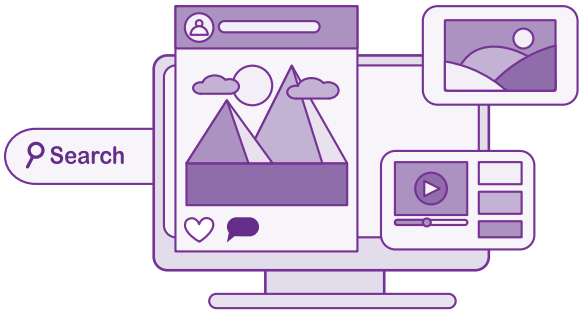
INFOCOMM MEDIA DEVELOPMENT AUTHORITY

# Online Safety Assessment Report 2024

## Designated Social Media Services

# Contents




# Preface

Social media has become ubiquitous in our everyday lives. While there are benefits, Singapore users are also encountering an increasing amount of harmful content. The Ministry of Digital Development and Information's 2024 Online Safety Poll of Singapore residents showed that 66% reported encountering harmful content on Designated Social Media Services (DSMSs).

Enhancing online safety is a shared responsibility between government, industry and users. DSMSs bear an especially significant responsibility given their reach and impact. In putting together this inaugural Online Safety Assessment Report, emphasis was placed on real-world outcomes and the effectiveness of DSMSs' measures for Singapore users, rather than simply whether measures were present or not. This was done to better equip users with relevant information so that they can decide for themselves which DSMSs to use, and manage the risks of online harms.

As this is the inaugural Report, we recognise the efforts by the DSMSs to enhance online safety for Singapore users. However, there is room for improvement, particularly in the effectiveness of safety measures for children and user reporting measures.

IMDA has engaged the DSMSs on the Report's findings. The intent is to enhance online safety over time. In the spirit of transparency, DSMSs were given the opportunity to respond to the findings and these are reflected in the Report.

DSMSs are expected to remain vigilant to online harms and enhance their measures as part of their responsibility to users, in particular children. DSMSs should continue to share transparent, up-to-date and easily understood information on their online safety measures with the public. We will continue to work with all stakeholders to enhance online safety in Singapore.

**INFOCOMM MEDIA DEVELOPMENT AUTHORITY**
**SINGAPORE**
**17 FEBRUARY 2025**

# Executive Summary

**01** The inaugural Online Safety Assessment Report aims to inform Singapore users of the online safety measures Designated Social Media Services (DSMSs) have in place, as required by the Code of Practice for Online Safety – Social Media Services (Code). It assesses the comprehensiveness and effectiveness of these measures to mitigate risks from harmful content, and highlights areas for improvement. This allows users to make informed decisions about the risks and available safety measures, and ensures that DSMSs are accountable for providing a safe user experience. The six DSMSs are Facebook, HardwareZone, Instagram, TikTok, X and YouTube.

## Code of Practice for Online Safety – Social Media Services

**02** The Code was published in July 2023. The six categories of harmful content DSMSs must address are:

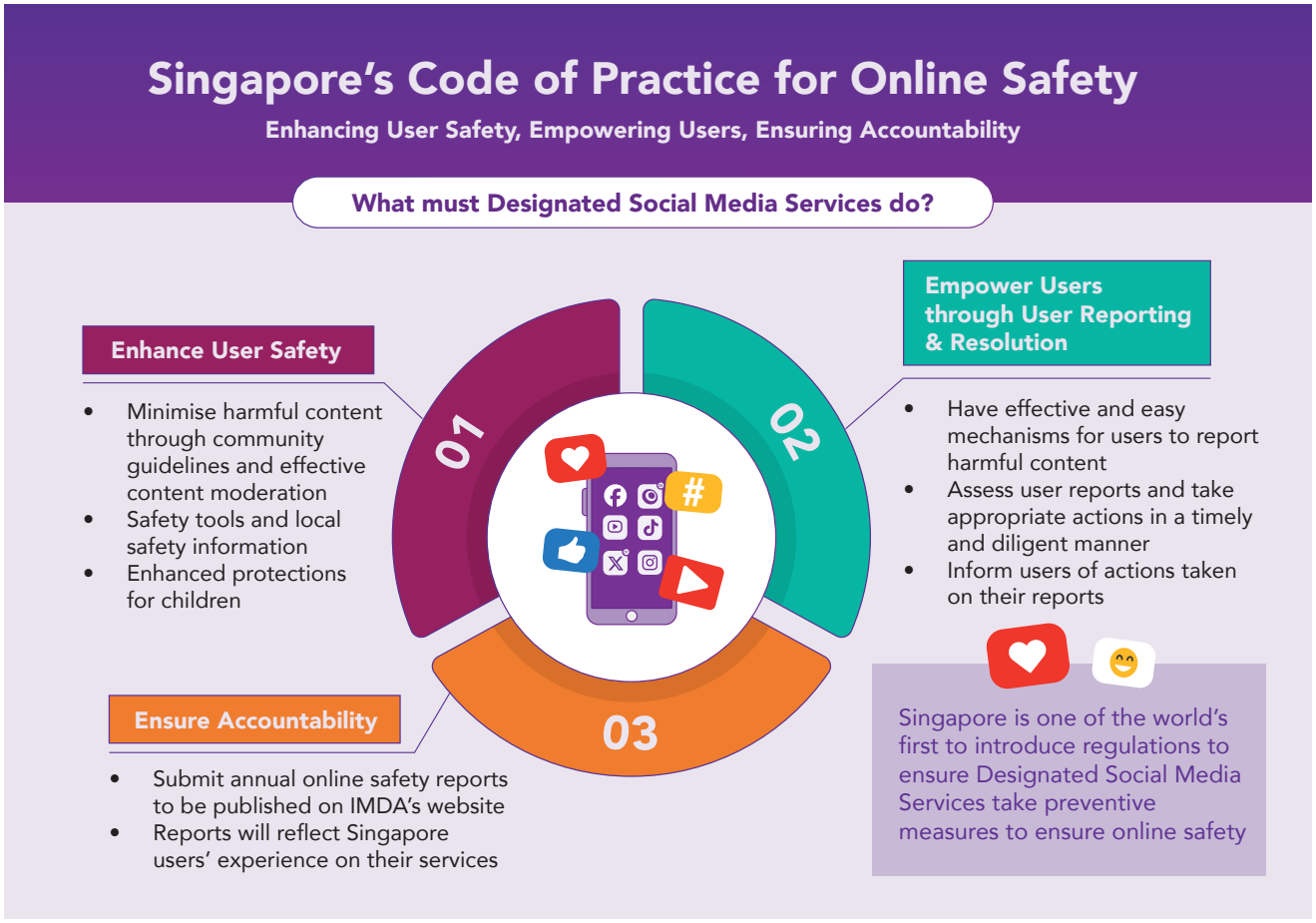

1. Sexual content
2. Violent content
3. Suicide and self-harm content
4. Cyberbullying content
5. Content endangering public health
6. Content facilitating vice and organised crime

**03** 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.




## Methodology

**04** DSMSs were assessed on whether their measures were **comprehensive** and **effective** in achieving the Code's safety outcomes. To assess effectiveness, test accounts were set up to simulate real-world user experiences. "Mystery shopper" tests were also conducted, by reporting harmful content that violated the DSMSs' own community guidelines, to assess if user reporting and resolution mechanisms were effective. Please refer to the main report for details on the methodology.

