.

Areas of Weakness

Child Safety

09 Facebook, YouTube and HardwareZone were found to have weaknesses in the effectiveness of their child safety measures, which could lead to children easily accessing age-inappropriate content. The comprehensiveness of child safety measures across different DSMSs also varied greatly. Instagram and TikTok reported the most comprehensive child safety measures, while HardwareZone and X only had a few baseline measures. Given the rapidly evolving online safety risk landscape, especially for children, DSMSs must continue to prioritise enhancing the comprehensiveness and effectiveness of their measures to minimise children's exposure to harmful and age-inappropriate content.

a. Facebook and YouTube continued to have instances where children's accounts could access some harmful and age-inappropriate content. This included digital imagery of adult nudity on Facebook, and videos with partial nudity and audio sexual depictions on YouTube. However, such instances were low in number, with no indications suggesting that this was a systemic issue on either service. Nevertheless, IMDA has shared these instances with Facebook and YouTube to analyse and improve their measures.

b. HardwareZone's measures to restrict children from accessing its service were found to be ineffective. However, HardwareZone has taken steps to improve its measures after engagement by IMDA.

    i. In 2024, IMDA notified HardwareZone that it needed to either effectively restrict children from accessing its service or put in place comprehensive safety measures for children as required by paragraph 20 of the Code. This was because HardwareZone's Terms of Service prohibit users under the age of 18 from accessing the service, but its age-gating measure to enforce this could easily be bypassed. As a result, children could access age-inappropriate content such as sexually suggestive references or innuendos.

    ii. In 2025, IMDA's tests found that HardwareZone's age-gating measure could still be easily bypassed. However, after multiple engagements by IMDA, HardwareZone submitted a plan to implement an updated age verification system and content classification system, which came into effect in end-January 2026. HardwareZone's updated age verification system utilises the sgID service that is part of the Singpass app and it works together with its content classification system to determine access levels and prevent the viewing of posts that the system classifies as inappropriate for children. IMDA's preliminary assessment is that these new measures by HardwareZone to restrict children's access to age-inappropriate content meet the requirements of the Code. IMDA will assess the effectiveness of these measures after HardwareZone submits its next annual online safety report due on 30 June 2026.

Accountability: Provision of Singapore-Specific Data

10 DSMSs must be transparent and accountable to users by providing clear information on how they are keeping their service safe for users. This allows users to make informed choices on which DSMS to use and how to keep themselves and their children safe online. In addition, DSMSs are encouraged to provide data to demonstrate the effectiveness of their measures for Singapore users. The data submitted by the DSMSs in 2025 was similar to 2024.

a. Facebook, Instagram and YouTube still could not provide Singapore-specific data on the effectiveness and timeliness of their user reporting and resolution mechanisms. Facebook and Instagram stated that the focus of their data collection efforts was for proactive content moderation rather than for user reporting. YouTube also reiterated that it did not collect the required data from its in-app user reporting system.

11 DSMSs are also responsible for providing a safe experience for their users. Therefore, it is important for DSMSs to be aware of the situation on their services so that they can have the appropriate measures to protect their users. In this regard, IMDA observed that DSMSs reported large changes in their year-on-year data on the volume of harmful content they removed in Singapore between 2024 and 2025, but they did not provide explanations to account for these changes in the first instance.

a. For example, Instagram reported a 212% increase in content removed under its "Child Endangerment" policy, from 4,712 pieces of content removed in 2024 to 14,700 removed in 2025, but it did not provide an explanation to account for this increase initially. YouTube also reported a 59% increase in the monthly average of videos removed under its "Child Safety" policy, from an average of 769 videos removed per month in 2024 (6,917 videos over 9 months) to 1,220 videos removed per month in 2025 (14,644 videos over 12 months). Similarly, YouTube did not provide an explanation to account for this increase initially.

b. Without explanations, it is challenging to make sense of the data and its impact on online safety in Singapore. Such large increases in harmful content removed on a particular service could be the result of either increased prevalence of harmful content or increased enforcement by DSMSs.

c. All DSMSs have since provided explanations when asked by IMDA. The reasons they cited included policy changes leading to increased enforcement, new detection technologies, and the identification of new types of harmful content.

d. IMDA has also been engaging DSMSs on improving their information transparency to Singapore users, so that Singapore users are better equipped with meaningful information to appreciate the online safety risks on the DSMSs and make informed decisions regarding the DSMSs they use.

Areas where DSMSs have Improved

User Reporting and Resolution

12 All DSMSs showed improvement in the effectiveness and timeliness of their responses to user reports, except for TikTok which declined in the effectiveness of its user reporting measures.

a. Based on IMDA's Mystery Shopper tests in 2025, all DSMSs except TikTok significantly improved their action rates on legitimate user reports, i.e. where content violated their own community guidelines (see results in Table 1). With the exception of TikTok, their action rates ranged from 54% to 93% in 2025. In contrast, the action rates on legitimate user reports in 2024 were approximately 50% or less for all DSMSs except HardwareZone. This indicates that most DSMSs took steps to improve their user reporting measures after IMDA highlighted this area of weakness to them in the first OSAR.

b. TikTok was the only DSMS to decline in the effectiveness of its user reporting measures. Its action rate declined from 39% in 2024 to 25% in 2025, which suggests that TikTok did not take the appropriate action on three out of four legitimate user reports. TikTok needs to demonstrate significant improvement in its user reporting measures and provide an update to IMDA on the steps it has taken to do so in its next annual online safety report.

c. All DSMSs also improved on the time they took to act on user reports of harmful content that violated their own community guidelines in 2025. Their average time to action was between 2 to 5 days (see results in Table 2).

DSMS	Action Rates on User Reports of Harmful Content that Violate their Community Guidelines	
	2024	2025
Facebook	53%	81% ▲
HardwareZone	89%	93% ▲
Instagram	2%	54% ▲
TikTok	39%	25% ▼
X	54%	74% ▲
YouTube	46%	68% ▲

Table 1: DSMSs' Action Rate on User Reports from IMDA's "Mystery Shopper" Tests in 2024 and 2025

DSMS	Average Time to Action[1]	
	2024	2025
Facebook	9 days	4 days
HardwareZone	3 days	2 days
Instagram	7 days	4 days
TikTok	5 days	4 days
X	10 days	5 days
YouTube	5 days	4 days

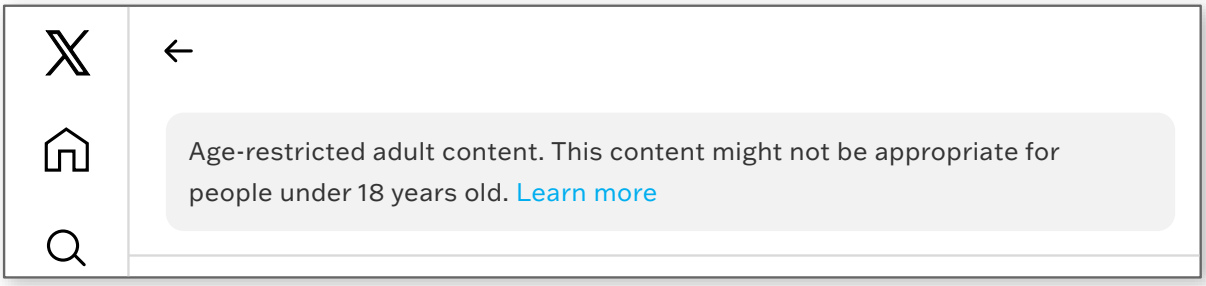
Table 2: DSMSs' Time to Action on User Reports from IMDA's "Mystery Shopper" Tests in 2024 and 2025

Child Safety

13 X improved the enforcement of its policies to restrict children's accounts from viewing adult sexual content.

a. In 2024, IMDA's tests found that X did not effectively enforce its policies to restrict children's accounts from viewing adult sexual content. Children's accounts could easily find and access explicit adult sexual content on X, especially hardcore pornography, with simple search terms.

b. In 2025, IMDA's tests found that X had improved the enforcement of its policies. It was more difficult to find and access explicit adult sexual content on X using children's accounts compared to 2024. Most explicit adult sexual content was appropriately age-restricted by X.



Screenshot of X's message for age-restricted adult content

c. Nevertheless, as some explicit adult sexual content could still be accessed using children's accounts, X should continue to improve its safety measures for children.

[1] The average time to act indicates the turnaround time for DSMSs to resolve user reports.

14 DSMSs have strengthened their safety policies and introduced new safety measures for children in 2025.

a. Instagram has enhanced its safety features for Teen Accounts. Users under 18 years old will be automatically placed into the most restrictive content setting, with parental permission required to change to a less restrictive setting.

b. TikTok introduced a new feature that allows parents to switch their teen's accounts back to default private setting if their teen has made their profile public.

c. YouTube introduced a new Family Center hub where parents can see shared insights into their teen's channel activity. YouTube also strengthened the enforcement of its policies on "violent or graphic content" and "illegal or regulated goods or services" by age-restricting additional types of content in these categories, such as fictional violence with graphic scenes and certain types of online gambling content like online casino promotions.

Online Safety Ratings

15 Each DSMS received (a) an **Overall Rating** and (b) **Ratings for Individual Sections of the Code**. The Overall Ratings show whether the DSMSs' online safety measures have met the baseline standard. The Ratings for Individual Sections of the Code indicate how well each DSMS's measures in that particular section have met the Code's requirements.

DSMS	Overall Rating	Ratings for Individual Sections of the Online Safety Code			
		Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
Facebook	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑
HardwareZone	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑
Instagram	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑
TikTok	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑
X	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑
YouTube	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑	☑☑☑☑☑

Prevalence of harmful content in Singapore in 2025

16 In their annual online safety reports submitted to IMDA, all DSMSs reported the number and types of harmful content in Singapore that was removed by them proactively and/or as a result of user reports. We have used this data to analyse the prevalence of harmful content in Singapore. This can inform public education and future regulatory efforts. DSMSs should also pay attention to both existing and emerging harms on their platforms and take the necessary measures to protect users.

17 Based on their data, **Sexual content**, **Violent content**, **and Cyberbullying content were the top three types** of harmful content in Singapore that were removed by the DSMSs proactively and/or as a result of user reports (see Table 3).

a. Sexual content was the most prevalent, ranked in the top three for all DSMSs.

b. Violent content was also prevalent, ranked in the top three for all DSMSs.

c. Cyberbullying content was the third most prevalent type of harmful content, ranked in the top three for four out of the six DSMSs, i.e. Facebook, HardwareZone, Instagram and X.

Rank	Facebook	HardwareZone	Instagram	TikTok	X	YouTube
1	Sexual content	Sexual content	Violent content	Sensitive and Mature Themes[2]	Sexual content	Harmful or Dangerous[3]
2	Violent content	Cyberbullying content	Cyberbullying content	Regulated Goods and Commercial Activities[4]	Cyberbullying content	Child Safety[5]
3	Cyberbullying content	Violent content	Sexual content	Mental and Behavioural Health[6]	Violent content	Nudity or Sexual

Table 3: Top three types of harmful content in Singapore that were removed by the DSMSs proactively and/or as a result of user reports

[2] TikTok's "Sensitive and Mature Themes" policy encompasses both Sexual and Violent content categories, where "some types of body exposure or sexual behaviour", "shocking and graphic content", and "animal abuse" are not allowed.