## Online Safety Ratings

**05** Each DSMS received (a) an **Overall Rating** and (b) **Ratings for Individual Sections of the Code**. Overall, the ratings show that DSMSs performed better in User safety measures for all users, and Accountability. They were weaker in User safety measures for children, and User reporting and resolution.

| 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 & resolution | Section C: Accountability |
| Facebook | ☑☑☑◪☐ | ☑☑☑☑☑ | ☑☑☑☑☐ | ☑☑◪☐☐ | ☑☑☑◪☐ |
| HardwareZone | ☑☑☑◪☐ | ☑☑☑☑☑ | ☑☑◪☐☐ | ☑☑☑☑☐ | ☑☑☑☑☑ |
| Instagram | ☑☑☑◪☐ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☐☐☐ | ☑☑☑◪☐ |
| TikTok | ☑☑☑☑☐ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑◪☐☐ | ☑☑☑☑☑ |
| X | ☑☑◪☐☐ | ☑☑◪☐☐ | ☑☑☐☐☐ | ☑☑☐☐☐ | ☑☑☑☑☑ |
| YouTube | ☑☑☑◪☐ | ☑☑☑☑☑ | ☑☑☑☑☐ | ☑☑☑☐☐ | ☑☑☑◪☐ |

## Summary of Key Findings by Section

### Section Ai: User safety measures for all end-users

**06** DSMSs have largely put in place the required user safety measures, including:


a. Community guidelines covering the six categories of harmful content in the Code, and both human and automated content moderation measures.

b. Tools for users to manage their own safety, such as tools to restrict visibility of harmful content and/or unwanted comments, limit the visibility of user accounts, posted content and interactions with others, and limit location sharing where applicable.


c. Easily accessible and understood online safety information such as Help Centres and pages. Facebook, Instagram, TikTok and YouTube also implemented and supported additional programmes and initiatives to educate and raise awareness of such information in Singapore.


d. Singapore-based safety resources for users who searched high-risk terms related to suicide and self-harm. Some DSMSs offered additional resources for search terms related to domestic violence, sexual violence and cyberbullying. Facebook, Instagram and TikTok also offered additional resources such as options to contact a friend for support, or links to professional and digital wellness resources.

**07** X needs to improve the effectiveness of its efforts to detect and remove CSEM on its service. CSEM is a very egregious type of harm, and the Code requires DSMSs to use technology to proactively detect and swiftly remove CSEM before users encounter such content. X's own publicly stated policy "prohibits any content that depicts or promotes child sexual exploitation". In its annual report, X stated that it proactively detected and removed 6 pieces of CSEM originating from Singapore. However, our tests detected considerably more cases of CSEM originating from Singapore on X during the same period. X will therefore need to provide IMDA with an update on steps taken to improve the effectiveness of its measures against CSEM.

## Section Aii: User safety measures for children

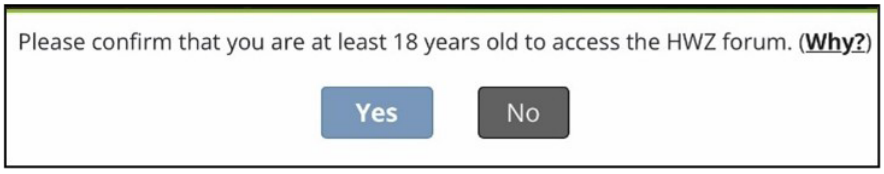
**08** DSMSs should do more to improve the effectiveness of their measures to protect children from harmful and age-inappropriate content.

a. All DSMSs had community guidelines appropriate for children. However, the enforcement of these community guidelines should be improved as children's accounts could still access harmful and age-inappropriate content on some DSMSs.

   i. Facebook and YouTube had instances where children's accounts could access content that should have been restricted under their own community guidelines. These include digital imagery of adult content on Facebook, and videos with sexually suggestive imagery on YouTube.

   ii. X did not effectively enforce its policies to restrict children's accounts from viewing adult sexual content. X's own publicly stated policy states that it restricts "viewers who are under 18, or who do not include a birth date on their profile, from viewing adult content". However, our tests found that children's accounts could easily find and access explicit adult sexual content, especially hardcore pornography, with simple search terms. X will need to provide IMDA with an update on steps taken to improve the effectiveness of its measures to prevent children from accessing age-inappropriate content, especially adult sexual content.

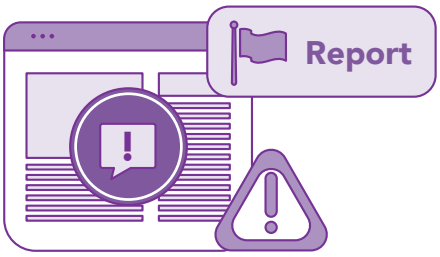
b. HardwareZone's Terms of Service prohibit users under the age of 18 from accessing the service. However, its age-gating measure to enforce this was easily bypassed. HardwareZone should either effectively restrict children from accessing its service or put in place comprehensive safety measures for children as required by the Code. HardwareZone will need to provide IMDA with an update on this.




Screenshot of HardwareZone's Age-Gating Measure

**09** Today, the Code already requires DSMSs to ensure that children receive age-appropriate content and experiences. However, we have allowed DSMSs to decide how best to achieve this outcome, and not required them to implement age assurance measures. From our monitoring, IMDA assesses that age assurance technology has improved considerably. IMDA has already made it a requirement for Designated App Distribution Services to implement age assurance measures to ensure children and youth do not download apps that are inappropriate for their age. We are also studying how Social Media Services should use age assurance technology to better protect children and youth online.

## Section B: User reporting and resolution




**10** DSMSs have prioritised technologies to automatically detect and remove harmful content at-scale. However, it is important that they take user reports seriously as it is a critical pathway for recourse for a user encountering online harms.

**11** DSMSs should improve the effectiveness and timeliness of their responses to user reports. The Code requires DSMSs to take appropriate action on user reports of harmful content that violates their community guidelines in a timely and diligent manner.

a. Our "Mystery shopper" tests (see results in Table 1) found that most DSMSs took appropriate actions on only approximately 50% or less of the content that violated their own community guidelines. When we subsequently notified the DSMSs to re-review the remaining pieces of harmful content, all were found to be violating their own community guidelines and subsequently removed. This means that a significant proportion of legitimate user reports were not actioned on in the first instance.

b. Most DSMSs also took an average of 5 days or more to act on user reports of harmful content that violated their own community guidelines. This was considerably longer than what was stated in their annual reports[1].

| DSMS | DSMS Action Rates on User Reports of Harmful Content that Violate their Community Guidelines | Average Time to Action |
|---|---|---|
| Facebook | 53% | 9 days |
| HardwareZone | 89% | 3 days |
| Instagram | 2% | 7 days |
| TikTok | 39% | 5 days |
| X | 54% | 10 days |
| YouTube | 46% | 5 days |

Table 1: Results from Mystery Shopper Test of DSMSs' User Reporting and Resolution Mechanisms

[1]HardwareZone reported an average time of 21 hours to take action on user reports; TikTok reported that 90.7% of videos taken down pursuant to user reports were removed within 24 hours; X reported that its median time for taking action on user reports was 15.06 hours. Facebook, Instagram and YouTube did not provide any relevant data on the time to take action on user reports.

12 These findings corroborate the Ministry of Digital Development and Information's 2024 Online Safety Poll which found that between 78% to 86% of respondents faced issues with user reporting on DSMSs, including: (a) the service did not take down the harmful online content reported or disable the account responsible for it, (b) the service did not provide updates on the outcomes of their reports and (c) the service allowed the removed content to be reposted. All DSMSs, with the exception of HardwareZone, will need to provide an update on steps taken to improve their effectiveness and timeliness of user reporting measures.

## Section C: Accountability

13 DSMSs should 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. As a baseline, the Code requires DSMSs to provide information on their measures and data to demonstrate the effectiveness of these measures. This is to be done through annual online safety reports.

14 In their first annual online safety reports, all DSMSs met this baseline standard. DSMSs submitted their annual reports on time and with clear information. In addition, the Code places emphasis on data which reflects the impact of the DSMSs' online safety efforts in Singapore. While some DSMSs were able to provide such data, other DSMSs did not or were unable to do so. Facebook, Instagram and YouTube should have provided data for Singapore users to understand the effectiveness and timeliness of their user reporting and resolution mechanisms. It is important for Singapore users to have access to up-to-date online safety data relevant to them. DSMSs should therefore improve on information transparency to Singapore users.

## Conclusion

15 The DSMSs' online safety measures were generally comprehensive. However, there are areas which require improvement, particularly in the effectiveness of these measures. DSMSs will need to provide updates on the steps taken to improve their areas of weakness in their next annual online safety reports.

IMDA's Online Safety Assessment Report 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 Globally, there is consensus on the need for social media services to take greater responsibility in protecting their users, especially children. Singaporeans have also expressed concern over the potential damage caused by harmful online content and expect social media services to take greater responsibility to protect their users.

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 (OCSs), such as social media services. The OSMAA empowers 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 (Code) which took effect from 18 July 2023. Six social media services with significant reach or impact in Singapore 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.

04 The key requirements of the Code are as follows:

a. **Section Ai: User safety measures for all end-users.** DSMSs are responsible for ensuring that their service is safe for their users. They must have measures in place to minimise users' exposure to harmful content and empower users with tools and information to manage their safety on the services. This includes published community guidelines, effective content moderation, self-help tools to enable users to manage their own safety, easily accessible and understood online safety information, and proactive detection and removal of child sexual exploitation and abuse material (CSEM) and terrorism content.

b. **Section Aii: User safety measures for children.** DSMSs must have enhanced protections for children (users under the age of 18) to minimise their exposure to age-inappropriate content. These measures include age-appropriate community guidelines for children, additional tools for children or their parents/guardians to manage children's safety, and differentiated accounts for children with more restrictive default settings.

c. **Section B: User reporting and resolution.** DSMSs must provide their users with effective, transparent, easy to access and easy to use mechanisms to report harmful content and unwanted interactions. DSMSs are required to take appropriate action on user reports of harmful content that violate their community guidelines in a timely and diligent manner, and inform the relevant users of their decision and any action taken in response to the user reports.

d. **Section C: Accountability.** DSMSs must be transparent and accountable to their users. They should submit annual reports with clear information on their safety measures, supported by suitable data that reflect 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.

Please refer to Annex A for the full text of the Code.

## Aim of the Online Safety Assessment Report

**05** This Report aims to inform Singapore users of:

a. The DSMSs' online safety measures in place as required by the Code.

b. The comprehensiveness and effectiveness of these online safety measures to mitigate risks from harmful social media content.

c. The areas for improvement required from the DSMSs.

**06** 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 ensure that the DSMSs are accountable for providing a safe experience for their users.

## Methodology

**07** The Report was prepared using the following sources:

a. Information and data from the DSMSs' annual online safety reports covering the period 1 August 2023 – 31 July 2024, 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.

**08** For each requirement in the Code, IMDA assessed the DSMSs according to whether the online safety measures were **present**, **comprehensive** and **effective**.

**09** 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, we tested how easily child 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 child 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 used 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.*

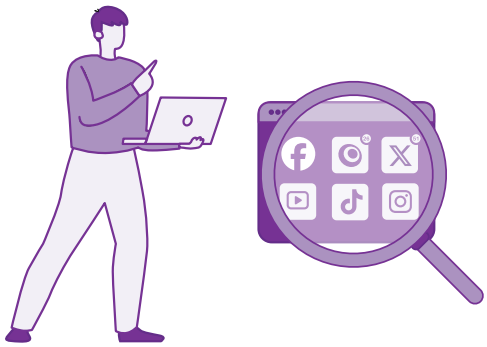
*Findings: The test found that all six DSMSs offered relevant online safety information to varying degrees. When keywords related to suicide and self-harm were searched, all DSMSs offered information such as the Samaritans of Singapore hotline. Some DSMSs also provided resources for keywords related 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, over 1,000 pieces of harmful content that violated the DSMSs' own community guidelines were 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. As this is the first year of the Report, the sample sizes for each DSMS ranged from between 35 to 365 pieces, depending on the ease of detecting harmful content with a Singapore nexus on each DSMS. For future years, we hope to increase the sample sizes.*

*Findings: Most DSMSs did not take appropriate and timely action on a significant proportion of legitimate user reports that violated their own community guidelines. Please refer to the Executive Summary and Main Report for more details on the findings.*

IMDA's Online Safety Assessment Report 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.