[3] YouTube's "Harmful or Dangerous" policy encompasses Violent content, where content such as "acts performed by adults that have a risk of serious harm or death", "mutilation and blunt force trauma", and "instructions to kill or severely harm others" are not allowed.

[4] TikTok's "Regulated Goods and Commercial Activities" policy does not allow the "trading, marketing, or providing access" to regulated good or services, such as "illegal drugs", "firearms", and "gambling-like activities".

[5] In addition to YouTube's "Nudity or Sexual" policy, its "Child Safety" policy also encompass Sexual content, where "sexually explicit content featuring minors" is not allowed.

[6] TikTok's "Mental and Behavioural Health" policy does not allow content that promotes "Suicide and Self-Harm", "Disordered Eating, Risky Weight Management, and Body Image", and "Dangerous Activity and Challenges".

Conclusion and Next Steps

18 Since the introduction of the Code in July 2023, DSMSs have stepped up efforts to implement online safety measures for Singapore users. The Code has also resulted in greater accountability from the DSMSs. However, as outlined in OSAR 2025, areas of serious weakness remain or have emerged, and DSMSs must do more to improve the effectiveness of their measures. This is especially important because online safety risks continue to evolve, as technology has made it easier to create and disseminate harmful content. DSMSs must remain vigilant and continually improve the effectiveness of their online safety measures to better protect Singapore users from online harms, especially children.

19 IMDA is constantly monitoring the rapidly evolving online safety risk landscape and reviewing the relevance of its regulations including the Code. We also expect platforms to continuously raise their standards and improve their online safety measures to tackle emerging online safety risks. In 2025, IMDA made it a requirement for Designated App Distribution Services to implement age assurance measures to ensure children do not download apps that are inappropriate for their age. To ensure that online safety measures are effectively and accurately applied to children, IMDA plans to extend age assurance requirements to DSMSs. We are also studying how online safety requirements for children can be further enhanced. IMDA is currently in discussions with DSMSs and more details will be announced later this year.

IMDA's Online Safety Assessment Report 2025 and the DSMSs' annual online safety reports are published in full on IMDA's website at www.imda.gov.sg/online-safety for public reference.

Main Report

Introduction

01 Social media has become ubiquitous in our everyday lives. While there are benefits, Singapore users are increasingly exposed to harmful content. The Ministry of Digital Development and Information's (MDDI) Perceptions of Digitalisation Survey published in 2025 found that more than four in five Singapore residents reported encountering harmful content.

02 In November 2022, Singapore passed the Online Safety (Miscellaneous Amendments) Act (OSMAA) which introduced a new section to the Broadcasting Act to regulate Online Communication Services, such as Social Media Services and App Distribution Services. The OSMAA empowered IMDA to issue legally binding Codes of Practice and designate services with significant reach or impact in Singapore to comply with such Codes.

03 IMDA issued the Code of Practice for Online Safety – Social Media Services (the Code) which took effect from 18 July 2023. Six social media services were designated for a start: Facebook, HardwareZone, Instagram, TikTok, X, and YouTube. These Designated Social Media Services (DSMSs) must have in place system-wide measures to minimise Singapore users' exposure to, and mitigate the impact of, harmful content on their services. The six categories of harmful content that the DSMSs must address are: (1) Sexual content, (2) Violent content, (3) Suicide and self-harm content, (4) Cyberbullying content, (5) Content endangering public health, and (6) Content facilitating vice and organised crime.

Code of Practice for Online Safety – Social Media Services

04 The Code outlines system-level measures to minimise users' exposure to harmful content. The Code takes an outcomes-based approach and gives DSMSs the flexibility to design their measures to meet the intent. See Annex A for the full text of the Code. The Code requires DSMSs to:

a. **Enhance online safety by minimising users' exposure to harmful content, with enhanced protections for children (users under the age of 18)**

   i. Put in place systems and processes to address harmful content, including community guidelines and effective content moderation measures.

   ii. Empower users with tools to manage their own safety.

iii. Proactively detect and swiftly remove child sexual exploitation and abuse material (CSEM) and terrorism content.

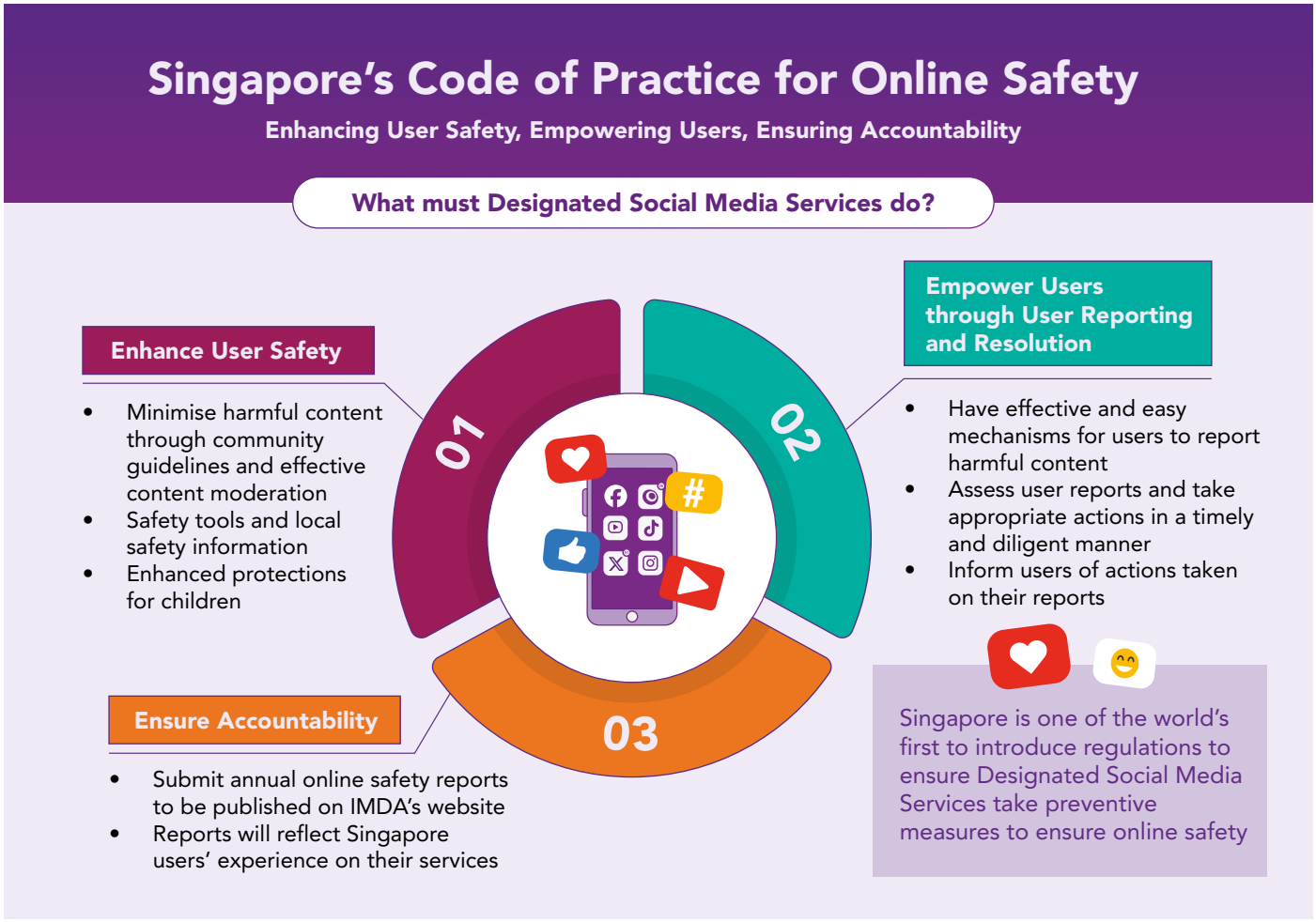
iv. Have enhanced protections for children including age-appropriate policies and tools for parents/guardians to manage their children's safety.

b. **Empower users with effective and easy-to-use mechanisms to report harmful content**

i. Take appropriate action on user reports in a timely and diligent manner and inform these users of the decision and any action taken in response to the reports.

c. **Ensure transparency and accountability to users by submitting annual online safety reports**

i. The reports must contain clear information on the DSMSs' safety measures, supported by suitable data that reflects the impact of their safety efforts in Singapore. This will enable users to make informed choices on which DSMSs would be best placed to provide safe user experiences.



Aim of the Online Safety Assessment Report

05 The first Online Safety Assessment Report (OSAR) published by IMDA in February 2025 aimed to inform Singapore users of the online safety measures DSMSs have in place, as required by the Code.

06 This second edition of the OSAR aims to update Singapore users of:

a. The latest assessment of the DSMSs' online safety measures as required by the Code.

b. The progress DSMSs have made in enhancing their online safety measures since the first OSAR.

c. The continued and new areas of weakness in the DSMSs' online safety measures that they must address.

07 This allows Singapore users to make informed decisions for themselves and their children about the online safety risks and available safety measures when using the various DSMSs, and ensures that the DSMSs are accountable for providing a safe experience for their users.

Methodology

08 The OSAR was prepared using the following sources:

a. Information and data from the DSMSs' annual online safety reports covering the period from 1 April 2024 to 31 March 2025, and

b. Empirical data such as: (i) harmful content detected by or reported to IMDA by members of the public and other public agencies, and (ii) data from our testing of the DSMSs' online safety measures.

09 For each requirement in the Code, IMDA assessed the DSMSs according to whether the online safety measures were **Present**, **Comprehensive**, and **Effective**.

10 The **Effectiveness** of the DSMSs' measures was assessed via the following methods:

a. **Setting up test accounts** to simulate the real-world experience of Singapore users. For example, to assess the effectiveness of the DSMSs' child safety measures, IMDA tested how easily children's accounts could access content that was restricted for them under the DSMSs' own community guidelines and policies, as well as the presence of tools in place to manage the experience of children's accounts.

b. **Mystery Shopper tests** were conducted to: (i) test the effectiveness of the DSMSs' user reporting and resolution mechanisms, by reporting harmful content that violated the DSMSs' own community guidelines and measuring if they took the appropriate action and in a timely manner, and (ii) assess if CSEM and terrorism content were proactively removed by the DSMSs.

Examples of How Testing was Conducted

Example 1: Comprehensiveness of DSMSs' measures to actively offer relevant safety information to users who used high-risk search terms

Methodology: The Code requires DSMSs to actively offer relevant online safety information to users who use high-risk search terms. For this test, common keywords related to harmful content such as "suicide", "cyberbullying", "domestic violence", "sexual assault", "depression", etc. were searched to see what relevant safety information was actively offered to users.

Findings: The test found that all six DSMSs offered relevant online safety information to varying degrees. When keywords related to suicide and self-harm were searched, all DSMSs offered information such as the Samaritans of Singapore hotline. Some DSMSs also provided resources for keywords related to depression, domestic violence and sexual violence, such as the contact details of the Institute of Mental Health and AWARE. In addition, some DSMSs also provided mental well-being resources, help pages, tips from professionals or prompts such as to contact a trusted person.

Example 2: Effectiveness and timeliness of DSMSs' user reporting and resolution mechanisms [Mystery Shopper Test]

Methodology: The Code requires DSMSs to assess user reports and take the appropriate action in a timely manner that is proportionate to the severity or imminence of the potential harm. The DSMSs are expected to enforce their own community guidelines effectively. For this test, harmful content that violated the DSMSs' own community guidelines was reported via their user reporting mechanisms. The action and time taken were recorded. Harmful content that was not actioned on when reported was then flagged to the DSMSs by IMDA. The remaining content was subsequently actioned on by DSMSs for violating their own community guidelines.

Findings: Since the first Report, most DSMSs have shown improvement in the effectiveness and timeliness of their response to legitimate user reports of harmful content that violated their own community guidelines. Please refer to the Executive Summary and Main Report for more details on the findings.

Detailed Assessments of Designated Social Media Services

Facebook

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
☑☑☑☑☐	☑☑☑☑☑	☑☑☑☑☐	☑☑☑☑☐	☑☑☑◪☐☐

Overall Assessment

01 Facebook's overall online safety rating improved slightly from 2024. Notably, the effectiveness and timeliness of Facebook's user reporting and resolution mechanisms improved from 2024.

02 However, Facebook needs to improve on the enforcement of its community guidelines for children. Similar to 2024, Facebook had instances where children's accounts could access harmful and age-inappropriate content that should have been restricted under its community guidelines.

03 Furthermore, Facebook still did not provide Singapore-specific data on the effectiveness and timeliness of its user reporting and resolution mechanisms.

Section Ai: User safety measures for all end-users

04 Facebook had the required user safety measures for all users. Measures that were newly reported in Facebook's latest annual online safety report include the following:

a. Facebook worked with the Mental Health Coalition to establish Thrive, a program for participating technology companies to share signals about violating suicide and self-harm content.

b. Facebook introduced a new "friends tab" that is designed to show content only from a user's Facebook friends on the service. This allows the user to only view updates and posts from people they know.

c. Facebook partnered with EYEYAH! to conduct two workshops for educators in Singapore. Participants learned new tools and strategies to enrich student learning on digital well-being themes, such as anxiety, addiction, and cyber bullying.

05 Beyond the data provided by Facebook in its annual report, IMDA observed an emerging trend on Facebook involving adult sexual content, where users exploited gaps in Facebook's automated content moderation systems by incorporating explicit content within seemingly benign videos. IMDA had flagged this content to Facebook, which confirmed that it was implementing measures to proactively detect and remove such content, including enhancing classifiers, updating data labelling protocols, and providing additional training for internal review teams.

Section Aii: User safety measures for children

06 Facebook had the required user safety measures for children. Facebook reported that it has updated its Transparency Centre with a new page on age-appropriate content, which aims to help parents and teens understand its multi-layered approach to prevent teens from seeing sensitive or mature content.

07 However, Facebook needs to improve on the enforcement of its community guidelines for children. Children could still access some harmful and age-inappropriate content on Facebook, despite the safety measures it has reported.

a. Similar to 2024, IMDA detected a few instances on Facebook where children's accounts could access harmful and age-inappropriate content that should have been restricted under its community guidelines. These included digital imagery of sexual content and adult nudity depicting genitalia. Following IMDA's flagging of this violative content, Facebook subsequently took the appropriate action on all of them.

b. IMDA has engaged Facebook on the need to improve the enforcement of its community guidelines for children. Facebook will need to provide an update on this in its next annual online safety report.

Section B: User reporting and resolution

08 The effectiveness and timeliness of Facebook's user reporting and resolution mechanisms showed improvement from 2024. IMDA's Mystery Shopper tests found that Facebook had a higher action rate on user reports of harmful content and a lower average response time to act on these user reports in 2025. However, Facebook should continue to improve the performance of its user reporting and resolution mechanisms.

a. Facebook took action on 81% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 53% in 2024.

    i. Notably, as compared to its average action rate, Facebook took action on only 60% of Sexual content that was user reported and failed to act on the remaining 40% until IMDA notified Facebook to review them again.

b. Facebook's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 4 days in 2025 from an average time of 9 days in 2024.

Section C: Accountability

09 Facebook should improve the provision of data in its annual online safety report. In particular, Facebook was unable to provide Singapore-specific data to demonstrate the effectiveness and timeliness of its user reporting and resolution mechanisms for specific categories of harmful content.

a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", Facebook was unable to provide the breakdown of harmful content it removed reactively as a direct result of user reports from Singapore users. Facebook reported that during the reporting period, there were 3,200,000 pieces of content reported by users from Singapore and it took action on 130,100 pieces of user reported content for violating its community guidelines. However, Facebook also reported that it was unable to provide a breakdown of this data by the six harmful content categories, as its focus has been on logging data for violative content that it proactively detected and removed before the content was user reported.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, Facebook was unable to provide this data due to the same reason provided in paragraph 9(a).

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", Facebook provided data on the harmful content it had removed proactively, broken down by the six harmful content categories, both globally and in Singapore.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", Facebook reported that during the reporting period, it disabled over 924,300 user accounts on Facebook created in Singapore for violating its community guidelines (excluding fake accounts). However, it did not provide a breakdown of this data by the six harmful content categories.

10 Based on the Singapore-specific data provided by Facebook and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on Facebook were Sexual content, Violent content and Cyberbullying content.

b. From 2024 to 2025, for several types of harmful content categories, Facebook's data showed significant changes in removals of these types of harmful content created in Singapore.

    i. When asked by IMDA, Facebook explained that a major contributing factor to the general increase in numbers across content categories for Singapore could be the changes in its measurement methodologies, which resulted in a larger pool of Singapore users being captured.

    ii. For its "Child Endangerment: Nudity and Physical Abuse" policy, there was a 21% increase in content removed under this policy, from 10,200 pieces of content removed in 2024 to 12,300 removed in 2025. Facebook explained that this increase was due to changes to its policy and enforcement approach, such that it removed more types of sexual content than it did previously.

    iii. For its "Dangerous Organisations and Individuals (Organised Hate)" policy, there was a 133% increase in content removed under this policy, from 1,500 pieces of content removed in 2024 to 3,500 removed in 2025. Facebook explained that it made significant investments to take action against criminal scam organisations under this policy, and that global events, such as the Israel-Gaza conflict and instability in the Middle East, might have influenced the prevalence of violating content.

iv. As part of Facebook's responsibility to provide a safe experience for its users, Facebook is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks. IMDA will continue to engage Facebook to improve its information transparency to Singapore users.

Harmful Content Categories	Facebook's Community Standards	Number of content actioned that was created in Singapore	Percentage of content that was proactively detected	Remaining percentage of content that was not proactively detected[7]
Sexual content	Adult Nudity and Sexual Activity	215,600	92.9%	7.1%
	Child Endangerment: Nudity and Physical Abuse	12,300	98.5%	1.5%
	Child Endangerment: Sexual Exploitation	43,500	98.6%	1.4%
Violent content	Violent and Graphic Content	89,200	98.8%	1.2%
	Violence and Incitement	48,400	97.5%	2.5%
Suicide and self-harm content	Suicide and Self-Injury	13,900	96.9%	3.1%
Cyberbullying content	Bullying and Harassment	79,300	83.1%	16.9%
	Hateful Conduct	26,300	90.6%	9.4%
Content endangering public health	Regulated Goods (Drugs)	5,800	97.3%	2.7%
Content facilitating vice and organised crime	Regulated Goods (Firearms)	6,300	96.0%	4.0%
	Dangerous Organisations and Individuals (Organized Hate)	3,500	93.4%	6.6%
	Dangerous Organisations and Individuals (Terrorism)	16,900	96.1%	3.9%

Table 1: Data provided by Facebook for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

[7] Facebook reported that the content that was not proactively detected as reflected in this column included content that it took action on after user reports were received, but this did not necessarily indicate that Facebook took action on them only because of those user reports.

Harmful Content Categories	Facebook's Community Standards	Number of content actioned globally	Percentage of content that was proactively detected	Remaining percentage of content that was not proactively detected[8]
Sexual content	Adult Nudity and Sexual Activity	193,900,000	94.5%	5.5%
	Child Endangerment: Nudity and Physical Abuse	6,400,000	98.3%	1.7%
	Child Endangerment: Sexual Exploitation	27,600,000	96.8%	3.2%
Violent content	Violent and Graphic Content	50,500,000	98.7%	1.3%
	Violence and Incitement	22,300,000	97.5%	2.5%
Suicide and self-harm content	Suicide and Self-Injury	25,600,000	99.1%	0.9%
Cyberbullying content	Bullying and Harassment	27,200,000	84.0%	16.0%
	Hateful Conduct	22,900,000	94.5%	5.5%
Content endangering public health	Regulated Goods (Drugs)	8,200,000	96.8%	3.2%
Content facilitating vice and organised crime	Regulated Goods (Firearms)	6,200,000	98.0%	2.0%
	Dangerous Organisations and Individuals (Organized Hate)	4,400,000	96.1%	3.9%
	Dangerous Organisations and Individuals (Terrorism)	40,500,000	99.4%	0.6%

Table 2: Data provided by Facebook for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

11 Facebook's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

Meta's Response

Meta appreciates the continued engagement and collaboration with the Singapore government on online safety. Meta remains committed to keeping people safe on our platforms.

[8] Facebook reported that the content that was not proactively detected as reflected in this column included content that it took action on after user reports were received, but this did not necessarily indicate that Facebook took action on them only because of those user reports.

HardwareZone

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
☑☑☑☑☐	☑☑☑☑☑	☑☑☑☐☐	☑☑☑☑☑	☑☑☑☑☑

Overall Assessment

01 HardwareZone's overall online safety rating improved slightly from 2024. Notably, the effectiveness and timeliness of HardwareZone's user reporting and resolution mechanisms improved from 2024. In addition, HardwareZone provided clear information in its annual online safety report with Singapore-specific data.

02 HardwareZone's measures to restrict children from accessing its service were found to be ineffective. However, HardwareZone has taken steps to improve its measures after engagement by IMDA. IMDA's preliminary assessment is that HardwareZone's improved measures meet the requirements of the Code.

Section Ai: User safety measures for all end-users

03 HardwareZone had the required user safety measures for all users. HardwareZone did not report new measures for this section in its annual online safety report for 2025.

Section Aii: User safety measures for children

04 HardwareZone was initially found to be ineffective in restricting access to its service by children. However, HardwareZone took steps to improve its measures after engagement by IMDA.

a. In 2024, IMDA notified HardwareZone that it needed to either effectively restrict children from accessing its service or put in place comprehensive safety measures for children as required by paragraph 20 of the Code. This was because HardwareZone's Terms of Service prohibit users under the age of 18 from accessing the service, but its age-gating measure to enforce this could easily be bypassed. As a result, children could access age-inappropriate content such as sexually suggestive references or innuendos.

b. HardwareZone's measures to restrict access by children were assessed to be ineffective again in 2025. Similar to 2024, HardwareZone's age-gating measure could still be easily bypassed.

i. There was no option for users to self-declare their age at the account creation stage. The use of Singpass to verify a user's age was also optional, with this option for age verification via Singpass only appearing via a pop-up box for registered users who were logged in. Users who were not logged in were able to access HardwareZone's forum and thus access harmful or age-inappropriate content.

ii. Based on IMDA's tests in 2025, children who bypassed the age-gating measure could still easily access age-inappropriate content, such as posts using sexualised, derogatory, and discriminatory terms.

c. After multiple engagements by IMDA, HardwareZone submitted its plans to implement an updated age verification system and an updated content classification model on its service, which came into effect in end-January 2026.

i. On age verification, HardwareZone explained that its current age query pop-up has been phased out and replaced with an updated age verification system utilising the sgID service that is part of the Singpass app. Under this updated system, users who do not have an account and do not choose to undergo age verification would not be able to view content marked as inappropriate for children.

ii. On its updated content classification system, HardwareZone explained that it was developed and trained by an experienced forum moderation team that factored in local lingo, linguistic jokes that required mastery of particular languages, and an understanding of acceptable social and cultural norms. The classification system would go through each piece of content on the service and classify them into three categories: (i) highly objectionable (which would be automatically blocked from publication and sent to the moderation team for review), (ii) suitable for mature audiences only, or (iii) suitable for all age groups.