# 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 & resolution | Section C: Accountability |
| ☑☑☑☑☐ | ☑☑☑☑☑ | ☑☑☑☑☐ | ☑☑☑☐☐ | ☑☑☑☑☐ |

### Section Ai: User safety measures for all end-users

**01** **Facebook had the required user safety measures for all users.**

a. Facebook had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. Facebook had a variety of tools for users to manage their own online safety. Users are provided multiple options to restrict visibility of harmful content and/or unwanted comments, limit the visibility of their accounts and content, and limit location sharing.

c. Facebook provided easily accessible and understood online safety information such as the Meta Safety Centre, Women's Safety Hub, Facebook Help Centre and a Transparency Centre. Meta also implemented and supported various programmes and initiatives to educate and raise awareness of online safety in Singapore such as the EYEYAH! x Meta Youth Digital Wellness Program, participating in the annual Digital for Life Festival, hosting webinars for parents on "Navigating the digital world with your teen", supporting the Ministry of Education with bi-annual learning journeys to Meta, and organising an APAC Youth Safety Summit.

d. Facebook actively offered local safety resources to users who search for high-risk terms. For example, users who search for terms such as "suicide" and "depression" are provided links to the Samaritans of Singapore and Institute of Mental Health respectively. Users are also provided the option to "Contact a Friend" or see suggestions from professionals outside of Meta.

e. Facebook reported comprehensive measures to proactively detect and remove CSEM and terrorism content. During the period of assessment, there were no CSEM or terrorism cases on Facebook detected by or reported to IMDA.

## Section Aii: User safety measures for children

**02** **Facebook had most of the required user safety measures for children.**

a. Facebook permits users aged 13 and above to use its service and applied more restrictive default settings to children's accounts, such as hiding age-inappropriate content, limiting content recommendations, prohibiting advertisements about restricted topics, who can see their friends list and who is allowed to comment on their public posts.

b. Facebook provided parents/guardians with tools to manage children's safety. For example, Parental Supervision is a set of tools and insights that parents/guardians can use to support their children on Facebook with more positive and age-appropriate experiences, including seeing some of their child's privacy settings and content preferences.

**03** Facebook's enforcement of its community guidelines for children should be improved. Age-inappropriate content that should have been restricted for children, such as digital imagery of adult content, was found to be accessible via a child's account.

## Section B: User reporting and resolution

**04** **Facebook needs to improve the effectiveness of its user reporting systems.** Based on our tests, Facebook took action on only 53% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified Facebook directly, Facebook took action on the remaining 47% of harmful content. This suggests that all the content was violative of Facebook's community guidelines and should have been actioned on when initially user reported.

a. Facebook did not provide data on the number and types of harmful content removed as a result of Singapore user reports and should be more transparent in this area. Facebook stated in its annual report that "Our focus has been on logging data on how well we are able to proactively detect and remove content that violates our Community Standards, before a user reports it to us. The metrics we have invested in reporting have therefore centred on the enforcement and proactive rates for specific violation types in our Transparency Centre. As such, we are unable to provide a breakdown per violation type arising from reports by end-users in Singapore alone."

05 **Facebook needs to improve the timeliness of its user reporting systems.** Based on our tests, Facebook took an average time of 9 days to take action on harmful content that violated its community guidelines when reported by user accounts.

a. Facebook took faster action on Suicide and Self-Harm content, with an average time of 2 days compared to the other categories of harmful content which ranged from 7 to 15 days.

b. Facebook did not provide data on the time it took to take action on Singapore user reports and should be more transparent in this area. Facebook stated in its annual report that "We are unable to provide metrics on our response times to user reports. We would additionally note that not all user reports are equal in terms of the level of risk and harm it may cause. Some reports may be benign, where there is no immediate harm or there are no or only a small number of views, while other posts may be spreading rapidly and pose a greater harm to individuals. We therefore do not prioritise the review of content reported to us by chronology. Instead, we prioritise content for review based on the severity of harm, whether it was reported to us or detected by our proactive systems."

06 **Facebook should do more to educate its users on its community guidelines and address public perception that its user reporting systems are ineffective.** In its annual report, Facebook stated, "Users often do not understand our policies, and the majority of reports from users is content that does not violate our policies. End-user reporting metrics therefore is not a reliable indicator of a policy violation." However, our tests above show that this is not necessarily true – Facebook did not take appropriate action on almost half of legitimate reports of violating content.

## Section C: Accountability

07 **Facebook provided clear information in its annual online safety report, although there is room for improvement in its supporting data.**

a. Facebook did not provide data to demonstrate whether it assesses and takes the appropriate action on user reports in a timely manner. Examples of such data which Facebook should have provided include:
   i. Data on the number and types of harmful and inappropriate content removed as a result of user reports, pursuant to paragraph 26(a) of the Code; and
   ii. Data on the time it took to take action on user reports, pursuant to paragraph 26(b) of the Code.

b. 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 broken down by the six harmful content categories, both globally and in Singapore.

c. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore", Facebook reported that it disabled over 337,600 user accounts on Facebook created in Singapore for violating its community guidelines (excluding fake accounts).

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

## Meta's Response

*At Meta, the online safety of our users is a top priority. We take a comprehensive approach to making our technologies a better place for everyone and have invested significantly in safety tools for users; detection technologies to reduce the prevalence of harmful content; teams of experts that work on safety and security around the globe, including in Singapore; and partnerships with safety, youth, and other organisations. We welcome IMDA's assessment of our first report for the Code of Practice for Online Safety, which recognises the extensive efforts Meta has made over the years in user safety measures for all our users (Section Ai), including for children (Section Aii), on Facebook and Instagram.*

*On user reporting and resolution (Section B), Meta has made significant investments to build easily accessible, intuitive user reporting tools, while continuously working to improve our review processes and technologies to ensure we are properly responding to user reports and prioritising the most urgent ones. We welcome IMDA's feedback and always take this into account in our continuous work to enhance our user reporting systems.*

*Given the sheer volume of user reports Facebook and Instagram receives everyday, our review systems use technology to prioritise high-severity content with the potential for imminent offline harm (e.g., posts related to terrorism and suicide) and viral content that is spreading quickly and has the potential to reach a large audience, in order to prevent as much harm as possible. We, therefore, note that the "Mystery Shopper" approach used to test and assess our user reporting systems might not have considered platforms' prioritisation of user reports.*

*On accountability (Section C), Meta's goal is to minimise the impact caused by violations of our policies on our users by reducing the prevalence (or views) of that content. As such, our years of investment in measuring our content moderation efforts has focused on how effectively we detect and take action on policy-violating content before users report them to us. Therefore, the metrics we have published in our quarterly Community Standards Enforcement Report (CSER) on our Transparency Center focus on content actioned and proactive rates for specific categories of harmful content, and not metrics around user reporting.*

*We took the same approach for this report, in which we measured and reported on the action and proactive rates for policy-violating content in Singapore. Because of this, as well as how we prioritise the review of user reports as noted above, we were unable to provide a breakdown per violation type arising from user reports in Singapore alone or time taken to respond to user reports from Singapore as an aggregate.*

## 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 & resolution | Section C: Accountability |
| ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☐☐ | ☑☑☑☑☐ | ☑☑☑☑☑ |

### Section Ai: User safety measures for all end-users

**01** **HardwareZone had the required user safety measures for all users.**

a. HardwareZone had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. HardwareZone provided multiple tools for users to restrict visibility of harmful content and/or unwanted comments and limit the visibility of their accounts and content. HardwareZone does not capture or show the location of its users, and therefore does not have any tools to limit location sharing.

c. HardwareZone provided easily accessible and understood online safety information such as its Help & Safety Resources page and Forum Content Policy. However, HardwareZone did not implement any programmes and initiatives to educate and raise awareness of such information for its users as required by Paragraph 13 of the Code.

d. HardwareZone actively offered local safety resources to users who search for high-risk terms. For example, users who search for terms such as "suicide", "depression", "cyberbullying" and "sexual harassment" are provided links to the Samaritans of Singapore hotline, Mindline.sg and AWARE respectively.

e. HardwareZone had measures to proactively detect and remove CSEM and terrorism content. HardwareZone also noted in its annual report that the "incidence rate of such content is very low and even if present, gets promptly reported by end-users to the HWZ team and removed." During the period of assessment, there were no CSEM or terrorism cases on HardwareZone detected by or reported to IMDA.