05 IMDA's preliminary assessment is that these new measures by HardwareZone to restrict children's access to age-inappropriate content meet the requirements of paragraph 20 of the Code. Age verification via Singpass is now mandatory in order to view content meant for users above 18, and IMDA also found that it is now more difficult for a child (or users who are not logged in) to find and access harmful and age-inappropriate content. IMDA will assess the effectiveness of these measures after HardwareZone submits its next annual online safety report due on 30 June 2026.

Section B: User reporting and resolution

06 HardwareZone continues to have effective user reporting and resolution mechanisms. IMDA's Mystery Shopper tests found that HardwareZone had a higher action rate on user reports of harmful content and a lower average response time to act on these user reports in 2025.

a. HardwareZone took action on 93% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 89% in 2024.

b. HardwareZone's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 2 days in 2025 from an average time of 3 days in 2024.

Section C: Accountability

07 HardwareZone provided clear information in its annual online safety report with Singapore-specific data.

a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", HardwareZone provided the data broken down by the six harmful content categories.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, HardwareZone provided the median time it took to take action on user reports in Singapore broken down by the six harmful content categories.

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", HardwareZone provided the data broken down by the six harmful content categories.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", HardwareZone provided data on the number of accounts suspended in Singapore, broken down by the six harmful content categories.

08 Based on the Singapore-specific data provided by HardwareZone and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on HardwareZone were Sexual content, Cyberbullying content and Violent content.

b. From 2024 to 2025, HardwareZone's data showed a 62% decrease in the number of Violent content and a 79% decrease in the number of Cyberbullying content removed as a result of user reports. HardwareZone explained that this decrease was due to a shift in user behaviours as errant users were banned from its forum over time.

c. As part of HardwareZone's responsibility to provide a safe experience for its users, HardwareZone is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks.

Content Categories	Volume of end-user reports received	Volume of content removed as a result of end-user reports
Sexual content	133	105
Violent content	11	8
Suicide and self-harm content	4	2
Cyberbullying content	25	9
Content endangering public health	4	4
Content facilitating vice and organised crime	1	1

Table 1: Data provided by HardwareZone for paragraph 26(a) of the Code: The number of end-user reports received from end-users in Singapore and the number of harmful and inappropriate content removed as a result of end-user reports by harmful content category

Content Categories	Median/average time taken to act on end-user reports in Singapore	Volume of content removed as a result of end-user reports
Sexual content	3 hrs	105
Violent content	1 hr	8
Suicide and self-harm content	3 hrs	2
Cyberbullying content	7 hrs	9
Content endangering public health	11 hrs	4
Content facilitating vice and organised crime	12 mins	1

Table 2: Data provided by HardwareZone for paragraph 26(b) of the Code: Time between receiving end-user reports from end-users in Singapore on harmful and inappropriate content and taking action

Content Categories	Volume of content proactively removed that are accessible by end-users in Singapore	Volume of content proactively removed that originated from Singapore
Sexual content	47	47
Violent content	0	0
Suicide and self-harm content	1	1
Cyberbullying content	7	7
Content endangering public health	1	1
Content facilitating vice and organised crime	1	1

Table 3: Data provided by HardwareZone for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

Content Categories	Volume of accounts suspended or banned in Singapore
Sexual content	7
Violent content	3
Suicide and self-harm content	1
Cyberbullying content	4
Content endangering public health	1
Content facilitating vice and organised crime	1

Table 4: Data provided by HardwareZone for paragraph 26(d) of the Code: The number of accounts suspended in Singapore

09 HardwareZone's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

HardwareZone's Response

We acknowledge IMDA's assessment of HardwareZone Forum's annual report findings. As of 26th January 2026, we've rolled out an updated age-verification-based access to content viewability, alongside an updated content classification model. HWZ is committed to ensuring the online safety of all users, and we'll continue to monitor this fast-changing internet landscape to consider relevant updates for our platform.

Instagram

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
☑☑☑☑☐	☑☑☑☑☐	☑☑☑☑☑	☑☑☑☐☐	☑☑☑☑☐

Overall Assessment

01 Instagram's overall online safety rating remained the same as 2024. The effectiveness and timeliness of Instagram's user reporting and resolution mechanisms improved from 2024.

02 In 2025, IMDA detected 4 cases of child sexual exploitation and abuse material (CSEM) originating from or targeting Singapore users on Instagram that were not proactively detected and removed until IMDA flagged those cases to Instagram. CSEM is a very serious harm that must be proactively detected and removed by the Designated Social Media Services (DSMSs) expeditiously.

03 Similar to 2024, Instagram still did not provide Singapore-specific data on the effectiveness and timeliness of its user reporting and resolution mechanisms.

Section Ai: User safety measures for all end-users

04 Instagram had the required user safety measures for all users. Measures that were newly reported in Instagram's latest annual online safety report include the following:

a. Instagram worked with the Mental Health Coalition to establish Thrive, a program for participating technology companies to share signals about violating suicide and self-harm content.

b. Instagram partnered with EYEYAH! to conduct two workshops for educators in Singapore. Participants learned new tools and strategies to enrich student learning on digital well-being themes, such as anxiety, addiction and cyber bullying.

05 Instagram reported more measures to proactively detect and remove CSEM. These measures, as reflected in Instagram's annual online safety report, include the use of technology to review more than 60 different signals to identify potentially suspicious adults, and hiring specialists with law enforcement and online child safety backgrounds to find and remove predatory networks on the service.

06 Despite this, IMDA detected 4 cases of CSEM originating from or targeting Singapore users on Instagram in 2025 that were not proactively detected and removed by Instagram. In contrast, IMDA detected 1 case of CSEM that was not proactively detected and removed in 2024.

a. Following IMDA's flagging of the 4 CSEM cases in 2025, Instagram took appropriate action on all of them.

b. IMDA has engaged Instagram regarding its efforts to improve the proactive detection and removal of CSEM. Instagram will need to provide an update on the steps taken to improve its measures to proactively detect and remove CSEM in its next annual online safety report.

c. Instagram must continue to be vigilant for CSEM cases on its service and the evolving modus operandi regarding the sale or sharing of sexual or sexualised imagery of individuals under 18, whether self-generated or produced by a third party.

07 Instagram had the required user safety measures for children. Measures that were newly reported in Instagram's latest annual online safety report include the following:

a. Instagram introduced Teen Accounts, which automatically places under 18 users into more restrictive settings, such as the strictest messaging settings where teens can receive messages only from people they follow, and sensitive content restrictions. For under 16 users, parental permission is required to change any default protections to less strict settings.

Section B: User reporting and resolution

08 The effectiveness and timeliness of Instagram's user reporting and resolution mechanisms showed improvement from 2024. IMDA's Mystery Shopper tests found that Instagram had a higher action rate on user reports of harmful content and a lower average response time to act on these user reports in 2025. Regardless, Instagram should continue to improve the performance of its user reporting and resolution mechanisms.

a. Instagram took action on 54% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 2% in 2024.

    i. Notably, as compared to its average action rate, Instagram took action on only 20% of Cyberbullying content and Content endangering public health that was user reported. It failed to act on the remaining 80% until IMDA notified Instagram to review them again.

b. Instagram's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 4 days in 2025, from an average time of 7 days in 2024.

    i. Notably, Instagram took a higher-than-average time of 5.7 days to remove Sexual content that violated its community guidelines. Instagram should improve the timeliness of removing such content when user reported.

Section C: Accountability

09 Instagram should improve the provision of data in its annual online safety report. In particular, Instagram was unable to provide Singapore-specific data to demonstrate the effectiveness and timeliness of its user reporting and resolution mechanisms for specific categories of harmful content.

a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", Instagram was unable to provide the breakdown of harmful content it removed reactively as a direct result of user reports from Singapore users. Instagram reported that during the reporting period, there were 1,200,000 pieces of content reported by users from Singapore and it took action on 42,900 pieces of user reported content for violating its community guidelines. However, Instagram also reported that it was unable to provide a breakdown of this data by the six harmful content categories, as its focus has been on logging data for violative content that is proactively detected and removed before the content was user reported.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, Instagram was not able to provide this data due to the same reason provided in paragraph 9(a).

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", Instagram provided data on the harmful content it had removed proactively, broken down by the six harmful content categories, both globally and in Singapore.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", Instagram reported that during the reporting period, it disabled over 1,300,000 user accounts on Instagram created in Singapore for violating its community guidelines (excluding fake accounts). However, it did not provide a breakdown of this data by the six harmful content categories.

10 Based on the Singapore-specific data provided by Instagram and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on Instagram were Violent content, Cyberbullying content and Sexual content.

b. From 2024 to 2025, for several types of harmful content categories, Instagram's data showed significant changes in removals of these types of harmful content created in Singapore.

   i. When asked by IMDA, Instagram explained that a major contributing factor to the general increase in numbers across content categories for Singapore could be the changes in its measurement methodologies, which resulted in a larger pool of Singapore users being captured.

   ii. For its "Child Endangerment" policy, there was a 212% increase in content removed under this policy, from 4,712 pieces of content removed in 2024 to 14,700 removed in 2025. Instagram explained that this increase was due to changes to its policies and enforcement approach, such that it removed more types of sexual content than it did previously.

   iii. For its "Bullying and Harassment" policy, there was a 76% increase in content removed under this policy, from 38,500 pieces of content removed in 2024 to 67,800 removed in 2025. Instagram explained that this increase was due to changes in its policy, such that content shared by an unwanted contact (when confirmed by the recipient) and videos of physical bullying against minors shared in any context were removed.

iv. For its "Suicide and Self-Injury" policy, there was a 67% increase in content removed under this policy, from 10,500 pieces of content removed in 2024 to 17,500 removed in 2025. Instagram explained that this increase was due to changes in its policy, such that content depicting a person who attempted or died by suicide, as well as imagery depicting body modification when shared in the context of suicide or self-injury, were prohibited.

v. As part of Instagram's responsibility to provide a safe experience for its users, Instagram is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks. IMDA will continue to engage Instagram to improve its information transparency to Singapore users.

Harmful Content Categories	Instagram's Community Standards	Number of content actioned that was created in Singapore	Percentage of content that was proactively detected	Remaining percentage of content that was not proactively detected[9]
Sexual content	Adult Nudity and Sexual Activity	54,300	97.9%	2.1%
	Child Endangerment: Nudity and Physical Abuse	2,100	97.3%	2.7%
	Child Endangerment: Sexual Exploitation	12,600	96.5%	3.5%
Violent content	Violent and Graphic Content	129,500	99.4%	0.6%
	Violence and Incitement	122,300	99.6%	0.4%
Suicide and self-harm content	Suicide and Self-Injury	17,500	94.8%	5.2%
Cyberbullying content	Bullying and Harassment	67,800	94.0%	6.0%
	Hateful Conduct	116,300	98.9%	1.1%
Content endangering public health	Regulated Goods (Drugs)	808	87.7%	12.3%
Content facilitating vice and organised crime	Regulated Goods (Firearms)	1,200	99.3%	0.7%
	Dangerous Organisations and Individuals (Organized Hate)	2,300	93.5%	6.5%
	Dangerous Organisations and Individuals (Terrorism)	6,400	96.2%	3.8%

Table 1: Data provided by Instagram for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

[9] Instagram reported that the content that was not proactively detected as reflected in this column included content that it took action on after user reports were received, but this did not necessarily indicate that Instagram took action on them only because of those user reports.

Harmful Content Categories	Instagram's Community Standards	Number of content actioned globally	Percentage of content that was proactively detected	Remaining percentage of content that was not proactively detected[10]
Sexual content	Adult Nudity and Sexual Activity	43,300,000	94.5%	5.5%
	Child Endangerment: Nudity and Physical Abuse	2,800,000	98.8%	1.2%
	Child Endangerment: Sexual Exploitation	12,000,000	97.5%	2.5%
Violent content	Violent and Graphic Content	31,900,000	98.6%	1.4%
	Violence and Incitement	31,700,000	99.1%	0.9%
Suicide and self-harm content	Suicide and Self-Injury	30,800,000	99.2%	0.8%
Cyberbullying content	Bullying and Harassment	31,200,000	94.8%	5.2%
	Hateful Conduct	29,200,000	98.2%	1.8%
Content endangering public health	Regulated Goods (Drugs)	4,000,000	95.6%	4.4%
Content facilitating vice and organised crime	Regulated Goods (Firearms)	811,100	99.2%	0.8%
	Dangerous Organisations and Individuals (Organized Hate)	937,800	90.3%	9.7%
	Dangerous Organisations and Individuals (Terrorism)	14,100,000	98.9%	1.1%

Table 2: Data provided by Instagram for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

11 Instagram's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

Meta's Response

Meta appreciates the continued engagement and collaboration with the Singapore government on online safety. Meta remains committed to keeping people safe on our platforms.

[10] Instagram reported that the content that was not proactively detected as reflected in this column included content that it took action on after user reports were received, but this did not necessarily indicate that Instagram took action on them only because of those user reports.

TikTok

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
☑☑☑◪☐☐	☑☑☑☑☐	☑☑☑☑☑	☑☑☐☐☐	☑☑☑☑☑

Overall Assessment

01 TikTok's overall online safety rating declined from 2024. Notably, the effectiveness of TikTok's user reporting and resolution mechanisms declined significantly in 2025 while its average time to action improved slightly.

02 In 2025, IMDA detected terrorism content shared by Singapore-based accounts on TikTok that were not proactively removed, and also not removed upon user reporting. Terrorism content is very serious harm that must be proactively detected and removed by Designated Social Media Services (DSMSs) before users encounter them.

Section Ai: User safety measures for all end-users

03 TikTok's measures to proactively detect and remove terrorism content were assessed to have serious weaknesses. IMDA detected 17 cases of terrorism content shared by Singapore-based accounts on TikTok in 2025 that were not proactively removed, and also not removed upon user reporting. There were no cases of terrorism content detected by or reported to IMDA in 2024.

a. The 17 cases of terrorism content detected in 2025 featured well-established indicators that the content was terrorism-related, in particular the use of edited footage or audio related to known transnational terrorist organisations that were in some cases blended with benign content. In some cases, the terrorism-related audio content was also concealed under TikTok's "original sound" label, which adds the audio to TikTok's database for others to use in their posts as well. IMDA has shared with TikTok our analysis of the content and indicators associated with the terrorism content.

b. Furthermore, when some of these 17 cases of terrorism content were user reported via its in-app user reporting mechanism, TikTok responded in 30 minutes stating that the content did not violate its community guidelines. This demonstrated that TikTok's automated content moderation systems did not accurately assess the terrorism content when user reported. TikTok only removed the terrorism content after IMDA's flagging. While these cases should have been proactively detected and removed by TikTok in the first place, as required by paragraph 15 of the Code, it also demonstrates the weakness of TikTok's user reporting and resolution mechanisms.

04 This is despite TikTok's measures reported in its annual online safety report to proactively detect and remove terrorism content on its service.

a. These measures included computer vision models to detect signals, emblems, logos, and objects known to be associated with terrorism groups; text-based technologies to detect language used to promote extremist ideologies and terrorist groups; as well as blocking searches for names and organisations associated with terrorism to disrupt the discoverability of terrorism content.

05 IMDA has issued a Letter of Caution to TikTok regarding its measures to proactively detect and remove terrorism content. TikTok has accepted IMDA's findings and committed to put in place specific measures to rectify these issues. TikTok has also been placed under Enhanced Supervision, in which it must meet with IMDA regularly to account for its progress in implementing the rectification measures it has committed to, until IMDA is satisfied that the issues are adequately resolved. In addition, TikTok will need to provide supporting data and information to IMDA, in its next annual online safety report due on 30 June 2026, to demonstrate the effectiveness of its implementation of the rectification measures.

06 Should TikTok fail to satisfy IMDA that it has improved its measures to address the specific types of terrorism content that IMDA has detected, IMDA will not hesitate to explore further options, including potential regulatory action under the Broadcasting Act.

Section Aii: User safety measures for children

07 TikTok had the required user safety measures for children. Measures that were newly reported in TikTok's latest annual online safety report include the following:

a. TikTok's Family Pairing feature allows parents or guardians to customise safety and privacy settings for their children, such as enabling "Restricted Mode" to limit their children's exposure to content with mature or harmful themes in their "For You" feed.

b. Children between 13 to 15 will be required to review the "who can watch this video" setting when they make their first post.

Section B: User reporting and resolution

08 The effectiveness of TikTok's user reporting and resolution mechanisms declined significantly from 2024. IMDA's Mystery Shopper tests in 2025 found that while TikTok's average response time to act on user reports improved, TikTok had a lower action rate on user reports of harmful content.

a. TikTok took action on 25% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 39% in 2024.

i. Notably, as compared to its average action rate, TikTok took action on only 4% of Content endangering public health that was user reported and failed to act on the remaining 96% until IMDA notified TikTok to review them again.

ii. TikTok was the only DSMS to decline in the effectiveness of its user reporting and resolution mechanisms. TikTok needs to demonstrate significant improvement in its response to user reports and provide an update to IMDA on the steps it has taken to do so in its next annual online safety report. IMDA will monitor TikTok's performance in this respect closely.

b. TikTok's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 4 days in 2025 from an average time of 5 days in 2024.

    i. Notably, TikTok took a higher-than-average time of 4.8 days to remove Cyberbullying content that violated its community guidelines. TikTok should improve the timeliness of removing such content when user reported as such content can result in direct harm to others.

Section C: Accountability

09 TikTok provided clear information in its annual online safety report with Singapore-specific data.

a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", TikTok reported that it evaluated 954,289 reports submitted by users in Singapore, of which 36,602 reports were for videos that were found to be violative of its community guidelines. TikTok also reported that 20,493 videos originating from Singapore were removed as a result of user reports globally and provided the data broken down by its own policies.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, TikTok reported that the median time it took to remove videos reported by users in Singapore was 22.5 hours.

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", TikTok provided the data broken down by the six harmful content categories.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", TikTok provided data on the number of accounts suspended globally and in Singapore.

10 Based on the Singapore-specific data provided by TikTok and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on TikTok were content that violated its community guidelines on "Sensitive and Mature Themes", "Regulated Goods and Commercial Activities" and "Mental and Behavioural Health" respectively.

b. From 2024 to 2025, TikTok's data showed a decrease in the number of harmful content originating from Singapore that was proactively removed for six out of its seven reported harmful content categories.

    i. TikTok explained that this decrease from 2024 to 2025 could be attributed to several factors that occurred during the period from 2023 to 2024. For example, global events, in particular the Israel-Hamas conflict, resulted in higher enforcement activity in that period. TikTok also reported a substantial increase in the number of inauthentic and fake accounts and related videos removed in Q4 2023 and Q1 2024.

    ii. The only content category that had an increase in the number of harmful content proactively removed was content that violated TikTok's "Youth Safety and Well-Being" policy. There was a 138% increase in violative content proactively removed under this policy, from 36,343 pieces removed in 2024 to 86,247 pieces removed in 2025. TikTok explained that this increase could be due to adjustments in its enforcement strategy, such that it removed more content related to alcohol, tobacco and drugs under this policy.

    iii. As part of TikTok's responsibility to provide a safe experience for its users, TikTok is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks.

TikTok's Community Guidelines	Number of Singapore-originated content removed due to user reports
Integrity and Authenticity	95
Mental and Behavioural Health	858
Privacy and Security	8,022
Regulated Goods and Commercial Activities	5,460
Safety and Civility	3,205
Sensitive and Mature Themes	5,937
Youth Safety and Well-Being	531

Table 1: Data provided by TikTok for paragraph 26(a) of the Code: The number of end-user reports received from end-users in Singapore and the number of harmful and inappropriate content removed as a result of end-user reports by harmful content category

TikTok's Community Guidelines	Number of content proactively removed at the global level	Number of Singapore-originated content proactively removed
Integrity and Authenticity	16,337,977	20,740
Mental and Behavioural Health	168,879,921	120,850
Privacy and Security	73,905,372	55,470
Regulated Goods and Commercial Activities	236,548,930	271,742
Safety and Civility	132,137,690	72,119
Sensitive and Mature Themes	259,418,915	284,259
Youth Safety and Well-Being	168,368,582	86,427

Table 2: Data provided by TikTok for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

Number of user accounts removed for violating TikTok's Community Guidelines	115,530,334 (Of which, 86,525,704 accounts were removed on the basis that users were suspected to be under the age of 13.)
Number of Singapore-originated user accounts removed for violating TikTok's Community Guidelines	69,101 (Of which, 37,610 accounts were removed on the basis that users were suspected to be under the age of 13.)

Table 3: Data provided by TikTok for paragraph 26(d) of the Code: The number of accounts suspended globally and in Singapore

11 TikTok's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

TikTok's Response

TikTok is committed to ensuring the safety and integrity of the platform, and this is a responsibility with no finish line. We continue to refine and improve our approach to remove violative content quickly, accurately, and at scale. We are committed to strengthening our systems based on IMDA's feedback as we continue the work to support a safe online environment for all.

On User Safety (Section Ai), TikTok's goal is to identify and remove harmful content before it reaches our users. TikTok does not allow violent and hateful organisations or individuals on our platform. In Q4 2025, our proactive systems successfully removed over 99% of content violating our Violent and Hateful Organisation and Individual policy before it was reported, with over 93% of content being removed within 24 hours.

As violent extremist methodologies and evasion techniques continue to evolve, so do we. We continually evaluate and strengthen our policies and systems, and are committed to partnering with IMDA under the Enhanced Supervision scheme. We view this as a constructive path forward and remain focused on strengthening our systems to ensure TikTok remains a safe and positive space for our community in Singapore.

On User Safety measures for children (Section Aii), we are pleased that IMDA recognised our longstanding youth safety measures, including our Family Pairing features and privacy safeguards for younger users. These remain a cornerstone of our commitment to supporting Singaporean families online, and we continue to improve on these tools with new features, like those introduced in July 2025.