## Section Aii: User safety measures for children

**02** **HardwareZone should either effectively restrict children from accessing its service or put in place comprehensive safety measures for children as required by the Code.**

a. HardwareZone's Terms of Service prohibit users under the age of 18 from accessing the service and therefore stated in its annual report that Section Aii was not applicable to it. However, its age-gating measure to enforce this was easily bypassed.

b. Some of HardwareZone's measures for all users partially met the requirements for children, such as its community guidelines which is set at a stricter threshold that covers age-inappropriate content for children, general safety tools for all users that could also be used by children, and online safety information in its Help Centre that was easily accessible and understood. However, our tests found that HardwareZone needs to improve the enforcement of their policies as children who bypass the age-gating measure could access age-inappropriate content such as sexually suggestive references/innuendos that should have been prohibited for all users.

c. HardwareZone did not have additional measures for children such as accounts with more restrictive default settings and dedicated tools for parents/guardians to manage children's safety.

## Section B: User reporting and resolution

**03** **HardwareZone had effective user reporting systems.** Based on our tests, HardwareZone took the appropriate action on 89% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified HardwareZone directly, HardwareZone took action on the remaining 11% of harmful content. This suggests that all the content was violative of HardwareZone's community guidelines and should have been actioned on when initially user reported.

**04** **HardwareZone took timely action on user reports albeit with some exceptions.** Based on our tests, HardwareZone took an average time of 3 days to take action on user reports of harmful content that violated its community guidelines when reported by user accounts.

a. In contrast, HardwareZone's annual report stated that its average time to take action on user reports was 21 hours.

b. HardwareZone's report also stated that it took an average time of 129 hours to resolve 5 reports related to Suicide and Self-harm content and explained that HardwareZone "generally found the content to be vague and the HWZ Team had to monitor the forum members and the discussion for a longer period of time to determine the validity of the reports, authenticity and severity of the content."

## Section C: Accountability

**05** **HardwareZone provided clear information in its annual online safety report with supporting 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 was able to provide the data broken down by the six harmful content categories.

b. For paragraph 26(b) of the Code on time it took to take action on user reports, HardwareZone was able to provide the data 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's report stated that it proactively removed a total of 115 pieces of harmful content, including 76 pieces of Sexual content, 34 pieces of Cyberbullying content and 5 pieces of Violent content respectively without user reports.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore," HardwareZone banned a total of 38 accounts including 25 for posting Sexual content, 6 for posting Cyberbullying content, 5 for posting Violent content and 2 for posting Content endangering public health.

**06** 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 HWZ's annual report and compliance with the Online Safety Code. HWZ is committed to ensuring the online safety of all users in the fast-changing internet landscape.*

## 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 & resolution | Section C: Accountability |
| ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☐☐☐ | ☑☑☑☑☐ |

### Section Ai: User safety measures for all end-users

**01** **Instagram had the required user safety measures for all users.**

a. Instagram had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. Instagram had a variety of tools for users to manage their own online safety. Users are provided multiple options to restrict visibility of harmful content and/or unwanted comments, limit the visibility of their accounts and content, and limit location sharing.

c. Instagram provided easily accessible and understood online safety information such as the Meta Safety Centre, Women's Safety Hub, Instagram Help Centre and a Transparency Centre. Meta also implemented and supported various programmes and initiatives to educate and raise awareness of online safety in Singapore such as the EYEYAH! x Meta Youth Digital Wellness Program, participating in the annual Digital for Life Festival, hosting webinars for parents on "Navigating the digital world with your teen", supporting the Ministry of Education with bi-annual learning journeys to Meta, and organising an APAC Youth Safety Summit.

d. Instagram actively offered local safety resources to users who search for high-risk terms. For example, users who search for terms such as "suicide" and "depression" are provided links to the Samaritans of Singapore and Institute of Mental Health respectively. Users are also provided the option to "Contact a Friend" or see suggestions from professionals outside of Meta.

e. Instagram reported comprehensive measures to proactively detect and remove CSEM and terrorism content but should be vigilant for CSEM on its service. During the period of assessment, there was 1 case of CSEM on Instagram that was not proactively detected and removed until IMDA notified Instagram.

## Section Aii: User safety measures for children

**02** **Instagram had the required user safety measures for children.**

a. Instagram permits users aged 13 and above to use its service and applies more restrictive default settings to children's accounts, such as hiding age-inappropriate content, limiting content recommendations, prohibiting advertisements about restricted topics, and more restrictive direct messaging settings. Our tests also found that Instagram's safeguards to prevent children from accessing harmful and inappropriate content were generally effective.

b. Instagram provided parents/guardians with access to tools that enable them to manage children's safety. For example, Parental Supervision is a set of tools and insights that parents and guardians can use to help support their teens on Instagram with more positive and age-appropriate experiences, including seeing some of their child's privacy settings and content preferences.

## Section B: User reporting and resolution

**03** **Instagram needs considerable improvement on the effectiveness of its user reporting systems.** Based on our tests, Instagram took action on only 2% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified Instagram directly, Instagram took action on the remaining 98% of harmful content. This suggests that all the content was violative of Instagram's community guidelines and should have been actioned on when initially user reported.

a. Instagram did not provide data on the number and types of harmful content removed as a result of Singapore user reports and should be more transparent in this area. Instagram stated in its annual report that "Our focus has been on logging data on how well we are able to proactively detect and remove content that violates our Community Standards, before a user reports it to us. The metrics we have invested in reporting have therefore centred on the enforcement and proactive rates for specific violation types in our Transparency Centre. As such, we are unable to provide a breakdown per violation type arising from reports by end-users in Singapore alone."

**04** **Instagram needs to improve the timeliness of its user reporting systems.** Based on our test, Instagram took an average time of 7 days to take action on harmful content that violated its community guidelines when reported by user accounts.

a. Instagram did not provide data on the time it took to take action on Singapore user reports and should be more transparent in this area. Instagram stated in its annual report that "We are unable to provide metrics on our response times to user reports. We would additionally note that not all user reports are equal in terms of the level of risk and harm it may cause. Some reports may be benign, where there is no immediate harm or there are no or only a small number of views, while other posts may be spreading rapidly and pose a greater harm to individuals. We therefore do not prioritise the review of content reported to us by chronology. Instead, we prioritise the most critical content to be reviewed first, whether it was reported to us or detected by our proactive systems."

**05** **Instagram should also do more to educate its users on its community guidelines and address public perception that its user reporting systems are ineffective.** In its annual report, Instagram stated that "Users often do not understand our policies, and the majority of reports from users is content that does not violate our policies. End-user reporting metrics therefore is not a reliable indicator of a policy violation." However, our tests above show that this is not necessarily true – Instagram did not take appropriate action on a significant proportion of legitimate reports of violating content.

## Section C: Accountability

**06** **Instagram provided clear information in its annual online safety report, although there is room for improvement in its supporting data.**

a. Instagram did not provide data to demonstrate whether Instagram assesses and takes the appropriate action on user reports in a timely manner. Examples of such data which Instagram should have provided include:
   i. Data on the number and types of harmful and inappropriate content removed as a result of user reports, pursuant to paragraph 26(a) of the Code; and
   ii. Data for the time it took to take action on user reports, pursuant to paragraph 26(b) of the Code.

b. 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 broken down by the six harmful content categories, both globally and in Singapore.

c. 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 549,900 user accounts on Instagram created in Singapore for violating our Community Standards (excluding fake accounts).

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

## Meta's Response

*At Meta, the online safety of our users is a top priority. We take a comprehensive approach to making our technologies a better place for everyone and have invested significantly in safety tools for users; detection technologies to reduce the prevalence of harmful content; teams of experts that work on safety and security around the globe, including in Singapore; and partnerships with safety, youth, and other organisations. We welcome IMDA's assessment of our first report for the Code of Practice for Online Safety, which recognises the extensive efforts Meta has made over the years in user safety measures for all our users (Section Ai), including for children (Section Aii), on Facebook and Instagram.*

*On user reporting and resolution (Section B), Meta has made significant investments to build easily accessible, intuitive user reporting tools, while continuously working to improve our review processes and technologies to ensure we are properly responding to user reports and prioritising the most urgent ones. We welcome IMDA's feedback and always take this into account in our continuous work to enhance our user reporting systems.*

*Given the sheer volume of user reports Facebook and Instagram receives everyday, our review systems use technology to prioritise high-severity content with the potential for imminent offline harm (e.g., posts related to terrorism and suicide) and viral content that is spreading quickly and has the potential to reach a large audience, in order to prevent as much harm as possible. We, therefore, note that the "Mystery Shopper" approach used to test and assess our user reporting systems might not have considered platforms' prioritisation of user reports.*

*On accountability (Section C), Meta's goal is to minimise the impact caused by violations of our policies on our users by reducing the prevalence (or views) of that content. As such, our years of investment in measuring our content moderation efforts has focused on how effectively we detect and take action on policy-violating content before users report them to us. Therefore, the metrics we have published in our quarterly Community Standards Enforcement Report (CSER) on our Transparency Center focus on content actioned and proactive rates for specific categories of harmful content, and not metrics around user reporting.*

*We took the same approach for this report, in which we measured and reported on the action and proactive rates for policy-violating content in Singapore. Because of this, as well as how we prioritise the review of user reports as noted above, we were unable to provide a breakdown per violation type arising from user reports in Singapore alone or time taken to respond to user reports from Singapore as an aggregate.*

## 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 & resolution | Section C: Accountability |
| ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ |

### Section Ai: User safety measures for all end-users

**01** **TikTok had the required user safety measures for all users.**

a. TikTok had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. TikTok had a variety of tools for users to manage their own online safety. Users are provided multiple options to restrict visibility of harmful content and/or unwanted comments, limit the visibility of their accounts and content, and limit location sharing.

c. TikTok provided easily accessible and understood online safety information such as the TikTok Safety Centre, Guardian's Guide for parents/guardians, a Digital Wellness Hub and a newsroom with Singapore-specific updates. TikTok also implemented programmes and initiatives to educate and raise awareness of such information in Singapore. For example, TikTok worked with the Ministry of Culture, Community and Youth, the Ministry of Education, National Youth Council and IMDA to organise the Youth for Good programme to raise awareness on mental and cyber wellness. They also partnered with the Straits Times to publish an article on how parents and guardians can keep their children safe on TikTok.

d. TikTok actively offered local safety resources to users who search for high-risk terms. For example, users who search for terms such as "suicide" and "sexual assault" are provided with links to the Samaritans of Singapore and the Institute of Mental Health, and a sexual abuse support centre respectively. TikTok also offers additional tips and support such as "Talk to someone you trust", "Take time out", etc. TikTok also has a local edition of its Digital Wellness Hub which provides resources on digital wellness and self-care issues.

e. TikTok reported comprehensive measures to proactively detect and remove CSEM and terrorism content. During the period of assessment, there were no CSEM or terrorism cases on TikTok detected by or reported to IMDA.

## Section Aii: User safety measures for children

**02** **TikTok had the required user safety measures for children.**

a. TikTok permits users aged 13 and above to use its service and applies more restrictive default settings to children's accounts. For example, all children between 13 to 17 have their accounts set to private and will not be recommended to adult users by default. Children are also not allowed to host live content, engage in financial transactions or download videos by default. In addition, children between 13 to 15 do not have access to direct messaging, will not be shown content in the "For You" feed from people that they do not know, and only their friends are allowed to comment on their content.

b. TikTok's Family Pairing feature allows parents/guardians to link their accounts to their children's accounts to customise a variety of safety and privacy settings for their children. For example, parents/guardians can make their children's account private, control who can message their children, set password-protected screen time limits and content preferences for their children's accounts.

c. In addition to the TikTok Safety Centre, TikTok also has a Youth Portal which covers the basics of digital literacy and shares information on its online safety tools for children in an easily understood manner.

## Section B: User reporting and resolution

**03** **TikTok needs to improve the effectiveness of its user reporting systems.** Based on our tests, TikTok took action on 39% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified TikTok directly, TikTok took action on the remaining 61% of harmful content. This suggests that all the content was violative of TikTok's community guidelines and should have been actioned on when initially user reported.

a. While it was not possible to compare IMDA's data with TikTok's data in its annual report, TikTok's data provides an indication of how many pieces of harmful content were removed as a result of user reports. TikTok stated that it "evaluated 740,801 videos that were reported by end-users in Singapore and removed 49,559 of these videos. During the same period, 106,851 videos originating from Singapore were removed as a result of end-user reports globally". Of these videos originating from Singapore that TikTok removed, the top two categories of harmful content were "Regulated Goods & Commercial Activities" and "Sensitive & Mature Themes" under TikTok's community guidelines.

**04** **TikTok needs to improve the timeliness of its user reporting systems.** Based on our tests, TikTok took an average time of 5 days to take action on harmful content that violated its community guidelines when reported by user accounts.

a. TikTok took action on Violent content, and Content facilitating vice and organised crime faster with an average time of 2 and 3 days respectively compared to the other categories of harmful content which ranged from 4 to 8 days.

b. In contrast, the data provided in TikTok's annual report stated that 90.7% of videos taken down pursuant to user reports were removed within 24 hours.

## Section C: Accountability

**05** **TikTok provided clear information in its annual online safety report with supporting 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 was able to provide 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, TikTok stated that during the reporting period, 90.7% were removed within 24 hours, calculated from when the report was submitted to when the video was taken down.

c. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service," TikTok stated that it proactively removed a total of 566,476,953 videos globally and included a breakdown by the harmful content categories in its community guidelines. Out of these videos, 2,616,072 videos originating from Singapore were proactively removed, with 92.7% removed within 24 hours and 88.8% removed before they were viewed. The top two categories of harmful content removed proactively were "Sensitive & Mature Themes" and "Regulated Goods & Commercial Activities" under TikTok's community guidelines.

d. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore," TikTok reported that 79,973 accounts originating from Singapore were removed. Of these accounts, 57,548 accounts were removed on the basis that users were suspected to be under the age of 13.

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

## TikTok's Response

*TikTok's strong partnerships with the local authorities and community partners supplement our moderation, user reporting and enforcement processes, and are critical in fostering a safe online environment for our ecosystem of users.*

*We appreciate our continued collaboration with IMDA, and IMDA's recognition of our efforts to create a safe online space for our community, including for our young users.*

*Safety has no finish line, and we are continuously refining our approach to prioritize moderation accuracy, minimize views of violative content, and remove such content quickly.*

## 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 & resolution | Section C: Accountability |
| ☑☑☑☐☐ | ☑☑☑☐☐ | ☑☑☐☐☐ | ☑☑☐☐☐ | ☑☑☑☑☑ |

### Section Ai: User safety measures for all end-users

**01** **X had most of the required user safety measures for all users.**

a. X had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. X listed a variety of tools for users to manage their own online safety without elaboration. X's annual report should better explain the functionality of these tools and how these tools meet the intended outcomes stated in the Code.

c. X provided easily accessible and understood online safety information such as the X Help Center and a dedicated page on its Help Center with resources for users if they or someone they know is struggling or in crisis. However, X did not implement any programmes and initiatives to educate and raise awareness of such information for its users as required by Paragraph 13 of the Code.

d. X actively offered local safety resources to users who search for high-risk terms. For example, users who search for terms such as "suicide" and "domestic violence" are provided links to the Samaritans of Singapore hotline, and to the Ministry of Social and Family Development's National Care hotline and AWARE Women's helpline respectively.

**02** **X's proactive detection and removal of CSEM requires considerable improvement.**

a. Paragraph 15 of the Code requires DSMSs to minimise users' exposure to CSEM through the use of technologies and processes that proactively detect and swiftly remove CSEM. CSEM is a very egregious type of harm that should be removed by the DSMSs expeditiously.

b. X's Child Safety policy states that "any content that depicts or promotes child sexual exploitation is strictly prohibited on X". In X's annual report, X reported that it had comprehensive measures in place against CSEM. It stated that it had proactively removed 6 pieces of CSEM and reactively removed 29 pieces of CSEM as a result of user reports for content accessible by Singapore users.

c. However, our tests detected considerably more cases of CSEM on X from Singapore during the same period. These cases included nudity (including self-generated nudity), solicitation of sexual services, and distribution of CSEM. X must step up its efforts to detect and remove CSEM on its service.

## Section Aii: User safety measures for children

**03** **X's measures did not effectively restrict children from accessing adult sexual content.**

a. X's annual report stated that it applies additional safety measures to children's accounts, including restricting minors from accessing sensitive content such as adult pornography and excluding sensitive content from search results.

b. However, X did not effectively enforce its own measures to restrict children's accounts from viewing adult sexual content. X's own publicly stated policy states that it restricts "viewers who are under 18, or who do not include a birth date on their profile, from viewing adult content". Our tests found that children's accounts could easily find and access explicit adult sexual content on X, especially hardcore pornography, with simple search terms.

c. X claims in its annual report that its service "is not primarily for children". However, as children above the age of 13 are still permitted to use the service, X must ensure that its safety measures for children are effectively implemented.

**04** In addition, X did not have dedicated tools for children or their parents/guardians to manage children's safety. X only has general safety tools for all users, which children can also use. Children or their parents/guardians are not provided clear warnings of implications if children opt out of their default settings.

## Section B: User reporting and resolution

**05** **X needs to improve the effectiveness of its user reporting systems.** Based on our tests, X took action on 54% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified X directly, X took action on the remaining 46% of harmful content. This suggests that all the content was violative of X's community guidelines and should have been actioned on when initially user reported.

**06** **X needs to improve the timeliness of its user reporting systems.** Based on our tests, X took an average time of 10 days to take action on harmful content that violated its community guidelines when reported by user accounts.

a. Specifically, X took an average of 7 days to take action on Sexual content, 9 days for Suicide and Self-Harm content, and between 10 to 20 days for other categories of harmful content.

b. In contrast, X's annual report stated that its median time for taking action on user reports was 15.06 hours.

## Section C: Accountability

**07** **X provided clear information in its annual online safety report with supporting 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," X provided the data broken down by the six harmful content categories.

b. For paragraph 26(b) of the Code on time it took to take action on user reports, X provided the median time it took to take action on end-user reports in Singapore for all Terms of Service violations, which was 15.06 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 harmful content category as well as how many pieces of content originated from Singapore.

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.

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

## X's Response

***Executive Summary***
*Our mission at X is to promote and protect the public conversation. We believe X users have the right to express their opinions and ideas without fear of censorship. We also believe it is our responsibility to keep users on our platform safe from content that violates our Rules.*

***Section Ai – User safety measures for all end-users***
*Our policies and enforcement principles are grounded in human rights, and we have been taking an extensive and holistic approach towards freedom of expression by investing in developing a broader range of remediations, with a particular focus on education, rehabilitation, and deterrence. These beliefs are the foundation of "Freedom of Speech, not Freedom of Reach" – our enforcement philosophy, which means we restrict the reach of posts, only where appropriate, to make the content less discoverable as an alternative to removal.*

*The X Rules and policies are publicly accessible on our Help Center. To enforce our Rules, we use a combination of machine learning and human review. Our systems either take action automatically, or surface content to human moderators based on user reports and/or proactive detection methods, who then use important context to make decisions about potential violations. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. We also have an appeals process for any potential errors that may occur.*

*X has zero tolerance towards any material that features or promotes Child Sexual Exploitation (CSE). X maintains extensive policies on content or actions that are prohibited and we continue to strengthen our enforcement with more tools and technology to prevent bad actors from distributing, searching for or engaging with CSE content.*

*We are building on the progress made in 2023 and continue our aggressive approach to enforcement. In the first half of 2024, we suspended 2,781,634 accounts for violating our Rules on Child Sexual Exploitation and actioned 14,571 pieces of content.*

***Section Aii – User safety measures for children***
*X, as a service, is not primarily for children, but anyone above the age of 13 can sign up for the service. However, we recognize that users between 13 and 17 years old are a more vulnerable group by virtue of their age. The X Rules enable minors to participate in the public conversation freely and safely. We also have measures in place to make sure minors' experience using the platform is safe and secure.*

*X restricts viewers who are under 18, or who do not include a birth date on their profile, from viewing adult content.*

***Section B – User reporting and resolution***
*X is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive and/or controversial. We strive to create an environment where users feel empowered to express themselves. When abusive behavior occurs, we make it easy for users to report violations of X Rules or local laws.*

*We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our platform and, in particular, promotes counterspeech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.*

*Thus, context matters. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether: the behavior is directed at an individual, group, or protected category of people; the report has been filed by the target of the abuse or a bystander; the user has a history of violating our policies; the severity of the violation; the content may be a topic of legitimate public interest.*

***Section C – Accountability***
*Transparency on X is of extreme importance. X is committed to the open exchange of information. It is now more important than ever that we shine a light on our own practices, including enforcement of the X Rules. The public and policy makers want to be better informed about our actions and we recognize these calls for greater transparency.*

## 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 & resolution | Section C: Accountability |
| ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ | ☑☑☑☑☑ |

### Section Ai: User safety measures for all end-users

**01** **YouTube had the required user safety measures for all users.**

a. YouTube had community guidelines covering the six categories of harmful content listed in the Code as well as human and automated content moderation measures to minimise users' exposure to harmful content.

b. YouTube had a variety of tools for users to manage their own online safety. Users are provided multiple options to restrict visibility of harmful content and/or unwanted comments and limit the visibility of their accounts and content. YouTube does not display the location of its creators in their respective videos and the location of users on any comments, and therefore does not have any tools to limit location sharing.

c. YouTube provided easily accessible and understood online safety information through its YouTube Help Centre, YouTube for Families Help Centre and Creator Safety Centre. YouTube also implemented initiatives to educate and raise awareness of online safety in Singapore, such as the "Safer with Google" event that covers various initiatives focused on enhancing online safety and digital resilience.

d. YouTube actively offered local safety resources to users who searched for high-risk terms. For example, users who searched for "suicide" are provided with a link to Samaritans of Singapore. Similarly, users that search for "domestic violence" are provided the National Anti-Violence helpline. YouTube stated in its annual report that it "recently expanded crisis resource panels into a new full-page experience that better helps viewers pause for a moment. This full-page experience allows viewers to more prominently see resources for the third-party crisis hotlines run by locally based organisations."

e. YouTube reported comprehensive measures to proactively detect and remove CSEM and terrorism content. During the period of assessment, there were no CSEM or terrorism cases on YouTube detected by or reported to IMDA.

## Section Aii: User safety measures for children

**02** **YouTube had all the required user safety measures for children.**

a. YouTube provided differentiated accounts to children with more restrictive default settings based on various age ranges. For example, YouTube Kids offered content settings for different age groups such as "Preschool" for ages 4 and under, "Younger" for ages 5 to 8, and "Older" for ages 9 to 12. YouTube also offered Supervised Experiences, a feature where children can access the main YouTube platform with parental supervision.

b. YouTube provided parents/guardians with tools to manage children's safety via YouTube Kids or Supervised Experiences. These tools allow parents to set content settings for children, and manage recommendations and search results. YouTube Kids is only available for parents/guardians with children under the age of 13, while Supervised Experiences is available for parents/guardians with children under the age of 18 but only if their child's account was created before they turned 13 years of age.

c. For children aged 13 to 17 who use the main YouTube platform, YouTube applies more restrictive default settings such as turning SafeSearch on, implementing content restrictions and turning autoplay off, limiting repeated recommendations of certain content to support the well-being and mental health of young people, and restrictive privacy settings on video uploads.

**03** **YouTube's enforcement of its community guidelines for children could be improved.** Age-inappropriate content that should have been restricted for children based on its own community guidelines, such as videos with sexually suggestive imagery, was found to be accessible via a child's account.

## Section B: User reporting and resolution

**04** **YouTube needs to improve the effectiveness of its user reporting systems.** Based on our tests, YouTube took action on 46% of harmful content that violated its own community guidelines when reported by user accounts. Subsequently, when IMDA notified YouTube directly, YouTube took action on the remaining 54% of harmful content. This suggests that all the content was violative of YouTube's community guidelines and should have been actioned on when initially user reported.

a. YouTube was unable to provide data on the number and types of harmful and age-inappropriate content removed as a result of Singapore user reports. YouTube was only able to provide data on the "Number of videos uploaded from a Singapore IP address that were removed for CG (community guidelines) violations, where the source of first detection was a user flag". YouTube stated that during the reporting period 9,476 such videos were removed. YouTube also offered to provide alternative data from its Legal Complaints webform, with which users can submit alleged violations based on country-specific laws or regulations such as Singapore's Online Safety Code. However, YouTube did not receive any relevant complaints during the period of review and thus had no useful data to provide.

**05** **YouTube needs to improve the timeliness of its user reporting systems.** Based on our tests, YouTube took an average time of 5 days to take action on harmful content that violated its community guidelines when reported by user accounts.

a. Specifically, YouTube took an average time of 3 days to take action on Content facilitating vice and organised crime and between 10 to 17 days for other categories of harmful content.

b. YouTube was unable to provide data on the time it took to take action on user reports as required by the Code for the same reasons as in paragraph 4a above.

## Section C: Accountability

**06** **YouTube provided clear information in its annual online safety report, although there is room for improvement in its supporting data.**

a. YouTube was not able to provide data to demonstrate whether YouTube assesses and takes the appropriate action on user reports in a timely manner. Examples of such data which YouTube should have provided include:
   i. Data on the number and types of harmful and inappropriate content removed as a result of user reports, pursuant to paragraph 26(a) of the Code; and
   ii. Data for the time it took to take action on user reports, pursuant to paragraph 26(b) of the Code.

b. For paragraph 26(c) of the Code on "the number and types of harmful or inappropriate content proactively removed by the Service," YouTube's annual report stated that it removed 24,405,862 videos globally where the source of first detection was automated flagging and 1,978,602 videos uploaded from a Singapore IP address where the source of first detection was automated flagging. However, it did not provide a breakdown by the harmful content categories in its community guidelines.

c. For paragraph 26(d) of the Code on "the number of accounts suspended or banned in Singapore," YouTube stated in its annual report that it removed 2,800,719 channels globally and 20,739 Singapore channels for violating YouTube's community guidelines, and provided a breakdown by the categories in its community guidelines. "Nudity or Sexual", "Misinformation" and "Child Safety" were the top three reasons why channels were removed globally and "Harmful or Dangerous", "Nudity or Sexual" and "Misinformation" were the top three reasons why Singapore channels were removed.

**07** 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 protect 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