On User Reporting (Section B), user reports play an important role within TikTok's safety ecosystem. We are encouraged that IMDA noted improvements in our response times, with our internal data showing a median removal time of 22.5 hours for violative videos reported by Singapore users. We recognise that the effectiveness of user reporting continues to be a challenge faced across the industry. While we continue to refine our user reporting mechanism, we want to highlight the following context on how these reports are processed at scale:

- *While we value the insights from IMDA's "Mystery Shopper" exercise, these findings represent a small snapshot of the larger content pool. During the reporting period, TikTok evaluated nearly one million reports from Singapore users - a volume that we believe more fully reflects the actual effectiveness of our reporting systems.*
- *Our Community Guidelines explain that we may take a range of enforcement actions depending on the type of rule violation. They also include For You eligibility standards that help ensure content that is recommended to people's For You feeds is suitable for a broad audience. This approach embodies our Community Principles of balancing freedom of expression with preventing harm.*
- *We continue to improve our content moderation strategies, including regular training for our teams, so we can better assess contextual information.*

X

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
☑☑☑☐☐	☑☑☐☐☐	☑☑☑☑☐	☑☑☑☑☐	☑☑☑☑☑

Overall Assessment

01 X's overall online safety rating improved slightly from 2024. Notably, X improved the enforcement of its policies to restrict children's accounts from viewing adult sexual content. The effectiveness and timeliness of X's user reporting and resolution mechanisms also improved from 2024. In addition, X provided clear information in its annual online safety report with Singapore-specific data.

02 However, X has not improved its proactive detection and removal of child sexual exploitation and abuse material (CSEM) in 2025, despite its weak performance in this area highlighted in the 2024 report. IMDA detected a 120% increase in CSEM cases on X originating from or targeting Singapore users from 2024 to 2025. CSEM is a very serious harm that must be proactively detected and removed by Designated Social Media Services (DSMSs) before users encounter them.

Section Ai: User safety measures for all end-users

03 X's measures to proactively detect and remove CSEM were assessed to have serious weaknesses. This is the second consecutive year that IMDA has detected CSEM on X despite informing X of this issue in 2024.

a. In 2024, IMDA detected 33 cases of CSEM on X originating from or targeting Singapore users that were not proactively detected and removed by X. These cases involved content sharing or linking to CSEM, and self-generated CSEM. These cases had a Singapore nexus and featured well-established indicators that are commonly associated with CSEM, such as coded words and number patterns indicating that a user is under 18 years of age.

b. In 2025, IMDA detected 73 cases of CSEM with the same characteristics as those detected in 2024 that were not proactively detected and removed, using the same detection methodology.

   i. The cases of content sharing or linking to CSEM (68% of the 2025 cases) involved coordinated networks of accounts using terms commonly associated with CSEM, with links that direct Singapore users to external sites hosting CSEM.

   ii. The cases of self-generated CSEM (32% of the 2025 cases) involved accounts sharing posts with self-generated explicit sexual imagery from users in Singapore who were purportedly under 18. These users included commonly known number patterns on their profile indicating their date of birth or age.

iii. A majority of these 73 cases of CSEM were also on the platform for an extended period of time, ranging from 9 to 31 weeks on average. X therefore had ample time to proactively detect and remove these cases.

c. All 73 cases violated X's own policies against CSEM and X only removed all of them upon IMDA's flagging. These cases should have been proactively detected and removed by X in the first place as required by paragraph 15 of the Code.

d. In both 2024 and 2025, IMDA shared with X its analysis of the CSEM cases and their indicators. Despite this, the measures X reported to address CSEM in its latest annual online safety report did not explain how it addressed the specific types of CSEM that IMDA flagged to X.

i. For example, X's annual online safety report stated that it was training additional agents for proactive keyword and media sweeps, using signals from Project Lantern[11] for account investigations, and launching an improved Direct Message reporting flow with intuitive child safety pathways to enhance reactive enforcement and proactive investigations via clearer signals.

04 IMDA has issued a Letter of Caution to X regarding its measures to proactively detect and remove CSEM. X has accepted IMDA's findings and committed to put in place specific measures to rectify these issues. X has also been placed under Enhanced Supervision, in which it must meet with IMDA regularly to account for its progress in implementing the rectification measures it has committed to, until IMDA is satisfied that the issues are adequately resolved. In addition, X will need to provide supporting data and information to IMDA, in its next annual online safety report due on 30 June 2026, to demonstrate the effectiveness of its implementation of the rectification measures.

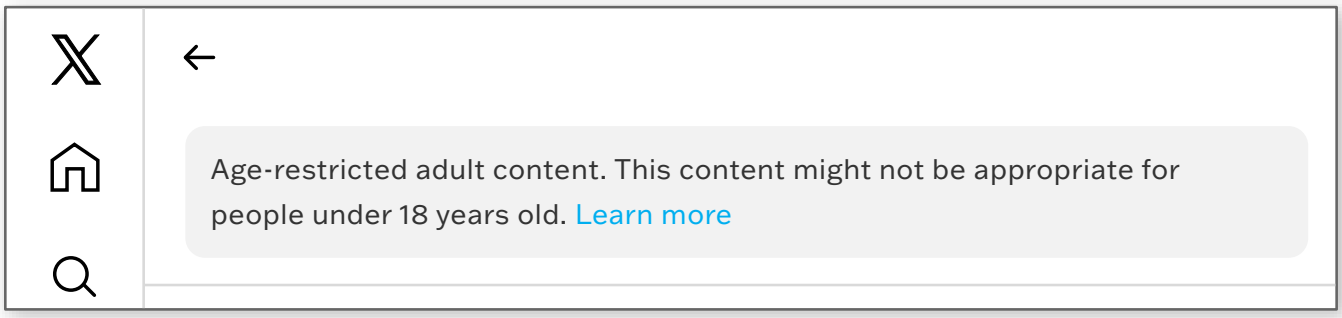
05 Should X fail to satisfy IMDA that it has improved its measures to address the specific types of CSEM that IMDA has detected, IMDA will not hesitate to explore further options, including potential regulatory action under the Broadcasting Act.

Section Aii: User safety measures for children

06 X has improved the enforcement of its policies to restrict children's accounts from viewing adult sexual content. In its annual online safety report, X reported that it takes steps to proactively detect and label sensitive media that has not been labelled in accordance with the X Rules, and that X has implemented the use of additional proactive heuristics to detect and label adult content and violent content.

07 Based on IMDA's tests in 2025, it was more difficult to find and access explicit adult sexual content on X using children's accounts compared to 2024. Most explicit adult sexual content was appropriately age-restricted by X (see below image for screenshot of X's message for age-restricted adult content).

[11] Project Lantern is an initiative that aims to bring together technology companies to securely and responsibly share intelligence and threat indicators related to CSEM. It was launched in 2023 by the Tech Coalition, a community of technology companies, including Google, Meta and Microsoft.



08 Nevertheless, as some explicit adult sexual content could still be accessed using children's accounts, X should continue to improve its safety measures for children.

Section B: User reporting and resolution

09 The effectiveness and timeliness of X's user reporting and resolution mechanisms showed improvement from 2024. IMDA's Mystery Shopper tests found that X had a higher action rate on user reports of harmful content and a lower average response time to act on these user reports in 2025. However, X should continue to improve the performance of its user reporting and resolution mechanisms.

a. X took action on 74% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 54% in 2024.

    i. Notably, as compared to its average action rate, X took action on only 27% of Content endangering public health that was user reported, and failed to act on the remaining 73% until IMDA notified X to review them again.

b. X's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 5 days in 2025 from an average time of 10 days in 2024.

    i. Notably, X took a higher-than-average time of 9 days to remove Sexual content. X should improve the timeliness of removing such content when user reported.

Section C: Accountability

10 a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", X provided the data broken down by the six harmful content categories.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, X reported that the median time it took to take action on user reports in Singapore for all relevant X Rules policies was 69 hours. For X's CSE policy specifically, the median time it took was 4.7 hours.

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", X provided the data broken down by the six harmful content categories.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", X provided data on the number of accounts suspended globally and in Singapore, broken down by the six harmful content categories.

11 Based on the Singapore-specific data provided by X and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on X were Violent content, Sexual content and Cyberbullying content.

b. There was a very low number of content removed as a result of user reports for several types of harmful content. In particular, for X's "Violent Content" and "Hateful Conduct" policies, IMDA noted that there was a significant difference between the number of user reports X received from Singapore users (>250,000 user reports) and the number of content X removed as a result of these user reports (<100 pieces of content removed).

    i. For the "Violent Content" policy, X explained that higher user report volumes did not necessarily result in higher action rates, as enforcement depended on the accuracy of the reports received. X would only take action on content that was user reported if it determined the content to be violative of its policies.

c. There were large year-on-year changes in the number of accounts originating from Singapore that were suspended for some types of harmful content. For example, X's data showed that the number of accounts suspended for violating its "Abuse and Harassment" policy increased from 1,500 accounts in 2024 to 2,600 accounts in 2025, reflecting a 73% year-on-year increase.

    i. X explained that this change might have resulted from a higher volume of violative activity occurring on the platform, a higher volume of accurate user reports received, or improvements to X's proactive detection and enforcement systems.

d. From 2024 to 2025, X's data also showed significant increases in the number of user reports received from Singapore users for several types of harmful content categories.

    i. For its "Violent and Hateful Entities" policy, there was a 70% increase in user reports from Singapore users, i.e. from 25,900 user reports in 2024 to 44,100 in 2025. X explained that the changes in user reporting behaviour might have been due to various factors, such as increased awareness of reporting mechanisms or off-platform events that triggered an increased proclivity among users to report content. Alongside this increase in user reports from Singapore users, there was a decrease in the number of accounts accessible by Singapore users that X removed under this policy, i.e. from 275,400 in 2024 to 101,800 in 2025. X noted that it primarily enforced the "Violent and Hateful Entities" policy at the account level by permanently suspending violative accounts, instead of removing individual pieces of content as a result of user reports.

    ii. For its "Violent Content" policy, there was also a 79% increase in user reports from Singapore users, i.e. from 65,800 user reports in 2024 to 118,008 in 2025 (excluding reports from spam accounts). As with the case for the "Violent and Hateful Entities" policy, X explained that the changes in user reporting behaviour might have been due to various factors, such as increased awareness of reporting mechanisms or off-platform events that triggered an increased proclivity among users to report content. Alongside the increase in user reports from Singapore users under this policy, there was no content removed in response to these user reports in 2025, as compared to the 2,700 pieces of content removed due to user reports from Singapore users in 2024. X explained that the reported content would only be actioned if it was determined that they were violative of its "Violent Content" policy.

e. As part of X's responsibility to provide a safe experience for its users, X is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks.

Harmful Content Categories	X Rules	Number of user reports received from end-users in Singapore	Number of harmful and inappropriate content removed as a result of end-user reports
Sexual content	Non-consensual Nudity	133	5,300 (X reported that the volume of content removed following user reports under the "Non-consensual Nudity" policy was higher than the number of user reports received under the same policy because the data included content user reported under other policies which X ultimately took action on under the "Non-consensual Nudity" policy.)
	Sensitive Media	7,600	1,800
	CSE	3,100,000 (X reported that 95.73% of the user reports received, i.e. approximately 2,967,630 user reports, were from accounts subsequently suspended for spam and other violations.)	26 (X primarily enforces this policy at the account level by permanently suspending violative accounts, and has suspended 17,800 Singapore accounts in the reporting period.)
Violent content	Violent and Hateful Entities	44,100	0 (X primarily enforces this policy at the account level by permanently suspending violative accounts, and has suspended 257 Singapore accounts in the reporting period.)
	Violent Content	297,700 (X reported that 60.63% of the user reports received, i.e. approximately 180,496 user reports, were from accounts subsequently suspended for spam and other violations.)	0
Suicide and self-harm content	Suicide and Self Harm	9,900	514
Cyberbullying content	Abuse and Harassment	558,300 (X reported that 30.72% of the user reports received, i.e. approximately 171,510 user reports, were from accounts subsequently suspended for spam and other violations.)	4,200
	Hateful Conduct	540,600 (X reported that 39.76% of the user reports received, i.e. approximately 214,943 user reports, were from accounts subsequently suspended for spam and other violations.)	79
	Private Content	74,500	316
Content endangering public health and Content facilitating vice and organised crime	Illegal and Regulated Behaviours	117	0
	Financial Scam	0	0
	Synthetic and Manipulated Media	N/A (X reported that users cannot submit reports under its Synthetic and Manipulated Media policy, and that it proactively detects violative content under this policy.)	

Table 1: Data provided by X for paragraph 26(a) of the Code: The number of end-user reports received from end-users in Singapore and the number of harmful and inappropriate content removed as a result of end-user reports by harmful content category

Harmful Content Categories	X Rules	Number of content proactively removed that are accessible by end-users in Singapore	Number of content proactively removed that originated from Singapore
Sexual content	Non-consensual Nudity	23,000	59
	Sensitive Media	769,100	5,200
	CSE	2,500	8
Violent content	Violent and Hateful Entities	0	0
	Violent Content	1,400,000	2,400
Suicide and self-harm content	Suicide and Self Harm	131	0
Cyberbullying content	Abuse and Harassment	2,300	1
	Hateful Conduct	2,200	5
	Private Content	3,700	2
Content endangering public health and Content facilitating vice and organised crime	Illegal and Regulated Behaviours	143	2
	Financial Scam	0	0
	Synthetic and Manipulated Media	0	0

Table 2: Data provided by X for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

Harmful Content Categories	X Rules	Number of accounts suspended globally and accessible by end-users in Singapore	Number of accounts suspended that originated from Singapore
Sexual content	Non-consensual Nudity	143,000	837
	Sensitive Media	31,800	44
	CSE	4,900,000	17,800
Violent content	Violent and Hateful Entities	101,800	257
	Violent Content	157,000	419
Suicide and self-harm content	Suicide and Self Harm	3,300	11
Cyberbullying content	Abuse and Harassment	1,980,000	2,600
	Hateful Conduct	4,700	3
	Private Content	4,600	18
Content endangering public health and Content facilitating vice and organised crime	Illegal and Regulated Behaviours	1,060,000	1,500
	Financial Scam	22,600	42
	Synthetic and Manipulated Media	3	0

Table 3: Data provided by X for paragraph 26(d) of the Code: The number of accounts suspended globally and in Singapore by harmful content category

12 X's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

X's Response

X applies rigorous global child safety policies, proactive detection tools and enforcement to protect Singapore users from child sexual exploitation and abuse material.

Our mission is to promote and protect the public conversation, and ensuring the safety of our users is our top priority. X users have the right to express their opinions and ideas without fear of censorship, while we uphold our responsibility to protect them from content that violates our Rules.

X maintains a zero tolerance policy toward child sexual exploitation material ("CSEM") and eliminating it is one of our key goals. During the reporting period, we suspended approximately 4.9 million accounts globally for CSEM violations, including 17.8k Singapore accounts. For user reports of CSEM, the median resolution time was 4.7 hours.

We enforce our Rules through a combination of advanced machine learning and human review, supported by an international team providing 24/7 coverage across multiple languages, and a robust appeals process. These enforcement actions are made possible by our robust safeguards and enforcement mechanisms, powered by state-of-the-art technology and proactive detection tools.

Although X is not primarily a platform for children, we are deeply committed to child safety and have measures in place to ensure that minors' experience on X is safe and secure. We have made meaningful progress in restricting minors' access to adult content on X, and warmly welcome IMDA's recognition of this. We take our responsibility to young users very seriously and continue to strengthen safeguards in this area.

We remain deeply committed to continuously strengthening our CSEM detection and enforcement systems and building new defences that proactively reduce and eliminate such content. We acknowledge the IMDA's identification of 73 CSEM cases. We take these incidents extremely seriously and all cases were promptly actioned upon notification. In context, this number remains very small relative to the enormous volume of content shared daily on X, and the millions of accounts suspended globally, and in Singapore, for CSEM violations. We implemented targeted rectification measures for the issues identified by IMDA regarding Singapore users, reflecting our ongoing investment in advanced technology to minimise such occurrences.

As online behaviours evolve, we are also strengthening the effectiveness and timeliness of our user reporting and resolution processes. These efforts have delivered significant results, increasing our action rate from 54% to 74% and reducing our average response time from 10 days to 5 days. This progress has been recognised by IMDA in their findings.

We remain committed to enhancements in these areas, whilst upholding the human rights principles that underpin our policies and enforcement.

X remains committed to close collaboration with IMDA, in the spirit of transparency and open dialogue, to enhance online safety in Singapore while upholding our core principles.

YouTube

Overall Rating	Ratings for Individual Sections of the Online Safety Code			
	Section Ai: User safety measures for all end-users	Section Aii: User safety measures for children	Section B: User reporting and resolution	Section C: Accountability
				

Overall Assessment

01 YouTube's overall online safety rating improved slightly from 2024. Notably, the effectiveness and timeliness of YouTube's user reporting and resolution mechanisms improved from 2024.

02 However, YouTube needs to improve on the enforcement of its community guidelines for children. Similar to 2024, YouTube had instances where children's accounts could access harmful and age-inappropriate content that should have been restricted under its community guidelines.

03 Furthermore, YouTube still did not provide Singapore-specific data on the effectiveness and timeliness of its user reporting and resolution mechanisms.

Section Ai: User safety measures for all end-users

04 YouTube had the required user safety measures for all users. YouTube also reported a new initiative to educate and raise awareness of online safety in its latest annual online safety report:

a. YouTube shared that it would join regional content creators and distributors to participate in the Youth Digital Wellbeing Initiative and support its vision for the development of high-quality age-appropriate content for young people.

Section Aii: User safety measures for children

05 YouTube had the required user safety measures for children. Measures that were newly reported in YouTube's latest annual online safety report include the following:

a. YouTube reported that it strengthened the enforcement of its policies on "Violent or graphic content" and "Illegal or regulated goods or services" by age-restricting additional types of content in these categories, such as fictional violence with graphic scenes and certain types of online gambling content like online casino promotions.

b. YouTube introduced a new Family Center hub where parents can see shared insights into their supervised teens' channel activity, including the number of uploads, subscriptions, and comments.

06 However, YouTube needs to improve the enforcement of its community guidelines for children. Children could still access harmful and age-inappropriate content on YouTube, despite the safety measures it has reported.

a. Similar to 2024, IMDA detected a few instances on YouTube where children's accounts could access harmful and age-inappropriate content that should have been restricted under its community guidelines. These included videos with partial nudity and audio depictions of sexual acts. Following IMDA's flagging of this violative content, YouTube subsequently took the appropriate action on all of them.

b. IMDA has engaged YouTube on the need to improve the enforcement of its community guidelines for children. YouTube will need to provide an update on this in its next annual online safety report.

Section B: User reporting and resolution

07 The effectiveness and timeliness of YouTube's user reporting and resolution mechanisms showed improvement from 2024. IMDA's Mystery Shopper tests found that YouTube had a higher action rate on user reports of harmful content and a lower average response time to act on these user reports in 2025. However, YouTube should continue to improve the performance of its user reporting and resolution mechanisms.

a. YouTube took action on 68% of harmful content, on average, across all categories of harmful content that violated its own community guidelines when reported by user accounts in 2025, as compared to an action rate of 46% in 2024.

    i. Notably, as compared to its average action rate, YouTube took action on only 33% of Content endangering public health that was user reported and failed to act on the remaining 67% until IMDA notified YouTube to review them again.

b. YouTube's average response time to act on harmful content that violated its own community guidelines when reported by user accounts improved to 4 days in 2025 from an average time of 5 days in 2024.

    i. Notably, YouTube took a higher-than-average time to remove Sexual content and Content facilitating vice and organised crime that violated its community guidelines, at 5.1 days and 4.8 days respectively. YouTube should improve the timeliness of removing such content when user reported as such content can result in direct harm to others.

Section C: Accountability

08 YouTube should improve the provision of data in its annual online safety report. In particular, YouTube was unable to provide Singapore-specific data to demonstrate the effectiveness and timeliness of its user reporting and resolution mechanisms for specific categories of harmful content.

a. For paragraph 26(a) of the Code on "the number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports", YouTube was unable to provide the breakdown of harmful content it removed reactively as a direct result of user reports from Singapore users. YouTube reported during the reporting period, there were 445,998 video flags reported by users with a Singapore IP address. YouTube also reported that it removed a total of 72,126 videos that were uploaded from a Singapore IP address for violating its community guidelines, but it did not indicate how many of these were removed due to user reports. In addition, YouTube reported that none of the complaints received from its "Other Legal Complaint" webform were related to the Code or had sufficient basis for removal.

b. For paragraph 26(b) of the Code on the time it took to take action on user reports, YouTube was not able to provide this data.

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service", YouTube provided the number of videos it proactively removed both globally and in relation to videos uploaded from a Singapore IP address. However, it did not provide a breakdown of this data by the six harmful content categories.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", YouTube provided data on the number of global and Singapore channels that were removed.

09 Based on the Singapore-specific data provided by YouTube and IMDA's subsequent clarifications, IMDA observed the following:

a. The top three types of harmful content prevalent on YouTube were content that violated its community guidelines on "Harmful or Dangerous", "Child Safety" and "Nudity or Sexual" respectively.

b. From 2024 to 2025, for several types of harmful content categories, YouTube's data showed significant changes in removals of these harmful content types uploaded from a Singapore IP address. YouTube explained these changes when asked by IMDA.

    i. For its "Child Safety" policy, there was a 59% increase in the monthly average of videos removed under this policy, from an average of 769 videos removed per month in 2024 (6,917 videos over 9 months) to 1,220 videos removed per month in 2025 (14,644 videos over 12 months). YouTube explained that this increase aligned with global trends and was a result of improved proactive detection and not increased local prevalence of such content on its platform. YouTube also explained that the reported numbers may fluctuate from one reporting period to another for various reasons, including platform-level changes or enhancements, changes in the number of users on the platform, external events, and differences in reporting periods.

    ii. For its "Harmful or Dangerous" policy, there was a 98% decrease in videos removed under this policy, from 1,956,180 videos removed in 2024 to 42,104 removed in 2025. YouTube explained that there was a one-off spike in detection of violative videos under this policy in 2024.

iii. As part of YouTube's responsibility to provide a safe experience for its users, YouTube is expected to keep track of significant changes in its online safety data, account for the effectiveness of its online safety measures, and adapt its measures to deal with any emerging risks. IMDA will continue to engage YouTube to improve its information transparency to Singapore users.

YouTube's Community Guidelines	Number of videos flagged by end-users with a Singapore IP address
Spam or Misleading	233,115
Sexual	82,384
Violent or Repulsive	43,627
Harmful or Dangerous Acts	37,433
Promotes Terrorism	27,664
Child Abuse	21,685

Table 1: Data provided by YouTube for paragraph 26(a) of the Code: The number of end-user reports received from end-users in Singapore by harmful content category

YouTube's Community Guidelines	Number of videos uploaded from a Singapore IP address that were removed
Harmful or Dangerous	42,104
Child Safety	14,644
Nudity or Sexual	5,195
Violent or Graphic	4,197
Harassment and Cyberbullying	4,178
Promotion of Violence and Violent Extremism	1,202
Misinformation	606

Table 2: Data provided by YouTube for paragraph 26(a) of the Code: The number of harmful and inappropriate content removed as a result of end-user reports by harmful content category

Number of videos removed globally for violating YouTube's Community Guidelines, where the source of first detection was automated flagging	34,480,811
Number of videos uploaded from a Singapore IP address removed for violating YouTube's Community Guidelines, where the source of first detection was automated flagging	70,593

Table 3: Data provided by YouTube for paragraph 26(c) of the Code: The number of harmful or inappropriate content proactively removed by the Service by harmful content category

Number of global channels removed for violating YouTube's Community Guidelines (see table 4.1 for breakdown by content category)	**2,313,050**
Number of Singapore channels removed globally for violating YouTube's Community Guidelines (see table 4.2 for breakdown by content category)	**7,450**

Table 4: Data provided by YouTube for paragraph 26(d) of the Code: The number of accounts suspended globally and in Singapore

YouTube's Community Guidelines	Number of channels removed
Nudity or Sexual	588,165
Child Safety	532,568
Misinformation	460,191
Harassment and Cyberbullying	306,252
Harmful or Dangerous	220,169
Promotion of Violence and Violent Extremism	152,204
Violent or Graphic	53,501

Table 4.1 Breakdown by content category for global channels removed for violating YouTube's Community Guidelines

YouTube's Community Guidelines	Number of channels removed
Nudity or Sexual	3,646
Misinformation	2,041
Harmful or Dangerous	723
Harassment and Cyberbullying	485
Child Safety	440
Promotion of Violence and Violent Extremism	77
Violent or Graphic	38

Table 4.2 Breakdown by content category for Singapore channels removed for violating YouTube's Community Guidelines

10 YouTube's annual online safety report can be viewed on IMDA's website at www.imda.gov.sg/online-safety.

YouTube's Response

YouTube appreciates the continued engagement with the Singapore government on online safety and we remain committed to protecting Singaporeans online.

Annex A: Code of Practice for Online Safety – Social Media Services

BROADCASTING ACT 1994

CODE OF PRACTICE FOR ONLINE SAFETY – SOCIAL MEDIA SERVICES

1. In exercise of the powers conferred by section 45L of the Broadcasting Act 1994, the Info-communications Media Development Authority ("IMDA") hereby issues the following online Code of Practice ("Code").

Title and Commencement

2. This Code is called the Code of Practice for Online Safety – Social Media Services and shall come into effect on 18 July 2023.

Purpose of this Code

3. This Code specifies outcomes that Social Media Services ("Service") which are designated/will be designated under section 45K(1) of the Broadcasting Act 1994 have to meet to enhance online user safety, particularly for children, and curb the spread of harmful content on their service.

4. The categories of harmful content include:
   a. Sexual content
   b. Violent content
   c. Suicide and self-harm content
   d. Cyberbullying content
   e. Content endangering public health
   f. Content facilitating vice and organised crime

Application

5. This Code applies to Social Media Services which are designated/will be designated under section 45K(1) of the Broadcasting Act 1994.

Definitions

6. For the purpose of this Code, the following definitions shall apply:

   a. "community guidelines and standards" means guidelines issued by the Service on impermissible content and end-user activity.

   b. "content moderation" means processes developed and activities taken by the Service to (i) detect, whether through the Service's systems or in response to user reporting; (ii) assess; and (iii) address harmful content for end-users or content inappropriate for children on the Service in accordance with its community guidelines and standards such as by removing or restricting access to the content.

   c. "child" means an individual who is below 18 years of age.

   d. "end-user" means Singapore end-user.

Obligations

7. The obligations are categorised into three sections:

   Section A - User Safety;
   Section B - User Reporting and Resolution; and
   Section C - Accountability.

Section A – User Safety

8. End-users must be able to use the Service in a safe manner. In this regard, the Service must put in place measures to minimise end-users' exposure to harmful content, empower end-users to manage their safety on the Service and mitigate the impact on end-users that may arise from the propagation of harmful content.

9. Children in particular, may lack the capacity or experience to deal with the information and content available online and will need more protection to ensure a safer online space for them. In this regard, the Service must therefore also have specific measures to protect children from harmful content.

10. Measures to comply with the obligations in paragraphs 8 and 9 must include those found in (Ai) and (Aii) below.

(Ai.) Measures for all end-users

Community guidelines and standards and content moderation

11. End-users' exposure to harmful content must be minimised via reasonable and proportionate measures. These measures include, but are not limited to, a set of community guidelines and standards, and content moderation measures that are put in place and effected by the Service. The Service's community guidelines and standards must address the categories of harmful content in paragraph 4 and must be published.

Empower end-users and improve safety

12. End-users must have access to tools that enable them to manage their own safety and effectively minimise their exposure to, and mitigate the impact of, harmful content and unwanted interactions on the Service. Such tools may include:

    a. Tools to restrict visibility of harmful content and/or unwanted comments.

    b. Tools to limit visibility of the end-user's account, including profile and content, as well as contact and/or interactions with other end-users.

    c. Tools to limit location sharing.

13. End-users must be able to easily access information related to online safety on the Service. Such information must be easy to understand and must include the availability of tools and local information, including Singapore-based safety resources or support centres, if available. The Service should seek to implement, support and/or maintain programmes and initiatives to educate and raise awareness of such information.

14. End-users who use high-risk search terms such as, but not limited to, terms relating to self-harm and suicide on the Service must be actively offered relevant safety information (stated in paragraph 13) such as, but not limited to, local suicide prevention hotlines, if available.

Proactive detection and removal

15. End-users' exposure to child sexual exploitation and abuse material and terrorism content on the Service must be minimised through the use of technologies and processes. These technologies and processes must proactively detect and swiftly remove child sexual exploitation and abuse material and terrorism content as technically feasible, such that the extent and length of time to which such content is available on the Service is minimised.

16. End-users must be protected from preparatory child sexual exploitation and abuse activity and terrorism activity on the Service through reasonable and proportionate steps taken by the Service to proactively detect and swiftly remove preparatory child sexual exploitation and abuse activity (such as online grooming for child sexual abuse) and terrorism activity (such as glorifying or endorsing terrorist activities and recruitment).

(Aii.) Measures for children

Community guidelines and standards and content moderation

17. Besides harmful content, children's exposure to inappropriate content must also be minimised through reasonable and proportionate measures. These measures include, but are not limited to, a set of community guidelines and standards and content moderation measures put in place and effected by the Service that are appropriate for children. These community guidelines and standards must minimally address the following categories of content, and must be published:

    a. Sexual content
    b. Violent content
    c. Suicide and self-harm content
    d. Cyberbullying content

18. Children must not be targeted to receive content that the Service is reasonably aware to be detrimental to their physical or mental well-being. Such content includes the categories of harmful and/or inappropriate content in paragraphs 4 and 17. In this regard, content targeting refers, but is not limited to, advertisements, promoted content and content recommendations.

Protection for children

19. Children or their parents/guardians must have access to tools that enable them to manage children's safety, and effectively minimise children's exposure to, and mitigate the impact of, harmful and/or inappropriate content and unwanted interactions on the Service. These tools may include the following:

    a. Tools to effectively manage the content that children see and/or their experiences.

    b. Tools to:

        i. Limit the public visibility of children's accounts, including their profile and content;
        ii. Limit who can contact and/or interact with children's accounts; and
        iii. Limit location sharing.

20. Unless the Service restricts access by children, children must be provided differentiated accounts whereby the settings for the tools to minimise exposure and mitigate impact of harmful and/or inappropriate content and unwanted interactions are robust and set to more restrictive levels that are age appropriate by default. Children or their parents/guardians must be provided clear warnings of implications if they opt out of the default settings.

21. Children must be able to easily access information related to online safety on the Service. Such information must be easily understood by children and must include information on tools available to protect children from harmful and/or inappropriate content and unwanted interactions, as well as local information, including Singapore-based safety resources or support centres, if available. The Service should seek to implement, support and/or maintain programmes and initiatives to educate and raise awareness of such information.

22. Children who use high-risk search terms, such as, but not limited to, terms relating to self-harm and suicide, on the Service must be actively offered relevant safety information (stated in paragraph 21) such as, but not limited to, local suicide prevention hotlines, if available.

Section B – User Reporting and Resolution

23. Any individual must be able to report concerning content or unwanted interactions to the Service in relation to the categories of harmful and/or inappropriate content in paragraphs 4 and 17. In this regard, the reporting and resolution mechanism provided to end-users must be effective, transparent, easy to access, and easy to use.

    a. End-users' reports must be assessed, and appropriate action(s) must be taken by the Service in a timely and diligent manner that is proportionate to the severity or imminence of the potential harm. In particular, timelines must be expedited for content and activity related to terrorism. Appropriate action(s) may include:

        i. Swiftly removing the reported content or restricting access to the reported content; and

        ii. Warning, suspending, or banning the account(s) that generated, uploaded, or shared the reported content.

    b. Where the Service receives a report that is not frivolous or vexatious:

        i. The end-user who submitted the report must be informed of the Service's decision and action taken with respect to that report without undue delay.

        ii. Should the Service decide to take action against the reported content or account(s), the end-user holding the account(s) that generated, uploaded, or shared the reported content must be informed of the Service's decision and action without undue delay.

    c. The end-users referred to in sub-paragraphs (b)(i) and (b)(ii) must be allowed to submit requests to the Service for a review of the decision and action taken.

Section C – Accountability

24. End-users must have access to clear and easily comprehensible information that enable them to assess the level of safety and related safety measures afforded by the Service and make informed choices.

25. In this regard, the Service must submit to IMDA annual online safety reports on the measures the Service has put in place to combat harmful and inappropriate content, for publishing on IMDA's website. The annual online safety reports must reflect Singapore end-users' experience on the Service, including:

    a. What steps the Service has taken to mitigate Singapore end-users' exposure to harmful or inappropriate content, including descriptions of specific measures that the Service has in place to enhance online safety for end-users in Singapore in relation to obligations in Sections A and B;

    b. How much and what types of harmful or inappropriate content end-users in Singapore encounter on the Service; and

    c. What action(s) the Service has taken on end-user reports.

26. The Service may propose suitable information and metrics to be included in its annual online safety reports. These are subject to agreement by IMDA. These may include but are not limited to:

    a. The number and types of end-user reports received from end-users in Singapore, and the number and types of harmful and inappropriate content removed as a result of end-user reports;

    b. The time between the Service receiving end-user reports from end-users in Singapore on harmful and inappropriate content and taking action (if any) as an aggregate;

    c. The number and types of harmful or inappropriate content proactively removed by the Service that are:

        i. Accessible by end-users in Singapore; and

        ii. Originated from Singapore.

d. The number of accounts suspended or banned in Singapore, and the reasons for suspending or banning accounts in relation to the categories of harmful and inappropriate content in paragraphs 4 and 17.


INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY