Singapore Online Safety Report 2025

Introduction

This Online Safety Report (**Broadcasting Act 1994, Code Of Practice For Online Safety**) outlines our overall approach to online safety on X as well as the measures and activities to combat harmful and inappropriate content for the period of 1 April 2024 to 31 March 2025 for Singapore.

Our Approach to Online Safety

X's purpose is to serve the public conversation. In line with our mission to promote open conversation, we encourage a variety of perspectives on the platform. This is central to our *Freedom of Speech, Not Reach* philosophy, that moves us away from a binary take down/leave up approach to a more reasonable, proportionate and effective moderation process. Violence, harassment, and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation. We thus have clear Rules and Policies in place that are designed to ensure all people can participate in the public conversation freely and safely. They apply globally, including to Singapore, and are easily accessible on our Help Center.

Our Rules and Policies are dynamic, and we continually review them to ensure that they are up-to-date, necessary and proportional. Creating a new policy or making a policy change requires in-depth research around trends in online behaviour, developing clear external language that sets expectations around what's allowed, and creating enforcement guidance for reviewers that can be scaled across millions of pieces of content and accounts. We are also committed to providing transparency on our policy development process and rules enforcement philosophy, and provide explanations of these on our Help Center.

While we aim to enable open discussion of differing opinions and viewpoints, we are committed to the objective, timely, and consistent enforcement of the X Rules. To enforce the X Rules, we use a combination of machine learning and human review.

Our content moderation systems are designed and tailored to mitigate potential harms without unnecessarily restricting the use of our platform and fundamental rights, especially freedom of expression. Content moderation activities are implemented and anchored on principled policies and leverage a diverse set of interventions to ensure that our actions are reasonable, proportionate and effective. Our content moderation systems blend automated and human review paired with a robust appeals system that enables our users to quickly raise potential moderation anomalies or mistakes. This work is led by an international, cross-functional team with 24-hour coverage and the ability to cover multiple languages. These moderation activities are supplemented by scaled human investigations into the tactics, techniques and procedures that bad actors use to circumvent our rules and policies.

X strives to provide an environment where people can feel free to express themselves. If abusive behaviour happens, we want to make it easy for people to report it to us. When we take enforcement actions, we may do so either on a specific piece of content (e.g., an individual post or Direct Message) or on an account. We may employ a combination of these options. In most cases, this is because the behaviour violates the X Rules.

We always aim to exercise moderation with transparency. Where our systems or teams take action against content or an account as a result of violating our Rules or in response to a valid and properly scoped request from an authorised entity in a given country, we strive to provide context to users.

X is a place for users to share ideas and information, connect with communities, and see the world around them. In order to protect the very best parts of that experience, we provide tools designed to help users control what they see and what others can see about them, so that they can express themselves on X with confidence. Our diverse product-level safety features allow users to modify their experience and engagement on X to ensure each user is able to participate on the platform in a safe and meaningful way.

We recognise that minors are a more vulnerable group by virtue of their age. X, as a service, is not primarily for children, and those below the age of 13 are not permitted to sign up for the service. Users who do not meet our age requirements have their account locked. Parents and guardians are able to access our Rules and Policies to learn more about how to keep their child's account and experience on X safe, secure and welcoming.

2024 - 2025 Annual Online Safety Report to be submitted to IMDA under the Code of Practice for Online Safety

Guidelines

- **Please ensure that the report contains all the information requested in this template and is relevant to the reporting period.**
- **DSMSs are encouraged to follow the flow and format of this template when preparing their reports.**
- **Please include Singapore-relevant data and information where possible.**
- **Please ensure that all information is self-contained within the report. E.g. do not include hyperlinks to information outside of the report.**
- **Please try to map the content categories in your community guidelines to the harmful/inappropriate content categories under the Code of Practice for Online Safety.**
- **The annual online safety report must be submitted to OCO@imda.gov.sg by no later than 30 June 2025.**

Code Obligations:	**For DSMS's input**
SECTION A: USER SAFETY	
Paragraph 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: ● Minimize 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. *And* Paragraph 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.	**Please provide information on the measures in place for all End-users in Singapore:** X has implemented a robust and multi-layered framework to enable Singapore end-users to use the service in a safe manner, including by putting in place measures to minimise Singapore end-users' exposure to harmful content, to empower end users to manage their safety on the service, and to mitigate the impact on end-users that may arise from the propagation of harmful content. By integrating stringent content and service use policies (namely our Terms of Service and the X Rules), advanced automated detection, proactive enforcement, user reporting mechanisms, recommender system controls, and other user tools and resources, X ensures that harmful content is either prohibited or its reach is significantly restricted, and that users are empowered to take proactive steps to secure their safety on the service. The measures put in place by X are designed to create an environment where users can safely participate in the public conversation and benefit from a healthy experience on the platform, with their risk of encountering harmful content systematically reduced through both sophisticated technological systems and human oversight. **MEASURES RELEVANT TO ALL TYPES OF HARMFUL CONTENT** ***X Rules*** Each type of harmful content under the Singapore Online Safety Act is addressed by the X Rules. The specific X Rules that apply to each type of harmful content are set out in the sections below per type of harmful content. The effect of the relevant X Rules is that the majority of harmful content is expressly prohibited on the service, while the remainder is governed by policies that restrict its reach or otherwise ensure adequate protection from potential harm for users (e.g. labelling to prevent unwanted exposure to adult content and graphic media). ***User reporting mechanisms*** ● *Accessible reporting*: Users can easily report content that they encounter which they believe to be in violation of the X Rules. The reporting functionality is available in-app alongside individual items of content, user profiles, DMs, Spaces, Livestreams and Lists, which enables a user to directly report potential harmful content when they encounter it. Users can also report harmful content through the Help Centre reporting form. The reporting tools are intuitive and enable users to submit reports that are sufficiently precise. The Help Centre contains clear step-by-step instructions of how to report violations of the X Rules through all of the relevant reporting channels as iterated above. When a user reports an item of content, they will be notified that X has received the report and the content will be obscured from them behind a notice. ● *Timely action*: Once reported, X prioritises the review of content that presents the highest risk of harm to users. This ensures that harmful content which is most likely to severely negatively impact end-users is swiftly actioned when reported. The relevant content moderation team will enforce against violations in accordance with the X Rules. ● *Trusted partners*: X maintains partnerships with trusted organisations who are able to notify X of potential violative content on the platform. This includes SG Her Empowerment and AWARE Singapore, and Samaritans of Singapore for Self-Harm.

- *Training and guidance*: Operational guidance specifically related to handling user reports under the X Rules is provided to X's content moderators through regularly updated training materials. This includes guidance about escalating certain reports where appropriate to specific expert content moderator teams.

Key improvement: In the reporting period, X implemented an important enhancement to its DM reporting flow to enable users to more clearly and intuitively report certain child safety issues that may arise specifically in DMs. While users were previously able to report violations of X's Child Safety policy in DMs, X has introduced new reporting pathways including '*I am a minor and someone is engaging in inappropriate conversation*' and '*Someone is sharing or selling sexually explicit content involving a minor*'. This improvement has ensured that reports are immediately triaged to the correct internal teams and prioritised for swift review, investigation and enforcement.

Law enforcement reporting mechanism

X also provides a dedicated law enforcement reporting form through which applicable law enforcement agencies in Singapore can report content under applicable local laws.

Proactive detection and enforcement

X has implemented proactive automated detection systems under each of the X Rules policies relevant to harmful content under the Singapore Code of Practice for Online Safety (the "**Code**"). These systems vary between policies and include machine learning and heuristics-based models. X uses these proactive automated systems for automated enforcement in some cases (as described below) and to enqueue potentially violative content for human review.

Where X's automated detection systems meet a sufficient threshold, they are used to apply automated enforcement. This is generally applied in the case of egregious violations where the immediacy of enforcement is integral to preventing harm (e.g. CSEM). See below for further information regarding the proactive enforcement systems used in relation to specific types of harmful content, including CSEM.

Enforcement actions

X takes a graduated approach to enforcement such that the severity of enforcement depends on the severity and likelihood of potential harm. Generally, this means that harmful content is mostly prohibited on the service, and violations are enforced with content removal or account suspension. Where potential harm associated with certain content is less likely and less severe, X will restrict its reach rather than remove it entirely from the platform, to ensure adequate protection for freedom of expression and access to information. When X restricts the reach of a post, it will be excluded from search results and trending, notifications, and home timelines and the ability of users to engage with it will be restricted (i.e. users will be unable to like, share, or reply).

The specific enforcement actions applied under the X Rules policies relevant to each type of harmful content are detailed below under the specific sections per type of harmful content.

Recommender system controls

Every day, people come to X to keep up with what's happening. The choices they make when using X, such as the accounts and Topics they follow and the Lists they create or join, help determine the content they will see. For example, when a user follows an account or Topic, associated posts will appear in their Home timeline alongside others X thinks that the user may be interested in.

Recommendations may amplify content, so it's important that they are surfaced responsibly. While X's enforcement philosophy empowers people to understand different sides of an issue by allowing many forms of speech to exist on the platform, X also works hard to prevent the amplification of harmful content on the platform.

X has several ways of preventing potentially harmful content and accounts from being amplified, including using machine learning technology, reviewing user reports, and other tools.

- *X Rules enforcement*: At a baseline level, the X Rules and their enforcement ensure that Harmful Content under the Code (which is prohibited or restricted by the X Rules as explained above) is removed from the platform or restricted from being encountered by other users. Where X enforces certain content with 'restricted reach', the content will be excluded from search results, trends, and recommended notifications, and home timelines. Discoverability is restricted to the author's profile.
- *Eligibility requirements*: X implements specific eligibility criteria to ensure that appropriate content is surfaced in the "For You" and "Following" tabs. Sensitive content such as content labelled under the 'Adult Content' policy and graphic media labelled under the 'Violent Content' policy will not be accessible to U18 accounts, logged-out accounts, users without a birth date on their profile. Adult users are required to opt-in to see this content, and can do so by switching-on their 'display sensitive media' setting; while switched-off, sensitive media will be hidden behind an interstitial and not recommended to adult users. Further recommender system eligibility restrictions include:
    - Content that includes information known to be obtained through a hack and does not come from websites (specific articles or whole domains) that comment on or distribute materials in the course of some additional purpose, such as journalist coverage or commentary.
    - Content that violates any of the X Rules, but has been left on the platform due to the public-interest exception.
    - Content that promotes the use of regulated substances or weapons.
    - Content that is deemed marginally abusive and is ineligible for amplification under our safety policies, including our Abusive Behavior and Hateful Conduct policies.

- Harmful misleading information, including violations of the Civic integrity policy, Synthetic and manipulated media policy, and any other misleading information policies listed at help.twitter.com/misinfo.
- Content that automated systems have determined may violate the X Rules, but that has not yet been reviewed by a human and/or may have been identified in error.

Additional controls

- *Sensitive media labelling*: Under the Adult Content and Violent Content policies, users are required to label adult content and graphic content that they post. Users can do this on a one-time occasion by selecting a content warning while uploading an individual post, or by adjusting their media settings in the case that they regularly share sensitive media (i.e. such users can switch-on the following setting; '*Mark media you post as containing material that may be sensitive*'). If users fail to label sensitive media that they post, users can report this to X and X may take action to label that content itself. Repeat violations may result in further enforcement action. X also takes steps to proactively detect and label sensitive media that has not been labelled in accordance with the X Rules. When content has been labelled as sensitive media, it will not be accessible to U18 users (and logged out users and users without a birthdate on their profile), and will be obscured behind an interstitial and not recommended to adult users who have not opted-in to see sensitive media by switching-on the 'display sensitive media' setting on their profile. In any case, it is prohibited to share sensitive media in prominent or otherwise highly visible surfaces such as user profiles or banner photos.
- *User controls ('Not interested in this post')*: X provides users with the opportunity to indicate that they are 'Not interested' in a specific Post. When users select this option, the relevant post will be removed from their feed and X will prevent similar content from being recommended to the user in the future. This functionality enables users to provide dynamic and active feedback to shape their For You and Following feeds tabs, and help ensure that appropriate content based on their own preferences is recommended to them.
- *User controls ('Content you see')*: In their account settings, users have the option to select 'topics' and 'interests' of relevance to them. This feedback shapes the content that is recommended to them. Users can also see a record (and remove accordingly) of the topics that they have indicated through the 'Not interested' functionality to not be of interest to them. Further, users can alter their search settings, so that sensitive content is hidden from their search results.
- *User controls ('Mute')*: X provides a user control that enables users to mute accounts, words and notifications. When a user mutes a word, they will not see posts with those words in their "For You" or "Following" timelines, and will not receive any new notifications for posts that include such words. Further, users can mute accounts, with the effect that posts from such accounts will not appear in their "For You" or "Following" timelines. Further aspects of the Muting functionality are explained below in relation to user empowerment.
- *User controls ('Block')*: X provides a user control that enables users to actively block accounts. When a user blocks an account, posts from blocked accounts will not appear in their "For You" or "Following" tabs. Further aspects of the Blocking functionality are explained below in relation to user empowerment.

Advertising and monetisation safeguards

- *X Ads policies*: X has specific policies that apply to monetisation on X and X's paid advertising products. These policies prohibit the monetisation or promotion of adult sexual content, the promotion of drugs and drug paraphernalia, the promotion of hateful content, the promotion of weapons and weapons accessories, and the promotion of fraudulent or deceptive content, among other types of content. Additionally, X places further restrictions on advertising content by prohibiting ads that contain inappropriate content (e.g. offensive content such as content that is inflammatory or provocative, content that is inappropriate for the general audience, harmful weight loss content, and sensitive content such as content that sells goods/services while referencing a sensitive event). Finally, X prohibits knowingly marketing or advertising certain products and services to minors (e.g. alcohol, weight loss products, health and wellness supplements, etc) and X's paid advertising products and advertisements containing age-inappropriate content will be tagged as "not family safe", and restricted from being shown to users under the age of 21 and logged-out users. When advertisers on X choose to promote their content with X Ads, their account and content become subject to an approval process. The approval process is designed to support the quality and safety of the X Ads platform. This process helps X check that advertisers are complying with X's advertising policies.
- *X Rules*: In addition to the X Ads policies, advertisers must follow the X Rules. Ads can be reported for violating the X Rules, and X will review reports and enforce against violations. X also takes steps to proactively enforce against Ads that violate the X Rules, including through the X Ads approval process.
- *Creator Monetisation Standards*: X provides monetisation products, whereby creators can earn money or currency from X. There are two monetisation products on X: (i) Subscriptions, and (ii) Creator Revenue Sharing. Through Subscriptions, users can get paid a share of the revenue X earns through subscriptions (i.e. payments by users to subscribe to creators) by offering their most engaged followers an extra level of access and bonus content. Through Creator Revenue Sharing, eligible creators can earn money by creating content on X, based on verified engagements with their posts. Both of these monetisation products are governed by the X Creator Monetisation Standards, which include eligibility and content and conduct standards. The eligibility standards require such users to be 18+, reside in countries where monetisation is available, meet certain integrity and authenticity requirements, and be in good-standing with X, among other requirements. All content monetised on X must comply with the X Rules, as like any other content on X, additional conduct standards, and additional content standards (e.g. Adult Content, graphic, objectionable and violent content, sensitive content, and strong language may be ineligible for monetisation). The effect of the Creator Monetisation Standards is to ensure that X's monetisation products cannot be used to amplify harmful content or otherwise increase the likelihood that users encounter harmful content.

User tools and resources

X equips Singapore end-users with a comprehensive suite of tools and resources to actively manage their safety. By providing granular controls over content visibility, interactions, and privacy, X enables users to shape their experience on the platform in alignment with each user's unique preferences and safety needs. X's user empowerment tools - supported by educational resources provided by X on its Help Centre - ensure that Singapore end-users can confidently navigate the platform while minimising their exposure to harmful content and unwanted interactions, thereby empowering them to manage their own safety as required by the Code. These are as follows:

Content control tools

- *Unfollow*: Users can unfollow an account to stop seeing that account's posts in their 'Following' tab.
- *Mute account*: Users can mute an account, which will prevent any content from that account appearing in their 'Following' or 'For You' timelines.
- *Mute words*: Users can mute a word. Posts with muted words will not appear in that user's 'For You' or 'Following' timelines, and they will not receive any new notifications for posts that include such words.
- *Filter notifications*: Users have three options located in their notifications settings to filter the notifications that they receive.
    - *Quality filter*: When turned on, this setting filters lower-quality content from a user's notifications (for example, duplicate posts or content that appears to be automated). It does not filter notifications from people that the user follows or accounts that they have recently interacted with. For users who are new to X or who have re-installed the app, the quality filter setting will be turned on by default.
    - *Mute notifications*: When a user mutes a word or phrase through the 'mute words' setting, posts containing these words will not appear in their notification tab (from users that they do not follow). Users can also mute notifications for accounts that they would like to avoid seeing notifications from, by unfollowing and muting an account.
    - *Advanced filters*: Users can disable notifications from certain types of accounts. In addition, if a user receives a lot of sudden attention, X may insert a notification in their notifications tab inviting them to adjust their filters to give them more control over what they see. Users can choose to disable notifications from (i) accounts that are new (that the user does not follow), (ii) accounts that don't follow the user (and which the user does not follow), (iii) accounts that the user does not follow, (iv) accounts with a default profile photo that the user does not follow, (v) accounts without a confirmed email address (that the user does not follow), and/or accounts without a confirmed phone number (that the user does not follow).
- *Not interested in this post*: X provides users with the opportunity to indicate that they are 'Not interested' in a specific Post. When users select this option, the relevant post will be removed from their feed and X will prevent similar content from being recommended to the user in the future. This functionality enables users to provide dynamic and active feedback to shape their For You and Following feeds tabs, and help ensure that appropriate content based on their own preferences is recommended to them.
- *Report*: Users can report posts, Lists, and Direct Messages that are in violation of the X Rules or the X Terms of Service. When a user reports content, it will be replaced with a notice stating that the user has reported it. Reported messages and conversations will disappear from a user's inbox and cannot be recovered.
- *Safe Search*: X gives users control over what they see in their search results through the 'Safe Search' mode. This mode applies filters that will exclude potentially sensitive content, along with accounts that the user has muted or blocked, from their search results. Users can turn this setting off, or back on, at any time.

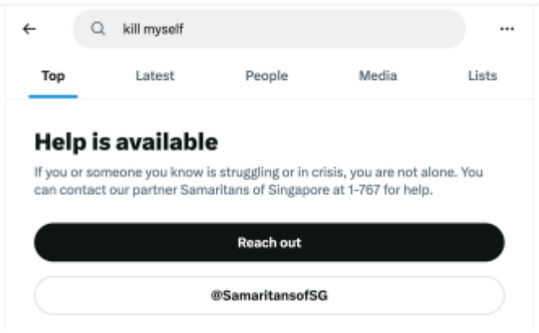
Privacy and interaction controls

- *Protected Posts*: Users can switch on 'protected posts' in their settings. This means that they will receive a request when any new accounts want to follow them, which can be approved or denied. When users switch on 'protected posts', their posts (including permanent links to their posts) will only be visible to their followers, and their followers will not be able to repost or 'repost with comment' their posts. Their posts will also not appear in third-party search engines (like Google search) and will only be searchable on X by the user and that user's followers. Further, only their followers will be able to see their replies to other users' posts. When a user turns on or turns *off* protected posts, this will apply retrospectively to all posts previously made by the user and prospectively, as long as Protected Posts remain switched off. Accounts belonging to known monitors are defaulted to 'Protected Posts'.
- *Block*: X provides a blocking functionality that enables users to control how they interact with other accounts. The functionality enables users to restrict specific accounts from following, direct messaging, and engaging with them. If a user's posts are set to public, the account they have blocked will still be able to see their posts,but they will not be able to engage with it. This ensures that the blocking function cannot be abused to disseminate harmful content about a blocked account holder. If a user does not want a blocked account to be able to see their content, they can set their account to 'Protected Posts'.
- *DM requests*: Users will always be able to receive messages from accounts that they follow. However, the DM settings menu allows users to select whether and which types of accounts that they do not follow can send them a 'message request'. The options are to allow requests from (i) 'no one', (ii) 'verified users', or (iii) 'everyone'. In any case, message requests always land in the 'request' inbox which is separate from the regular DM inbox. If a user wants to prohibit an account from messaging them entirely, they can block the relevant account.
- *DM low-quality filter*: In their DM settings menu, users can switch on the option to 'Filter low-quality messages'. Once enabled, this will hide message requests that have been detected as potentially being spam or of low quality. Such messages will be sent to a separate inbox at the bottom of the user's message requests.
- *DM graphic media filter*: By default, X will display a warning over graphic media detected in DMs, for both DMs sent from accounts that the user does and does not follow. Additionally, if a message with graphic content comes from an account that the user does not follow, X will treat it as spam and move it to the bottom of the user's 'request' inbox. Adult users have the possibility to turn off this filter in their settings if they wish to; once turned off, X will still display a warning over graphic media in DMs received from accounts that the user does not follow.
- *Disabling DM 'read' receipts*: DMs feature read receipts so users know when people have seen their messages. Users can turn off the 'show read receipts' setting, and this means that no one will be able to see when they have read a message, and they will not be able to see when another user has read their message.
- *Suspicious content warning in DMs*: X will hide messages that may be suspicious or contain suspicious URLs behind a warning label (e.g. spam). This is aside from X's other methods to proactively detect and block child sexual exploitation and abuse ("CSEA") URLs from being uploaded or shared on the platform.
- *Reporting*: As above, Singapore end-users can report violations of the X Rules. This includes violative interactions, such as grooming. The report functionality is available in DMs, with specific unique reporting pathways for grooming conduct and attempted CSEA solicitation. X will review user reports and take appropriate enforcement action against violations in line with its enforcement policies.

- *Controlling replies*: Users can choose who will be able to reply to their posts. The default position for public accounts is that anyone can reply, but options are available to turn off all replies or only allow accounts mentioned in the post to reply. A user can also change who can reply to their posts, or turn off replies, after they have published a post.
- *Hidden replies*: Users can hide replies to their posts. This helps users to have more control over the conversations on their posts and reduces the visibility of potentially harmful content.
- *Protected videos*: Users can switch on the 'Protect your videos' option in their settings, which means that videos in their posts or DMs will *not* be downloadable by default. Once enabled, this setting will apply to posts prospectively, but will not apply retroactively to videos already posted or shared in DMs. Users can also disable downloads for specific posts.
- *Photo tagging*: Users have a number of options with respect to photo tagging. In their privacy settings they can limit tagging such that only accounts that they follow can tag them. Alternatively, they can switch photo tagging off entirely such that no account will be able to tag them in a photo. When tagged, a user will receive a notification.
- *Discoverability settings*: In order to help users make connections with people on X that they already know, X may use the email address and phone number associated with their account to make their account discoverable to others (e.g. if someone allows X to suggest accounts to follow based on contacts held on their device, X may suggest accounts that have a phone number or email address from that person's contacts). However, users can control whether their account will be discoverable to others based on their phone number or email address by adjusting their discoverability settings, with this setting switched *off* by default. In any case, a user's email address and phone number will never be publicly displayed on X.
- *Location information*: X offers users a range of location-related settings to enhance and personalise their experience on the platform. By default, location tagging on posts is turned off, but users can opt in to attach a general location label—such as a city or neighborhood—or enable precise location, which uses device GPS data. Once enabled, location labels can be added manually to individual posts, and users retain full control over whether or not to include this information on a case-by-case basis. Additionally, X may personalise content based on users' current location, signup location, and, if the user consents, other places they've been. Users can manage their location history, delete past location data, and review which locations have been used through the platform's settings.

Safety information and resources

- *Help Centre*: X's Help Centre is available to end-users in Singapore. It sets out how to use the platform safely, and explains the X Rules.
- User safety information on policy changes and feature updates are regularly shared through the X Safety account.
- *High-risk search terms*: If users search for high risk terms or phrases, X may display crisis resources or link users to third party support resources. For example, if Singapore users input search terms for self-harm or suicide they will see the following pop-up:



- *User education*: Education is integral to X's enforcement philosophy. X achieves this by informing the violator that a X Rule was broken, and requiring them to remove the violating content themselves before they can post again (in the meantime, it is placed behind a 'tombstone' such that it is not visible to other users). By requiring users to acknowledge the violation and remove the relevant content themselves, X guides users towards a better understanding of the X Rules and how to practice appropriate behaviour on X. In less severe cases, where X restricts the reach of a post (rather than requiring the user to remove it), the user will also be presented with information about which X Rule has been violated.
- *Community Notes*: The Community Notes feature aims to create a better informed public conversation by empowering people on X to collaboratively add context to potentially misleading posts. Contributors can leave notes on any post and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a post. Community Notes are never authored by X and cannot be edited or modified by X. They are subject to the X Rules, and X will take enforcement action if a relevant violation is determined. External researchers found that users repost a post 61% less after it receives a Community Note[1], while another study found a 50% reduction in reposts and an 80% increase in post deletions after a post received a Community Note[2].
- *Global transparency reporting*: X believes that it is important to shine a light on its own practices, including enforcement of the X Rules. X hosts a comprehensive Transparency Centre, which features a bi-annual global transparency report. Users and policy makers can access meaningful metrics on the actions X takes on content and accounts that violate the X Rules, and the various legal requests that X receives from governments and law enforcement authorities.
- *Notifying users of enforcement actions*: When X takes enforcement action (whether at the individual content or account level), it will normally notify the affected account owner and provide them with the opportunity to appeal X's decision. X will not notify the relevant actor when service security and user safety necessitates so.

[1] https://osf.io/preprints/osf/3a4fe_v1
[2] https://arxiv.org/abs/2404.02803

Through the extensive array of content, privacy, and interaction controls detailed above, coupled with accessible safety resources and transparent user education, X empowers Singapore end-users to effectively manage their safety on X. The platform's user empowerment tools enable users to take proactive steps to reduce their risk of encountering harmful content and unwanted interactions, with the ability to meaningfully shape their experience on X according to their unique preference and needs.

In the sections below, X has provided information regarding the specific measures implemented in relation to each type of harmful content.

SEXUAL CONTENT

This type of harmful content is primarily restricted through the application of the X Rules policies of 'Child Sexual Exploitation', 'Non-Consensual Nudity' and 'Adult Content'. The 'Violent Content' and 'Abuse and Harassment' policies are also relevant to X's mitigations in this area.

X Rules

- X has zero tolerance for any form of child sexual exploitation and eliminating it is one of the company's key goals. X's robust 'Child Sexual Exploitation' policy ensures that ***any*** content that depicts or promotes child sexual exploitation (including real media, text, illustrated, or computer-generated media (including generative AI media)) is strictly prohibited. This includes, but is not limited to, directly or indirectly sharing, requesting, linking to, engaging with (such as liking or reposting, or following accounts that distribute), expressing a desire for, or describing child sexual exploitation and any conduct associated with child sexual exploitation. Under this policy, X also prohibits sending sexually explicit media to a child, or engaging or trying to engage a child in a sexually explicit conversation.
- Under the 'Non-Consensual Nudity' policy, users are prohibited from posting or sharing explicit images or videos that were taken, or appear to have been taken or shared without the consent of the people involved.
- Under the 'Adult Content' policy, users are permitted to share consensually produced and distributed adult nudity or sexual behaviour, provided it is properly labelled and not prominently displayed. Users who post adult content can label it as 'nudity' or 'sensitive' at the level of an individual post. Alternatively, if a user intends to regularly post sensitive media (including adult content or graphic media), they are required to adjust their media settings to select the option; *'Mark media you post as containing material that may be sensitive*'. When Adult Content is labelled, it will not be accessible to minors (or to logged out users or users without a birthdate on their profile), and it will be obscured behind an interstitial and not recommended to adult users who have not opted-in to see sensitive media by switching-on the '*display sensitive media*' setting on their profile. The overall effect of this policy is that it limits the display of compliant pornographic content to adults who actively want to engage with that content.
- Under the 'Violent Content' policy, violent sexual conduct is prohibited on the platform. This includes media that depicts violence, whether real or simulated, in association with sexual acts, including rape and other forms of violent sexual assault. Bestiality and necrophilia are also prohibited under this policy.
- Under the 'Abuse and Harassment' policy, unwanted sexual content and graphic objectification is prohibited on the platform.

Enforcement

- *'Child Sexual Exploitation'*: Users who violate this policy will be immediately and permanently suspended and the content would be reported to the National Centre for Missing and Exploited Children ("NCMEC") when applicable. In a limited minority of cases, where appropriate in line with X's zero tolerance approach to child sexual exploitation, X may determine that the appropriate enforcement action is to remove the post rather than suspend the account (for example, when shared with the intent to bring awareness or justice, or to express outrage or sharing content in a humoristic context). For the avoidance of doubt, the content would be reported to NCMEC when applicable, even if the user is allowed to regain access to their account after content removal. Subsequent further violations may lead to account suspension. This is further addressed below.
- *'Non-Consensual Nudity'*: X will immediately and permanently suspend any account that it identifies as the original poster of non-consensual intimate media, and any account that it identifies to have been created exclusively for that purpose. In other very limited cases, X may not suspend an account, however X will remove the associated content.
- *'Adult Content'*: Users can report unlabelled adult content, and X also maintains automated systems to proactively detect unlabelled content. Unlabeled Adult Content may be labelled by X, and X may adjust an account's settings so that any content posted by that account in the future will be appropriately labelled. Users who repeatedly fail to appropriately label adult content may have their account placed in read-only mode or suspended.

FOCUS: CHILD SEXUAL EXPLOITATION MATERIAL ("CSEM")

X maintains a zero-tolerance approach to CSEM under the 'Child Sexual Exploitation' policy, strictly prohibiting *any* content that depicts or promotes child sexual exploitation, including real media, text, illustrated, or computer-generated media (including generative AI media). Whenever there is evidence that a minor could be at potential risk, including when a case involves grooming, sextortion, or any other situation that indicates imminent risk of sexual abuse, X will take action and report relevant accounts to NCMEC. This includes, but is not limited to, directly or indirectly sharing, requesting, linking to, engaging with (such as liking or reposting, or following accounts that distribute), expressing a desire for, or describing:

- Depictions of a child engaging in sexually explicit or suggestive acts, which may include edits of such media, such as obscuring faces or body parts;

- Sexualised commentaries about or directed at a known (or unknown) minor;
- Links to third-party sites that host child sexual exploitation materials of any kind; and
- Fantasies about or promoting engagement in child sexual exploitation.

X prohibits:
- Identifying alleged victims of childhood sexual exploitation by name or image;
- Promoting or normalising sexual attraction to minors as a form of identity or sexual orientation;
- Sending sexually explicit media to a child, or engaging or trying to engage a child in a sexually explicit conversation;
- Trying to obtain sexually explicit media from a child, or trying to engage a child in sexual activity through blackmail or other incentives;
- Recruiting, advertising or expressing an interest in a commercial sex act involving a child, or in harboring and/or transporting a child for sexual purposes; and
- Threatening to share, or requesting that others share, sexual media of minors, including self-generated sexual media of minors, particularly when engaging in blackmail or bounty tactics.

Proactive detection and enforcement

To enforce these rules, X employs rigorous and advanced technical systems to proactively and aggressively detect, disrupt, and remove CSEM content and conduct, and the overwhelming majority of CSEM detected and enforced against is identified and removed proactively before a user report is received.. These systems include (a) automated enforcement against known CSEM using hash-matching, (b) automated enforcement with text-based detection tools, (c) automated blocking of known CSEM URLs, and (d) automated detection for human review of potential novel CSEM. During the reporting period (1 April 2024 - 31 March 2025), and as demonstrated by the data reported below, the vast majority of all enforcement actions under the Child Sexual Exploitation policy were taken proactively, illustrating the scale, rigor, and effectiveness of these systems. Proactive manual sweeps by moderation teams, triggered by on- or off-platform trends, are also used to manually recall and review posts with relevant keywords or URLs.

To enhance proactive CSEM detection, X continuously monitors tool effectiveness, addressing deficiencies to maximise recall. Automated enforcement systems undergo rigorous testing and human review before launch. Thereafter, automated enforcement systems are dynamically monitored and continuously enhanced.

During the reporting period, X implemented a number of improvements to its overall CSEM detection and enforcement systems, including training additional agents for proactive keyword and media sweeps. X used signals from Project Lantern - the cross-platform signal-sharing program launched by the Tech Coalition to combat online child sexual exploitation and abuse - for account investigations. As previously detailed, an improved DM reporting flow with intuitive child safety pathways was launched to enhance reactive enforcement and proactive investigations via clearer signals. X also maintains its hash-matching system to register new hashes for scalable automated enforcement. X continues to invest in proprietary technology to reduce the burden on users to report CSEM for enforcement, as demonstrated by the data reported below, the vast majority of enforcement actions taken under the Child Sexual Exploitation policy were enforced proactively. Further, X continuously works to detect evolving trends in CSEM and related activity to enhance its proactive detection and enforcement.

SUICIDE AND SELF-HARM CONTENT

Under the ‘Suicide & Self Harm’ policy, X recognises that suicide and self-harm are significant social and public health challenges that require collaboration between stakeholders - public, private, and civil society - and that X has a role and responsibility to help people access and receive support when they need it.

When developing its ‘Suicide & Self Harm’ policy, X consulted extensively with experts to ensure that people who have engaged in self-harm or experienced suicidal thoughts can share their personal experiences. X also recognised the need to protect people from the potential harm caused by exposure to content that could promote or encourage self-harm - intentionally or inadvertently.

Accordingly, X’s ‘Suicide & Self-Harm’ policy prohibits content that promotes or encourages self-harming behaviours and provides support to those undergoing experiences with self-harm or suicidal thoughts. X defines promotion and encouragement to include statements such as “the most effective”, “the easiest”, “the best”, “the most successful”, “you should”, “why don’t you”. Violations of the ‘Suicide & Self-Harm’ policy include, but are not limited to, encouraging someone to physically harm or kill themselves, asking others for encouragement to engage in self-harm or suicide, including seeking partners for group suicides or suicide games, and sharing information, strategies, methods or instructions that would assist people to engage in self-harm and suicide.

X does allow users to share personal stories and experiences related to self-harm or suicide, so long as they avoid sharing detailed information about specific strategies or methods related to self-harm, as this could inadvertently encourage self-harm behaviour. Additionally, X allows users to share coping mechanisms and resources for addressing self-harm or suicidal thoughts, and it allows discussions that are focused on research, advocacy, and education related to self-harm or suicide prevention.

Proactive detection and enforcement

X uses combinations of natural language processing models, image processing models and other machine learning methods to proactively detect potentially violative content or behavior under the Suicide and Self-Harm policy. Automations are monitored dynamically for ongoing performance and health. X also maintains feedback loops between frontline content moderations, escalations and policy teams

to ensure that its proactive detection and enforcement systems keep pace with evolving behaviours on the platform.

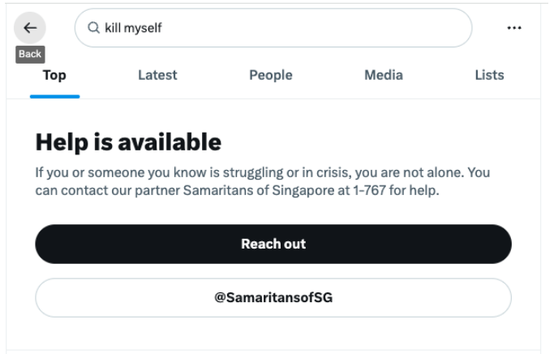
X may also take steps to prevent the spread of instructional material hosted on third-party websites by marking such links as unsafe.

Enforcement actions

X's enforcement approach depends on the type of content being shared, whether or not the reported account is encouraging or promoting self-harm or suicide, and the account's previous history of violations. If a user violates this policy by sharing content that intentionally encourages others to self-harm or suicide, asks others to encourage them to self-harm or suicide, or shares detailed information or instructions related to self-harm or suicide methods, X will require the user to remove the content. X will also temporarily lock the user out of their account before they can post again. If a user continues to violate this policy, or if their account is dedicated to promoting or encouraging self-harm or suicide, their account will be permanently suspended. If cases include images or videos related to self-harm or suicide, they will also be evaluated under the violent media policy.

Additional user support measures

A user attempts to search for suicide or self-harm related content, X will display and link to relevant crisis support resources. This includes the following Singapore-specific resource that is displayed to users in Singapore:



<u>CONTENT ENDANGERING PUBLIC HEALTH & CONTENT ENCOURAGING VICE AND ORGANISED CRIME</u>

This content is prohibited by the X Rules policies of 'Illegal and Regulated Behaviours', 'Scams' and 'Synthetic & Manipulated Media'.

- Under 'Illegal and Regulated Behaviours', it is prohibited to use the platform for any unlawful purpose or in furtherance of illegal behaviours. This includes attempts to purchase, manufacture, smuggle, sell, purchase or distribute drugs and substances, attempts to smuggle, sell, or facilitate the selling or trafficking of other human beings or body parts, or any other facilitation of illegal activity. Under this policy - which covers illegal drugs, non-prescription drugs and precursor chemicals or substances that are used for the production of these drugs - X prohibits encouraging the use of certain drugs (including by providing consumption instructions), where required by local laws.
- Under 'Scams', it is prohibited to use X to engage in scam tactics to obtain money, property, or private information.
- Under 'Synthetic & Manipulated Media', it is prohibited to share inauthentic media, including manipulated or out-of-context media, that may result in widespread confusion on public issues, impact public safety, or cause serious harm.

Enforcement

- 'Illegal and Regulated Behaviours': The consequences for violating this policy depend on the severity of the violation as well as any previous history of violations. X may either remove a violating post, or suspend an account whose sole purpose appears to be to engage in behaviors that violate this policy. In some additional cases, accounts may also be suspended on first review.
- 'Scams': When X detects suspicious levels of activity, accounts may be locked and prompted to provide additional information to complete anti-spam challenges. Additionally, X blocks or flags (provides warnings) URLs that may cause harm. Violations related to scam activity may also be enforced with restricted reach, temporary loss of access to X features of products, profile modifications, and account suspension.
- 'Synthetic & Manipulated Media': X may label posts containing synthetic and manipulated media to help people understand their authenticity and to provide additional context. X may also apply any of the enforcement actions detailed under 'Scams' above to relevant violations.

FOCUS: DRUG-RELATED CONTENT

Under the 'Illegal and Regulated Behaviours' policy ("IRB policy"), X prohibits attempting to purchase, manufacture, smuggle, sell, purchase or distribute drugs and substances trafficking, including encouraging use (including by providing consumption instructions) where prohibited by law. Online functionalities enabling supplier-customer interactions can be exploited for illicit drug distribution, offering greater substance variety, accessibility, and reduced detection risk compared to traditional dealing.

Users can report violations of the IRB policy. Violations regarding drug-related content distribution are enforced by requiring content removal or permanently suspending the associated account.

X implements ban evasion mitigations to prevent individuals previously suspended for violating the IRB policy - including in relation to prohibited behaviours in respect of drugs - from creating new accounts and reoffending on the platform.

VIOLENT CONTENT

This content is prohibited by the X Rules policies of 'Violent & Hateful Entities' and 'Violent Content'.

- Under X's 'Violent & Hateful Entities' policy, it is prohibited to affiliate with or promote the activities of violent and hateful entities. Examples of the types of content that violate this policy include, but are not limited to, indirectly or directly engaging in or promoting violent acts on behalf of a violent or hateful entity, or recruiting, or providing or distributing services to further stated goals on behalf of such entities.
- Under X's 'Violent Content' policy, X prohibits violent speech that is high in severity and likelihood of harm (e.g. violent threats, wish of harm, incitement of violence, or glorification of violence), and violent media that is high in severity and likelihood of harm (e.g. gratuitous gore). Relevant content that is of a lower severity and likelihood of harm will be enforced accordingly (e.g. where the harm is minor or non-deliberate, where the level of graphic detail does not rise to the threshold of excessively gory, or where public interest factors exist). In any case, content prohibited by these policies will be enforced by removing from the platform or by restricting its reach such that users are significantly less likely to be exposed to it.
- Under X's 'Violent Content' policy, X prohibits violent media that is considered high in severity and likelihood of harm. Additionally, some types of graphic media are permitted on X, so long as it is properly labelled. Content that has been labelled 'violent' will not be accessible to U18 users (or users without a birth date on their profile and logged out users) and it will be obscured behind an interstitial for adult users who have not opted-in to see sensitive media by switching-on the 'display sensitive media' setting on their profile . It will also be held behind a warning that needs to be acknowledged before the media can be viewed by any other 18+ user (who does not have their sensitive media setting switched off).

Proactive detection and enforcement

- *Automated enforcement against Violent & Hateful Entities:* X uses network and behavioural analysis, keyword normalisation, and image labelling to detect and remove content under 'Violent & Hateful Entities', with high-precision tactics resulting in automated enforcement.
- *Proactive enforcement against Violent Content*: X uses a combination of machine learning and human review. These systems either take action automatically, or surface content to human moderators based on user reports and/or proactive detection methods.

Enforcement actions

- *'Violent & Hateful Entities'*: X will immediately and permanently suspend any account that X determines to be in violation of this policy.
- *'Violent Content'* (speech): X will remove content that is prohibited under this policy, and subsequent violations may result in an account being placed in read-only mode or suspended. In the case of content that is addressed by this policy, but where the severity and potential likelihood of harm is low, X may restrict its reach such that it is much less likely to be encountered by users of the service. This approach aims to ensure that enforcement is responsive and proportionate to the level of potential harm caused by such content, while taking account of freedom of expression and access to information.
- *'Violent Content'* (media): X may adjust a user's media settings if they continue to post graphic media without an appropriate label such that all future posts by that user are labelled.

CYBERBULLYING CONTENT

This content is addressed by the X Rules policies of 'Abuse & Harassment', 'Hateful Conduct', and 'Private Content'.

- Under *'Abuse & Harassment'*, X prohibits behaviour and content that harasses, shames, or degrades others. This includes targeted harassment (i.e. the malicious, unreciprocated targeting of individual/s, particularly for the purpose of humiliating or degrading them), violent event denial, incitement of harassment, unwanted sexual content and graphic objectification.
- Under *'Hateful Conduct'*, X will take action against accounts who target an individual or group of people with hateful references (e.g. references to forms of violence or violent events where a protected category was the primary target or victim, where the intent is to harass), incitement, slurs and tropes, dehumanisation, hateful imagery, or a hateful profile (i.e. hateful images or symbols in profile image or header).

- Under *'Private Content'*, it is prohibited to share private information without the permission of the person it belongs to, threaten to publicly expose someone's private information, share information that would enable individuals to hack or gain access to someone's private information without their consent, or ask for or offer a bounty or financial reward in exchange for posting someone's private information or in exchange for not posting someone's private information (i.e. blackmail). Further, it is prohibited to share media of private individuals without their permission.

Enforcement actions

- *'Abuse and Harassment'*: When determining the penalty for violating this policy, X considers a number of factors including, but not limited to, the severity of the violation, whether someone has been targeted, and an individual's previous record of rule violations. As a result, X has a graduated list of potential enforcement options including restricted reach, post removal, and account suspension.
- *'Hateful Conduct'*: Under this policy, X takes action against behaviour that targets individuals or an entire protected category with hateful conduct, as described above. When determining the penalty for violating this policy, X considers a number of factors including, but not limited to, the severity of the violation and an individual's previous record of rule violations. As a result, X has a graduated list of potential enforcement options including restricted reach, post removal, and account suspension.
- 'Private Content': When determining the penalty for violating this policy, X will consider a number of factors including, but not limited to, the severity of the violation and an individual's previous record of rule violations. Potential enforcement options include requiring post removal or permanently suspending an account.

Please provide information on any differentiated or additional measures in place for Children:
X recognises that minors are a more vulnerable group by virtue of their age. X, as a service, is not primarily for children. However, anyone above the age of 13 can sign up., Accordingly, X has implemented a range of comprehensive measures to ensure that children present on the platform benefit from additional, tailored safety protections.

The controls and measures detailed above provide a significant level of protection for all users on the service in relation to harmful content, and children benefit from these protections as well. X takes the following additional steps to protect children from harmful content:

X Rules

Certain X Rules are in place on the service where the primary objective is the protection of children who may or may not be users of the service themselves. This includes the Child Safety policy, under which X operationalises its zero tolerance approach for any forms of child sexual exploitation. This policy is explained above, but to reiterate, X's Child Safety policy is designed to protect minors from sexual and physical abuse, as well as psychological harm that may result from the sharing of such content. When CSEM is shared on the platform, X will remove it and, in almost all cases, immediately and permanently suspend the associated account. Beyond CSEM, X will also take action to remove media that shows minors in a physical altercation, where harm from the sharing of such content is apparent. Additionally, X will remove most instances of media depicting physical child abuse, even if shared to raise awareness or express outrage.

Other X Rules (e.g., Non-Consensual Nudity, Violent Content, Suicide and Self-Harm, Abuse and Harassment, Hateful Conduct) apply uniformly to all users, but are enforced rigorously to minimize children's exposure to Harmful Content.

Content restrictions for minors

- *Sensitive content*: Accounts belonging to known minors on X are restricted from accessing adult and graphic content that has been labelled. This content will not be accessible in any way to accounts belonging to known U18s, logged out users, and accounts without a birth date attached to their profile.
- *Age-inappropriate ads*: Under the X Ads policies, X prohibits knowingly marketing or advertising certain products and services to minors. Paid advertising products and advertisements containing age-inappropriate content will be tagged as "not family safe", and restricted from being shown to users under the age of 21 and logged-out users. X infers age for accounts without a date of birth to help ensure that users under the age of 21 do not see inappropriate adverts. This use of age inference is limited to advertising only, and is not a solution for age verification or age estimation more broadly. Importantly, the proportion of accounts on X without a date of birth is significantly reducing in view of the fact that inserting a date of birth has been mandatory at account opening since 2020. An example of how this works is the following: Advertisers can reach X users under the specified age range (i.e. "age buckets") selected during campaign setup. For those users that still have not provided a date of birth to the platform, a multi-layered perceptron model uses user activity on X to predict a single age for each such user. X then puts the user into all applicable age buckets if the predicted age is between the lowest and highest age in the bucket. For example, if a user's predicted age (or provided age) is 25, the user will be placed into both a 24-40 advertiser bucket, as well as a 20-30 advertiser bucket.
    - When advertisers on X choose to promote their content with X Ads, their account and content become subject to an approval process. The approval process is designed to support the quality and safety of the X Ads platform. This process helps X to check that advertisers are complying with the advertising policies. X uses age inference to enhance its ability to identify the accounts of U21 users.

Minimum age

	While X is not primarily targeted for children, individuals over the age of 13 are able to create an account. This is in line with an understanding of the evolving levels of development of children and industry-wide standards which accept that children above the age of 13 have the ability to participate in online environments such as X, albeit with additional targeted protections. ● *Neutral age gate*: Users are required to enter their age through a neutral age gate when attempting to create an account (i.e. the page will not indicate that 13 is the minimum age, nor that dedicated content restrictions and account settings will apply to the accounts of 13-17 year olds). ● *Age lock*: Once a new user enters a date of birth that indicates they are under the age of 18, they will be prohibited from re-entering an alternate date of birth for that account. ***Reporting*** ● *DMs*: X provides in-app DM reporting functionality with specific and nuanced reporting options relevant to protection of minors. Having such reporting capability ensures that reports are immediately reviewed by relevant teams and prioritised enforcement. ● *Underage accounts*: To help X enforce its 13 year old minimum age requirement, users can report accounts that they believe to be operated by children under the age of 13. X suspends accounts of users who do not meet the minimum age requirement, and takes steps to prevent such individuals from creating new accounts. ***Protective account settings*** ● *Protected posts*: Accounts belonging to known minors are defaulted to 'protected' posts. This means that known minors will receive a request when new people want to follow them (which they can approve or deny), that their posts will only be visible to their followers, and that their posts will only be searchable by them and their followers (i.e. they will not appear in public searches). ● *DM requests*: Accounts belonging to known minors are restricted to receiving Direct Messages (DMs) from accounts they follow by default. ● *Precise location*: Accounts belonging to known minors will have their precise location disabled. ● *Further privacy settings switched on by default*: Various privacy settings are switched on by default for child accounts. This includes preventing ad personalisation, preventing personalisation based on inferred identity, preventing personalisation based on places that the user has been, and preventing data sharing with third party collaborators. ***Parental resources*** ● Parents/guardians can access X's Help Center for guides on safety tools and policies to manage their child's experience on X, including reporting an account operated by their child who is underage. ***Child-friendly information*** ● X's Help Centre contains accessible information on how to use relevant safety tools (e.g. content reporting) and other steps that can be taken to ensure a safe and healthy experience on the platform. This information is designed to be comprehensible for the youngest permitted user (i.e. 13 years old). ● X provides warnings about the implications of changing default settings, ensuring that child users are appropriately informed if they are considering changing protective settings that are switched on for child users by default. Through its comprehensive child-focused measures—including X Rules, default privacy protections, age-based content restrictions, accessible reporting and parental resources—X ensures that children in Singapore are effectively protected from Harmful Content. These controls, tailored to the needs of minors, not only mitigate access to inappropriate material, but also empower young users and their guardians to navigate the platform safely. By embedding these protections into the platform's design and operations, X provides a safer online space for children over the age of 13 who are able to create an account.
(Ai.) Measures for all end-users	
Community guidelines and standards and content moderation Paragraph 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	**Please provide information on the community guidelines and standards in place and how they address the categories of harmful content in paragraph 4** As set out in our response in relation to Paragraph 8 above, Singapore end-users' exposure to Harmful Content is minimised on X as a result of a suite of comprehensive measures spanning community guidelines and standards (i.e. X's Terms of Service and the X Rules), proactive detection and enforcement, user reporting mechanisms, law enforcement reporting mechanisms, recommender system controls, and advertising and monetisation safeguards. Please refer to our response above, where we have set out specifically how the X Rules address all types of Harmful Content under the Code, and apply a graduated enforcement approach that is reasonable and proportionate based on the risk of harm to users. To reiterate, each type of Harmful Content is addressed under the following X Rules and enforced in the following ways (referring to the response in relation to Paragraph 8 for extended detail): <u>Sexual Content</u>

• Content moderation measures that are put in place and into effect by the Service. The Service's community guidelines and standards must address the categories of harmful content in paragraph 4 and must be published.	• Within the scope of Sexual Content as defined by the Code, X expressly prohibits ***any*** content related to child sexual exploitation, non-consensual nudity, bestiality, sexual violence, and necrophilia under the ***'Child Sexual Exploitation'***, ***'Non-Consensual Nudity'***, ***'Violent Content'*** and ***'Abuse and Harassment'*** policies.[3] In most cases, users who post or share this content, or otherwise violate X's policies with respect to these types of content, will have their accounts immediately and permanently suspended. • In addition to the above types of egregious Sexual Content, X addresses relevant content that is inappropriate for children under the ***'Adult Content'*** policy. While users are permitted to share consensually produced and distributed adult nudity or sexual behaviour content, it must be properly labelled and not prominently displayed. When this content is labelled, it will not be accessible to children and it will be obscured behind an interstitial and not recommended to adult users who have not opted-in to see sensitive media by switching-on the 'display sensitive media' setting on their profile. Users can report adult content that has not been labelled, and X also has proactive detection systems to identify unlabelled adult content which, once identified, will be appropriately labelled by X. If a user fails to label adult content on multiple occasions, their account may be suspended. Violent Content • This type of Harmful Content is addressed by the X Rules of ***'Violent & Hateful Entities'*** and ***'Violent Content'***. Content defined under these policies to present a higher level of severity and likelihood of harm will be removed from the platform once detected, while content of a lower severity and likelihood will be visibility restricted such that users are significantly less likely to be exposed to it. Suicide and Self Harm • This content is prohibited under the X Rule of ***'Suicide and Self Harm'***. Generally, X will remove content that violates this policy, while it may also temporarily lock an offending user out of their account before they can post again or may otherwise permanently suspend an account that is dedicated to posting relevant material. Cyberbullying Content • This type of Harmful Content is addressed by the X Rules of ***'Abuse & Harassment'***, ***'Hateful Conduct'***, and ***'Private Content'***. Similar to Violent Content, X prohibits relevant content that presents a higher level of security and likelihood of harm, in which case it will be removed from the platform or may otherwise be enforced with account suspension (this includes all types of content addressed by 'Private Content'). In cases of lower severity and likelihood of harm, X may restrict the visibility of relevant content such that it is much less likely to be encountered by users on the platform. Content Endangering Public Health and Facilitating Vice/Organised Crime • These types of Harmful Content are prohibited under the common X Rules of **'Illegal and Regulated Behaviours'**, ***'Scams'***, and ***'Synthetic & Manipulated Media'***. Within ***'Scams'***, X takes a variety of tailored enforcement actions to disrupt bad actor behaviour (e.g. requiring suspicious accounts to complete anti-spam challenges). Under ***'Synthetic & Manipulated Media'***, certain content (e.g. inauthentic media that may impact public safety) may be labelled to help users understand its authenticity and relevant context. Under *'**Illegal and Regulated Behaviours**'*, violations will be enforced according to their severity; with content removal and account suspension enforcement actions available. **Please provide information on the content moderation measures in place and how they are enforced or effected. Please include screenshots or evidence where possible.** As set out in relation to Paragraph 8 above, the risk of Singapore end-users' exposure to Harmful Content is minimised on X by a comprehensive suite of reasonable and proportionate measures, including content moderation. To reiterate and expand on the information provided above, X's content moderation measures include the following: Proactive detection and enforcement As set out under paragraph 8, X employs a variety of sophisticated machine learning algorithms, natural language processing, and image processing models to proactively detect content that may violate the X Rules and enqueue this content for human review. X's variety of automated detection systems include machine learning and heuristics models. These systems aim to reduce the likelihood that users encounter the worst kinds of harmful content, by detecting and enforcing against such content before it is reported to X. These models vary in complexity and in the outputs they produce. For example, the model used to detect abuse on the platform is trained on abuse violations detected in the past. Content flagged by these machine learning models are either reviewed by human content reviewers before an action is taken or, in some cases, automatically actioned, based on the historical accuracy of the model's output. Heuristics are typically utilised to enable X to react quickly to new forms of violations that emerge on the platform. Heuristics are common patterns of behaviours, text, or keywords that may be typical of a certain category of violations. Pieces of content detected by heuristics may also get reviewed by human content reviewers before an action is taken on the content. These heuristics are used to flag content for review by human agents proactively. Automations are monitored dynamically for ongoing performance and health. Reporting mechanisms

[3] *Note that reports received and enforcement action taken under the 'Violent Content' and 'Abuse and Harassment' policies was not included in relation to 'Sexual Content' for the data reporting purposes below because they relate overwhelmingly to other types of Harmful Content such as 'Violent Content' and 'Cyberbullying Content'.*

- ***User reporting***: Users can report alleged violations of the above X Rules through our reporting tools. These tools are available in-app alongside individual items of content and accounts, or through the Help Centre.
- ***Law enforcement reporting***: X provides a dedicated law enforcement reporting form through which applicable Singapore law enforcement agencies can report content under applicable local laws.
- ***Trusted partners***: X maintains partnerships with trusted organisations who are able to notify X of potential violative content on the platform. This includes SG Her Empowerment and AWARE Singapore for gender based violence and non-consensual nudity, and Samaritans of Singapore for Self-Harm.

<u>Manual content moderation</u>

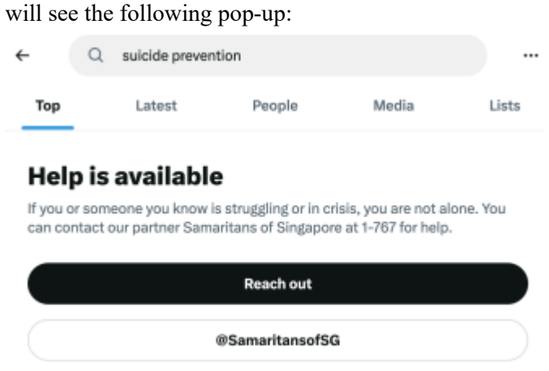
- As above, X uses human agents to review a combination of potentially violative content detected by a variety of advanced automated systems, and potentially violative content and accounts reported to X by users, law enforcement agencies, and other partners.
- X's scaled operations team possesses a variety of skills, experiences, and tools that allow them to effectively review and take action on reports across all of X's Rules and policies.
- X's content moderation teams conduct proactive sweeps for certain high-priority categories of potentially violative content both periodically and during major events, such as elections. Content moderators also proactively review content flagged by heuristic and machine learning models for potential violations of other policies, including our '*Adult content'*, *'Violent content'*, '*Child sexual exploitation (CSE)*' and *'Violent and hateful* entities' policies.
- ***Reactive manual review*:** Dedicated operational teams are responsible for responding to user reports in an accurate and timely manner. These teams are trained to look for contextual indicators to determine whether the content or account is in violation of X's policies. Comprehensive guidance for investigating accounts can be found in the training materials, with distinct sets of instructions provided for each workflow.
- ***Language expertise*:** X has operational teams with diverse language expertise to ensure accurate translations and a thorough understanding of regional and local contexts in the decision-making process.
- ***Training*:** Training is a critical component of how X maintains the health and safety of the public conversation by enabling content moderators to accurately and efficiently moderate content posted on our platform. Training at X aims to improve content moderators' enforcement performance and quality scores by enhancing their understanding and application of X Rules through robust training and quality programs, and the continuous monitoring of quality scores.

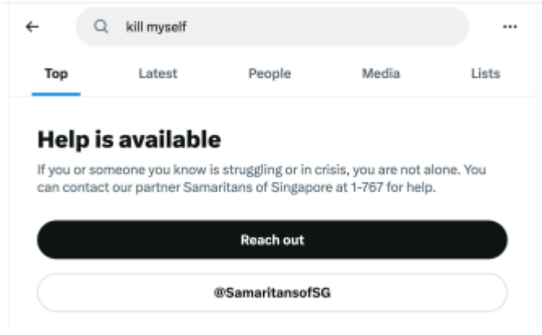
<u>Enforcement actions</u>

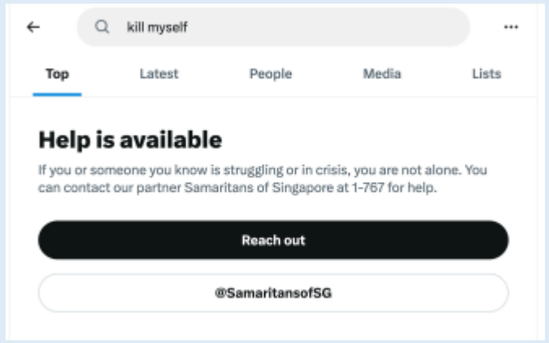
Above, X has set out the specific enforcement actions that it may utilise in relation to violations of X Rules relevant to harmful content. Set out below is a consolidated view of X's available enforcement actions, and additional information on what each enforcement action entails:

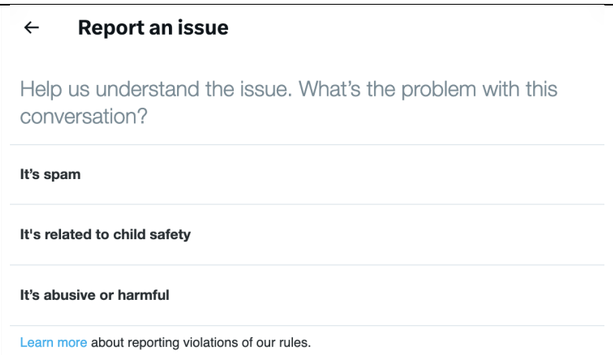
- ***Account suspension:*** X takes action to suspend an account if it determines that a user has engaged in repeated violations of X's policies, or in certain cases immediately suspend account for violating specific policies that cause significant risk to X or pose a threat to users (for example, child sexual abuse material ("CSEA material"), using the platform to incite violence, attempts to manipulate the platform or spam users, fraud, user privacy violations, violent threats, targeted harassment, etc.).
- ***Account lockdown*:** X may temporarily limit an account's ability to post, Repost, or Like. The user can read their timelines and will only be able to send Direct Messages to their followers.
- ***Post removal*:** When it is determined that a post violated the X Rules and the violation is severe enough to warrant post removal, X will require the violator to remove it before they can post again. They will need to go through the process of removing the violating post or appealing the removal request if they believe X made an error. The post will be hidden from public view with a notice during this process.
- ***Labelling a post*:** X may add a label to the content to provide context and additional information to users. Additionally, Community Notes may also be visible on posts to provide additional context.
- ***Direct Message-level enforcement*:** In a private Direct Message conversation, when a participant reports the other account, X will stop the violator from sending messages to the account who reported them. The conversation will also be removed from the reporter's inbox. However, if the reporter decides to continue to send Direct Messages to the violator, the conversation will resume.
- ***Placing a Direct Message behind a notice*:** In a group Direct Message conversation, the violating Direct Message may be placed behind a notice to ensure no one else in the group can see it again.
- ***Ban evasion detection*:** X has automated systems which detect attempts to evade suspension, automatically suspending a user creating a new account.
- ***Controls for inauthentic use or automated exploitation of the service*:** Accounts on X must be authentic. X does not allow the creation, operation or mass-registration of accounts that are not legitimate, genuine, and transparent as to their source, identity and popularity. This is intended to prevent the use of X's services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people's experience or platform manipulation defenses on X. As a matter of practice, we're consistently remediating spam and platform manipulation on X, and constantly testing new, and iterating on existing, automations that widen the scope of our proactive measures. As a result of X's efforts in the spam and platform manipulation space, it suspended ~335 million accounts in the past year.
- ***Restricted reach*:** In certain cases and where appropriate, X restricts the reach of posts that violate the Rules or create a negative experience for other users by making the posts less discoverable on X. X's enforcement policy outlines the measures that it takes to restrict the reach of a post, which include:
    - Excluding the post from search results, trends, and recommended notifications;
    - Removing the post from the For you and Following timelines;
    - Restricting the post's discoverability to the author's profile;

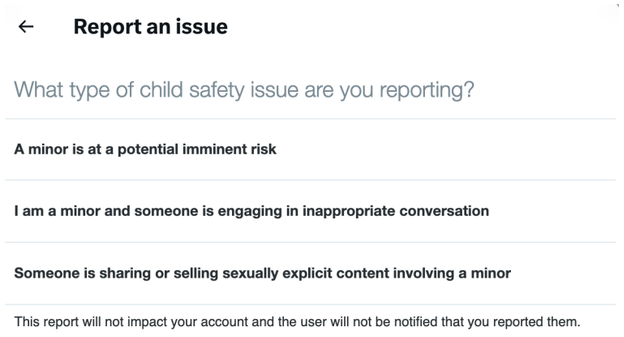
	○ Downranking the post in replies; and ○ Restricting Likes, replies, Reposts, Quotes, bookmarks, share, or pin to profile. ● ***Placing a post behind a notice*****:** X may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through. ● ***Withholding a post based on age*****:** X restricts views of specific forms of sensitive media such as adult content for viewers who are under 18 or viewers who do not include a birth date on their profile with interstitials. ● ***Withholding a post or account in a country*****:** X may withhold access to certain content in a particular country if it receives a valid and properly scoped request from an authorised entity in that country.
Empower end-users and improve safety Paragraph 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	**Please describe in detail the tools that address all three sub-categories and how they meet the outcomes described in paragraph 12 of the Code. You may also provide information on additional tools outside these sub-categories that meet the outcomes of the Code. Do include screenshots or evidence where possible.** Please refer back to the responses to paragraphs 8 and 9 for full comprehensive detail on relevant tools that X provides to Singapore end users to enable them to manage their safety and effectively minimise their exposure to, and mitigate the impact of, harmful content and unwanted interaction. This includes tools to restrict the visibility of harmful content and unwanted comments, to limit the visibility of the end-user's account and contact/interactions with other end-users, and to limit location sharing. **All End-users in Singapore:** **Paragraph 12(a): Tools to restrict visibility of harmful content and/or unwanted comments** **1. User tools to restrict the visibility of harmful content and/or unwanted comments** X equips Singapore end-users with intuitive tools to restrict harmful content and unwanted comments, ensuring a safer platform experience. To summarise, users can mute accounts or specific words, preventing such content from appearing in their "For You" and "Following" timelines or notifications. The block feature stops unwanted accounts from engaging, following, or messaging them, reducing exposure to harmful interactions. The "Not interested" option allows users to remove posts from feeds and prevent similar content recommendations. Accessible in-app reporting tools enable users to flag violations of relevant policies, triggering swift human review and appropriate enforcement, including account suspension, content removal or restricted reach. The "display sensitive media" setting is available to adult users but disabled by default, such that labelled content is obscured behind an interstitial. In any case, labelled content is inaccessible to U18 users, logged-out users and users without a birthdate on their profile. Safe Search mode filters sensitive content from results, and users can enable quality filters to exclude low-quality notifications. Protected Posts limit comment visibility to approved followers. Community Notes allow users to add context to misleading posts, reducing their impact. These tools, supported by clear Help Centre guides, empower users to proactively manage their exposure to harmful content and unwanted comments, fostering a healthy online environment. **All End-users in Singapore:** **Paragraph 12(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** **2. User tools to limit the visibility of the end-user's account, contact and/or interactions** X offers Singapore end-users a suite of user-friendly tools to limit account visibility, contact, and interactions, enhancing privacy and safety. To summarise, the Protected Posts setting can be opted-into by any user (and is enabled by default for child users, logged out users, and users without a birthdate on their profile), which restricts the visibility of posts to approved followers, prevents reposting, and limits searchability. Users can block accounts to stop engagement, following, or DMs from blocked accounts. If a user wants to prevent a blocked account from viewing their posts, they can enable Protected Posts - this requirement prevents abuse of the block functionality to disseminate harmful content about a blocked user. Muting accounts or words hides it from the user's timelines and notifications. DM settings allow users to restrict message requests to "verified users," "no one," or "everyone," with requests sent to a separate inbox. Users can enable low-quality and graphic media DM filters to hide spam or sensitive content and disable read receipts. Photo tagging can be limited to accounts that the user follows or turned off entirely. Default enabled discoverability settings for all users prevent accounts from being found via email or phone. Advanced notification filters block alerts from new or unverified accounts, and quality filters reduce low-quality content. In-app reporting empowers users to notify X of violative interactions, with specific child-friendly DM reporting options for grooming. These tools, accessible via clear settings and Help Centre guides, empower users to tailor their visibility and interactions effectively. **All End-users in Singapore:** **Paragraph 12(c): Tools to limit location sharing** **3. User tools to limit location sharing** X provides Singapore end-users with accessible tools to limit location sharing, prioritising privacy, especially for minors. To summarise, location sharing is disabled by default, requiring users to opt-in to attach location labels (e.g., city names) to posts, excluding precise coordinates. Users can choose not to share location for individual posts, even if they have previously enabled location sharing in settings.

	For in-app camera posts, an optional "precise location" feature includes latitude and longitude, but users must actively enable it. In their privacy settings, users can delete all past location data from previous posts with one click at any time, offering retroactive control. The Help Centre provides clear, age-appropriate guidance, cautioning users to consider risks before sharing location, with specific advice for minors. On X, minors benefit from default high-privacy settings. In-app prompts and intuitive menus make these tools easy to use, ensuring that users can manage their location data with clarity and confidence. These measures empower users to minimise geolocation exposure, balancing flexibility for those who choose to share with robust protections for privacy and safety.
Empower end-users and improve safety Paragraph 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 ● 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. *And* ***Protection for children*** Paragraph 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 ● Include information on tools available to protect children 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	**Please <u>explain</u> how your service has made information related to online safety easy to be accessed, including local information. You may also include details on programmes and initiatives to educate and raise awareness of such information.** **All End-users in Singapore:** Please refer back to the responses to paragraphs 8 and 9 for full comprehensive detail on easily accessible safety information provided to Singapore end-users. **1. Easy to understand online safety information** X ensures its safety tools are user-friendly and comprehensible for Singapore users, including those as young as the minimum permissible age of 13, through intuitive design and clear communication. Tools including muting, blocking, reporting, and Protected Posts are accessible via straightforward in-app menus with descriptive labels (e.g., "*Block*," "*Mute words*"). The Help Centre offers concise guides in simple, age-appropriate language, explaining each tool's purpose and usage, such as how to restrict DMs or hide sensitive content. Settings such as "Display sensitive media" or "Protect your videos" are clearly named, with toggle options accompanied by brief explanations. In-app prompts guide users through actions like reporting grooming or enabling Safe Search, which filters sensitive content. Warnings appear when users attempt to change protective settings, ensuring informed decisions. The X Safety account shares regular updates on features, enhancing awareness, and X publishes blog posts in its Help Centre to flag key changes. For high-risk searches, crisis resource links are presented clearly (and we have country specific resources available for Singapore as set out below). Community Notes empower the X community to clarify and add relevant context to misleading posts. These elements, combined with visual cues like interstitials for sensitive media where relevant, make tools accessible and easy to navigate, empowering users to manage their safety confidently, regardless of technical expertise or age. **2. Local Information** We have a dedicated page on our Help Centre which provides dedicated resources for users in Singapore if they or someone they know is struggling or in crisis. If users input relevant search terms they will see the following pop-up: 

initiatives to educate and raise awareness of such information.	**Please <u>explain</u> how your service has made information related to online safety easily accessible for children, including local information. You may also include details on programmes and initiatives to educate and raise awareness of such information.** **Children:** Please refer back to the response to paragraphs 8 and 9 for full comprehensive detail on easily accessible safety information provided to Singapore children. **1. Children can easily access safety information that is understandable** X ensures that child users in Singapore can easily access understandable safety information through tailored resources. The Help Centre provides guides in clear, concise language designed for users as young as the permissible youngest age of 13, explaining tools such as reporting, muting, blocking, and Protected Posts. These guides use simple terms to detail how to manage harmful content or interactions, such as in-app DM reporting tools with specific pathways for child safety related issues. In-app prompts offer immediate guidance, like warnings when attempting to alter default protective settings (e.g., disabling Protected Posts), ensuring comprehension. For high-risk search terms, X displays crisis resources and links to Singapore-specific support, like mental health services, in age-appropriate formats. The X Safety account shares feature updates in accessible language. Default settings, like restricted access to sensitive content and disabled precise location, are explained clearly in settings menus. These resources, combined with intuitive interfaces, empower children to navigate safety tools confidently, fostering a secure online experience aligned with their developmental needs. **2. Information on tools available to protect children** X's Help Centre offers comprehensive information for Singapore users, particularly parents/guardians, on tools to protect children. Guides detail the Child Safety policy, emphasising X's zero tolerance for child sexual exploitation, with immediate account suspension and content removal. The Help Centre explains default protections for minors, including Protected Posts, which limit the visibility of their posts to approved followers, and restricted DMs, which restricts inbound messages to followed accounts. Precise location sharing is disabled for minors, and sensitive content is inaccessible. The Help Centre outlines reporting mechanisms for underage accounts or grooming, including specific DM reporting options, and how parents can report accounts of children under 13. Information on age-inappropriate ad restrictions ensures that minors are not presented with inappropriate commercial communications. X's minimum age of 13 is set out clearly in the Help Centre. Guides also cover privacy settings, including ad personalisation, which is disabled by default for minors. These resources, written in clear language, empower child users to leverage X's robust safety tools, ensuring that they are effectively protected from harmful content and interactions.
Empower end-users and improve safety Paragraph 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. *And* ***Protection for children*** Paragraph 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	**Please provide information on safety resources provided to users who use high-risk search terms with new examples of search terms, that are not limited to self-harm and suicide. Do include screenshots or evidence where possible.** **All End-users in Singapore:** X provides safety resources for Singapore users who search high-risk terms related to self-harm and suicide, to mitigate potential harm. When users enter high-risk terms, X displays crisis resources and links to third-party support, such as Singapore's mental health services, tailored to the user's region. For example, if users input search terms for self-harm or suicide they will see the following pop-up: 

paragraph 21) such as, but not limited to, local suicide prevention hotlines, if available.	**Please provide information of safety resources provided to users who use high-risk search terms. The information provided is not limited to terms relating to self-harm and suicide. Do include screenshots or evidence where possible.** **Children:** As mentioned in response to paragraph 13 above, the X Rules and Policies (which protect user safety, as well as platform and account integrity) together with all other information related to online safety and security on X, are publicly accessible on the Help Center. X keeps the Help Center regularly updated anytime it modifies the X Rules. If users input search terms for self-harm or suicide they will see the following pop-up:  X provides safety resources to any user who searches for high-risk terms, irrespective of whether they are an adult or a child. The information that is presented is prominent, accessible, intuitive and easy to understand for users as young as the youngest permissible age of 13.
Proactive detection and removal Paragraph 15: End-users' exposure to child sexual exploitation and abuse material and terrorism content on the Service must be minimized 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.	**Please provide information on the measures, and include screenshots or evidence where possible.** **All End-users in Singapore:** **<u>Measures taken to minimise Singapore users' exposure to CSEA material</u>** As described in the responses to paragraphs 8 and 9, X enforces a zero-tolerance approach to CSEA, which minimises Singapore users' exposure to associated material. Please refer to the responses to paragraph 8 and 9 for comprehensive detail on the measures implemented by X in this regard. To summarise, the X Rules strictly prohibit CSEA material with immediate and permanent account suspension in the case of almost all violations. Automated hash-matching technology detects and removes known CSEA material, with reports made to the NCMEC where applicable. X blocks known CSEA URLs and suspends accounts who attempt to upload these URLs and text-based systems are also used to automatically enforce violations of the Child Sexual Exploitation policy. X further leverages image and text-based automated systems (including hash matching) to proactively detect and enqueue potential CSEA material for swift human review, increasing the volume of CSEA material that is proactively removed (and offending accounts suspended) before it can be disseminated on the platform. Partnerships with accredited partners, such as the NCMEC, enhance detection by continuously expanding X's databases of known CSEA hashes and CSEA URLs. In-app reporting tools, including specific child-friendly DM reporting pathways for grooming or CSEA solicitation, enable users to report the minority of CSEA material that is not proactively detected and enforced, with reports received in a priority queue and reviewed by dedicated child safety experts. Default settings - as iterated above - ensure a higher level of safety by default and mitigate the risk that bad actors can identify and engage with U18s on the platform. These measures ensure that CSEA content is swiftly detected and removed, and accounts are suspended to prevent re-offending. During the 2025 reporting period, X implemented a number of improvements to its overall CSEA detection and enforcement systems. These included: ● Training additional agents for proactive keyword and media sweeps; ● X used signals for account investigations from its membership of and participation in Project Lantern, the cross-platform signal-sharing program launched by the Tech Coalition to combat online CSEA; ● Launched an improved DM reporting flow with intuitive child safety pathways; ● Maintained its hash-matching system to register new hashes for scalable automated enforcement; ● Continued to invest in proprietary technology to reduce the burden on users to report CSEA for enforcement. As demonstrated by the data reported, the vast majority of enforcement actions taken under the CSE policy were enforced proactively; and ● X continuously worked to detect evolving trends in CSEA and related activity to enhance its proactive detection and enforcement. Please see screenshots of X's improved child-friendly DM reporting pathways:

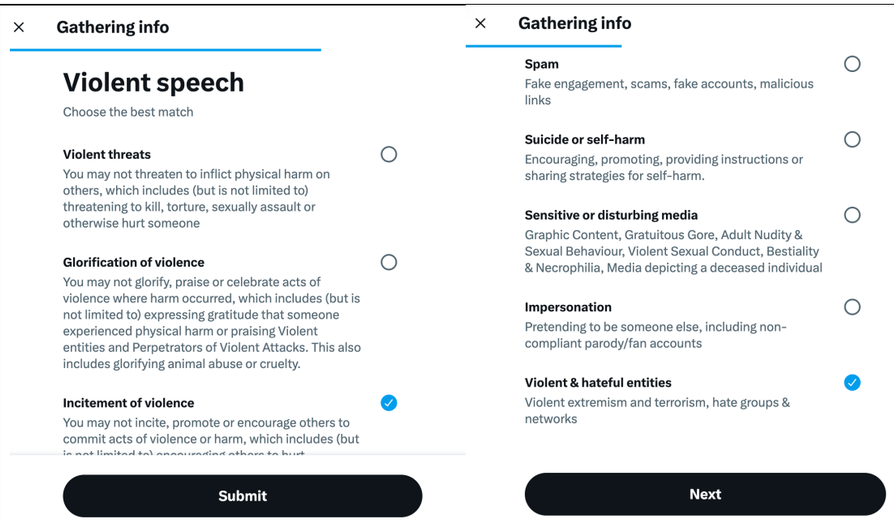




Measures taken to minimise Singapore users' exposure to terrorism material

As described in the responses to paragraph 8 and 9, X implements stringent measures to minimise Singapore users' exposure to 'terrorism material' under the Violent & Hateful Entities and Violent Content policies, which prohibit promoting or affiliating with terrorist activities, or sharing content that threatens, incites, glorifies, or expresses desire for violence or harm. Please refer to the responses to paragraph 8 and 9 for full details of the measures implemented by X in this regard. To summarise, violators face immediate and permanent account suspension under the Violent & Hateful Entities policy, and graduated enforcement under the Violent Content in line with the severity and likelihood of potential harm. Automated systems proactively detect and automatically enforce violating content with content removal and account suspension where appropriate. User reporting tools allow Singapore users to flag content under the Violent & Hateful Entities and Violent Content policies, with reports prioritised for review by expert moderators. Partnerships with third parties enhance X's ability to detect evolving potential terrorist activity on the platform. Ban evasion detection prevents suspended users from creating new accounts, enhancing the effectiveness of X's enforcement actions and reducing the risk of re-offending across the platform. These proactive and reactive measures ensure that terrorism material is swiftly removed from the platform.

Please see the below screenshots of X's intuitive and accessible reporting pathways for Violent & Hateful Entities and Violent Content.



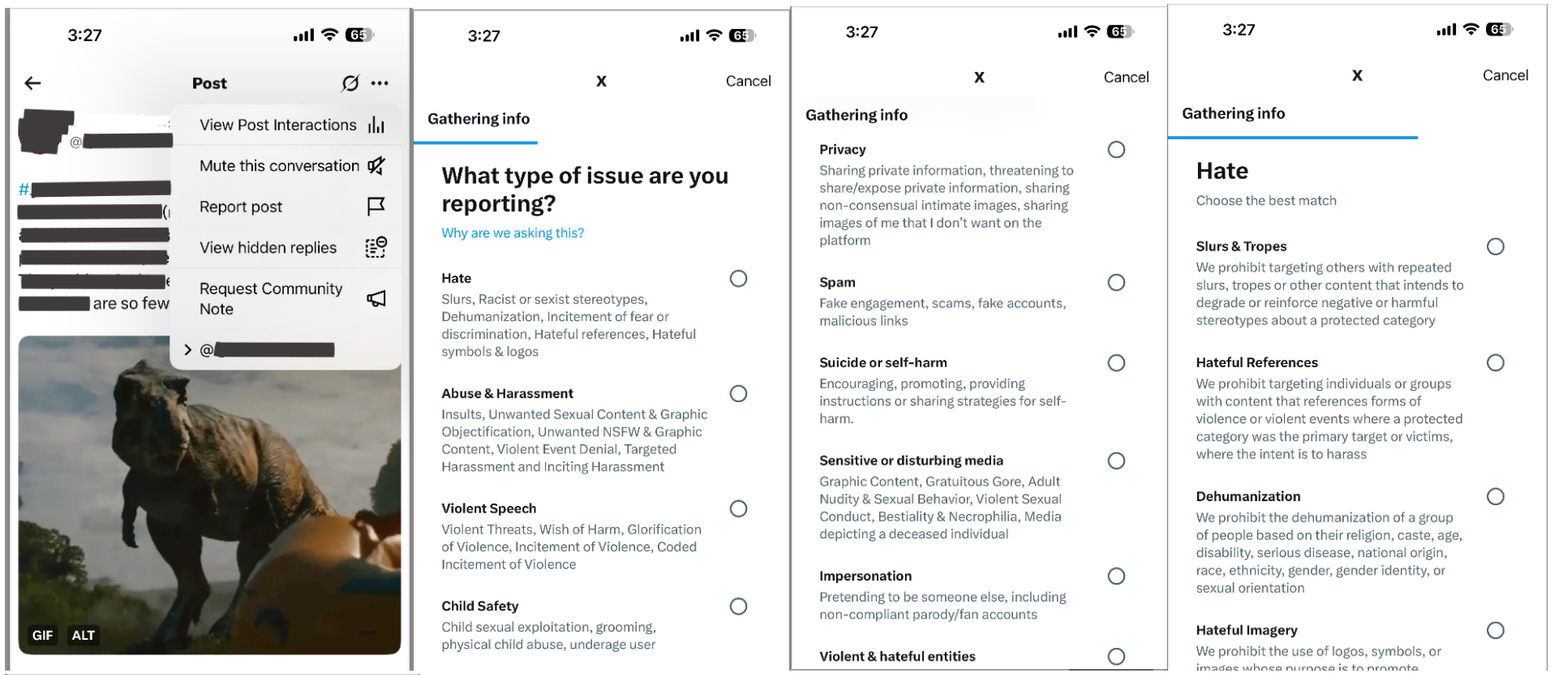
Proactive detection and removal Paragraph 16: End-users must be protected from preparatory child sexual exploitation and abuse activity and terrorism activity on the Service through reasonable and proportionate steps taken by the Service to proactively detect and swiftly remove preparatory child sexual exploitation and abuse activity (such as online grooming for child sexual abuse) and terrorism activity (such as glorifying or endorsing terrorist activities and recruitment).	**Please provide information on the measures, and include screenshots or evidence where possible.** **All End-users in Singapore:** **Measures taken to protect Singapore users from preparatory CSEA activity** Please see the responses to paragraphs 8 and 9, which set out in detail the comprehensive measures that X has implemented to protect end-users from preparatory CSEA activity. To summarise, this begins with the CSE policy which expressly prohibits relevant behaviour such as: ● Sharing, requesting or expressing a desire for, or describing: ○ Depictions of a child engaged in sexually explicit or suggestive acts, ○ Sexualised commentaries about or directed at a known (or unknown) minor, ○ Links to third-party sites that host child sexual exploitation materials, or ○ Fantasties about or promoting engagement in child sexual exploitation. ● Promoting or normalising sexual attraction to minors as a form of identity or sexual orientation; ● Sending sexually explicit media to a child, or engaging or trying to engage a child in sexually explicit conversation; ● Trying to obtain sexually explicit media from a child, or trying to engage a child in sexual activity through blackmail or other incentives; ● Recruiting, advertising or expressing an interest in a commercial sex act involving a child, or in harboring and/or transporting a child for sexual purposes; and ● Threatening to share, or requesting that others share sexual media of minors, including self-generated sexual media or minors, particularly when engaging in blackmail or bounty tactics. When X identifies accounts engaging in these types of behaviours, they are immediately and permanently suspended, and reported to the NCMEC where applicable. X employs a series of proactive automated detection tools to detect preparatory CSEA behaviour, including text-based models that scan posts for language, terms and phrases commonly associated with preparatory CSEA behaviour. Suspected violations of the CSE policy are enqueued for human review. Additionally, and as described on several instances above, X provides intuitive and accessible reporting pathways for users to report preparatory CSEA activity in DMs. X may leverage these reports to pursue further investigations and proactively detect other instances of preparatory CSEA behaviour in DMs. These systems and tools work collaboratively to enhance X's ability to detect and disrupt preparatory CSEA activity, and protect its users from such behaviour. X takes steps to ensure that individuals who are permanently suspended for violating the X Rules - including, for instance, engaging in preparatory CSEA behaviour - are prevented from creating any new accounts in the future. Attempts to circumvent an enforcement action by creating additional accounts or repurposing existing accounts to replace or mimic a suspended account are considered to violate X's ban evasion policy and will result in permanent suspension at first detection. X maintains sophisticated technological systems to detect ban evasion. As set out in paragraphs 8 and 9, X provides users with a variety of tools and controls that empower them to take steps to protect themselves from unwanted interactions, including preparatory CSEA behaviour. This includes default privacy settings for U18 accounts (e.g. 'Protected Posts') to significantly reduce the discoverability of minors on the service to bad actors, and tools such as 'Block' that enable users to limit engagement and interaction from other accounts.

	Measures taken to protect Singapore users from preparatory terrorism activity Please see the responses to paragraphs 8 and 9, which set out in detail the comprehensive measures that X has implemented to protect end-users from preparatory terrorism activity. To summarise, this begins with the 'Violent and Hateful Entities' and 'Violent Content' policies which expressly prohibits relevant behaviour such as: ● Affiliating with, or promoting, the activities of terrorist organisations, including recruiting, or providing or distributing services (such as media/propaganda) to further stated goals; ● Making violent threats; ● Wishing, hoping or expressing a desire for harm; ● Inciting, promoting or encouraging others to commit acts of violence; and ● Glorifying violence. In paragraphs 8 and 9, X sets out the sophisticated technologies and systems it has implemented to proactively detect and enforce against behaviour that violates the Violent & Hateful Entities and Violent Content policies, including preparatory terrorism which is prohibited as set out above. Users can also report content or other activity that violates these policies, and X has dedicated expert teams that review such reports of violent and hateful entities with priority.
(Aii.) Measures for children	
Community guidelines and standards and content moderation Paragraph 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 into effect 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	**Please provide information on the measures, and include screenshots or evidence where possible.** **All End-users in Singapore:** **Children:** Under the X Terms of Service, users below the age of 13 are not permitted to sign up for the service. Users who do not meet the minimum age requirement will have their account locked. In addition, parents and guardians are able to access the X Rules and Policies and other resources in the Help Center to learn more about how to keep their child's account and experience on X safe, secure and welcoming. This includes a form permitting them to report accounts (such as that of their child) who they suspect to be owned by individuals under the minimum age. While X, as a service, is not targeted at younger users, children above 13 are permitted to create an account. Accordingly, X has implemented comprehensive measures to ensure the safety of minors on the service and to minimise their exposure to inappropriate content. These measures include community guidelines and standards (i.e. X Rules) that cover inappropriate content, and content moderation measures. Most notably, under the Adult Content and Violent Content policies, users are required to label sensitive media that they post on X. If users fail to label sensitive media that they post, unlabelled content can be reported to X. Users who repeatedly fail to appropriately label sensitive media may have their account placed in read-only mode or suspended, and X may adjust an account's settings so that any content posted by that account in the future will be appropriately labelled. When content has been labelled as sensitive media, including content labelled under the Adult Content policy, as well as graphic media labelled under the Violent Content policy, it will not be accessible to U18 accounts, logged-out accounts, or to users without a birth date on their profile. Adult users are required to opt-in to see this content, and can do so by switching-on their 'display sensitive media' setting. In any case, it is prohibited to share sensitive media in prominent or otherwise highly visible surfaces such as user profiles or banner photos. X also takes steps to proactively detect and label sensitive media that has not been labelled in accordance with the X Rules, including proactive automated detection and enforcement systems. X has taken steps to further improve these automated systems, implementing additional proactive heuristics to detect and label Adult Content and Violent Content. These additional proactive heuristics are also employed to review sensitive media enqueued for manual review. ***Community guidelines and standards*** Inappropriate content is addressed by the same X Rules policies that prohibit or restrict harmful content as explained above in paragraphs 8, 9 and 11. These policies are published on the Help Centre. *Paragraph 17(a): Sexual content* ● This type of inappropriate content is governed by the X Rules policy of 'Adult Content'. Please see responses above which detail how this policy requires users who share adult content to apply labels to their posts so that such content can be made inaccessible to minors. *Paragraph 17(b): Violent content*

	• This type of inappropriate content is governed by the X Rules policy of 'Violent Content'. Please see responses above which detail how this policy requires users who share graphic media or other sensitive media to apply labels to their posts so that such content can be made inaccessible to minors. *Paragraph 17(c): Suicide and self-harm content* • Inappropriate content that contains depictions of self-harm, or that may encourage, mislead or instruct children in dangerous activities or self-harm, is governed by the X Rules policy of 'Suicide and Self Harm'. Please see responses above which detail how this policy prohibits any promotion, or otherwise encouragement, of suicide or self-harm. *Paragraph 17(d): Cyberbullying content* • This type of inappropriate content is governed by the X Rules policies of 'Abuse & Harassment' and 'Hateful Conduct'. Please see the responses above which detail how this policy prohibits, among other things, targeted harassment, inciting harassment, certain insults, hateful references, slurs and tropes, dehumanisation, hateful imagery, and hateful profiles. ***Content moderation*** *Paragraph 17(a): Sexual content* • Please see responses above which detail how labelled adult content will not be accessible to minors, how users can report unlabelled content, how X maintains proactive systems to detect unlabelled content, and how X may enforce against unlabelled content by labelling such content itself, adjusting the media settings of the relevant account (so that all future media shared by that account is automatically labelled), and suspend accounts who permanently fail to appropriately label adult content. *Paragraph 17(b): Violent content* • Please see responses above which detail how labelled graphic media under the 'Violent Content' policy will not be accessible to child users, and how X maintains proactive systems and reporting channels to detect and appropriately label unlabelled graphic media. *Paragraph 17(c): Suicide and self-harm content* • Please see responses above which detail how suicide and self-harm content is proactively detected - and how users can report such content - and how X will enforce against violations. *Paragraph 17(d): Cyberbullying content* • Please see responses above which detail how violations of the Abuse and Harassment and Hateful Conduct policies are proactively detected - and how users can report such content - and how X will enforce against violations. ***Other measures*** • X maintains a comprehensive variety of content and interaction controls that empower child users to protect themselves from inappropriate content. These controls are intuitive and accessible, and explained in language appropriate for the youngest permissible user in the Help Centre (with relevant information pages linked on in-app surfaces, such as settings options, where relevant). See the responses above, particularly under paragraphs 8 and 9, for further detail. • X maintains recommender system controls that systematically reduce the likelihood that inappropriate content is recommended to users, including child users. See the responses to paragraphs 8 and 9 for further detail. • X maintains default account settings for the accounts of child users to ensure a higher level of protection from inappropriate content and unwanted interactions.
Community guidelines and standards and content moderation Paragraph 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	**Where applicable, please <u>provide details</u> on how the measures apply to content targeting including, but not limited to, advertisements, promoted content and content recommendations.** **Children:** Please refer to the responses to paragraphs 8, 9 and 17 for detail of the measures that X has in place to prevent Singapore child users from being targeted with content that is detrimental to their physical or mental well-being, including measures relevant to advertisements and promoted content such as X's ad and monetisation policies and their enforcement, and recommender system controls.

4 and 17. In this regard, content targeting refers, but is not limited, to advertisements, promoted content and content recommendations.	
Protection for children Paragraph 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 Limit location sharing. iii. Limit location sharing	**Please <u>describe in detail</u> the tools available on your service for children and/or parents/guardians to effectively manage the content that children see and/or their experiences, and how they meet the outcomes described in paragraph 19 of the Code.** **Children:** **Paragraph 19(a): Tools to effectively manage the content that children see and/or their experiences.** ***Tools that enable children and their parents/guardians to effectively manage the content that a child sees and/or their experiences*** Please refer to paragraphs 8 and 9 for comprehensive detail on the user controls and tools that X provides to all user and child accounts - including content access controls - to effectively manage the content that they see. Parents and guardians are able to access the X Rules and Policies and other resources in the Help Center to learn more about how to keep their child safe on X. **Children:** Paragraph 19(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. ***Tools that enable children and their parents/guardians to limit the public visibility of a child's accounts, including their profile and content*** As set out in the response to paragraphs 8 and 9, accounts belonging to known minors will be defaulted to "*Protected posts*". This means that known minors will receive a request when new people want to follow them (which they can approve or deny), their posts will only be visible to their followers, and their posts will only be searchable by them and their followers (i.e. they will not appear in public searches). This is in addition to the other privacy settings and controls set out in paragraphs 8 and 9. ***Tools that enable children and their parents/guardians to limit who can contact and/or interact with a child's account*** As set out in the response to paragraphs 8 and 9, X maintains a variety of user controls - including interaction/privacy controls - that enable them and their parents/guardians to limit who can contact or interact with their account. ***Tools that enable children and their parents/guardians to limit location sharing*** Post location is off by default, and users need to opt in to location sharing on the service. Please see the responses to paragraphs 8 and 9 for detail on the tools provided to Singapore end-users to limit location sharing, and an explanation of how privacy is prioritised - especially for minors.
Protection for children Paragraph 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	**Please <u>describe</u> the settings/tools and <u>explain</u> how they are robust and set to more restrictive levels by default for children. Do include screenshots or evidence where possible.** **Children:** Please see the responses to paragraphs 9, 17, 18 and 19 in which X provided details of the tools that it provides to child accounts to minimise their exposure to and mitigate the impact of harmful and inappropriate content and unwanted interactions. In general, X does not have a line of business that is dedicated to children. Users on X must be aged 13 or above, and those below the age of 13 are not permitted to sign up for the service, as required by our Terms of Service. Users who do not meet the minimum age requirement have their account locked. Notwithstanding this, there are certain differentiated measures that apply if X detects that an account belongs to a minor, as described in our response to paragraphs 9, 17, 18 and 19 above. Parents and guardians are able to access the resources in our Help Center to learn more about how to keep their child's account and experience on X safe, secure and welcoming. This includes a form permitting them to report accounts holders who they suspect as being underage. X also has policies prohibiting users from promoting or encouraging harmful and inappropriate content and enforces age restriction on advertising content.

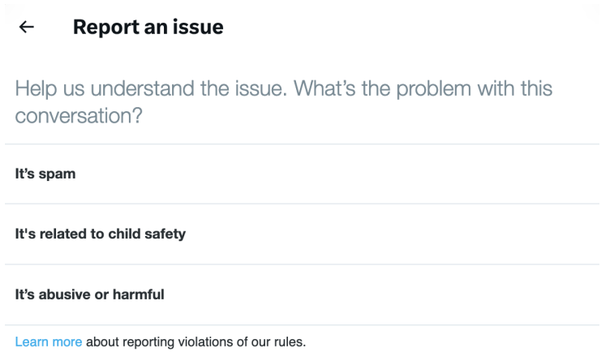
implications if they opt out of the default settings.	
Section B: User Reporting and Resolution	
Paragraph 23: Any individual must be able to report concerning content or unwanted interactions to the Service in relation to the categories of harmful and/or inappropriate content in paragraphs 4 and 17. In this regard, the reporting and resolution mechanism provided to end-users must be effective, transparent, easy to access, and easy to use.	**Please provide information on the measures, and include screenshots or evidence where possible.** **All End-users in Singapore:** **Paragraph 23(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.** Users can effectively and easily report alleged violative content in relation to the categories of harmful and/or inappropriate content in paragraphs 4 and 17 of the Code through the dedicated reporting tools. These reporting tools are easily accessible in-app through options alongside individual items of content, account profiles, DMs, Spaces, Livestreams and Lists, which enables users to directly report alleged violative content or behaviour when they encounter it. Users can also report harmful content through a dedicated Help Centre reporting form. The reporting tools are intuitive and enable users to submit reports that are sufficiently precise. The Help Centre contains clear step-by-step instructions on how to report violations of the X Rules through each of the available reporting tools, as iterated above. When a user reports an item of content, they will be notified that X has received the report and the content will be obscured from them behind a notice. The reporting tools available to users include: ● *Directly reporting an individual post:* If a user wants to report a post, they can do so via the overflow icon (i.e. 'three-dot menu'), through which they can follow an intuitive process to select the type of alleged violation. For example, under the 'child safety' option, users can select from a number of specific options including 'selling or distributing sexually explicit content involving a minor', 'sexualisation of minors', 'grooming or online enticement of minors', 'child sex trafficking', 'minor at risk', among more. ● *Directly reporting Direct Messages*: Users can either report an individual Direct Message via the information icon alongside the individual message that they would like to report, or an entire conversation via the three-dot menu option in the relevant conversation. In the case that a user reports an individual message, X may also ask the user to select additional messages to report to provide appropriate context for evaluating the report. ● *Directly reporting an account profile*: Users can report specific accounts by opening the profile they would like to report, selecting the three-dot menu, selecting 'report', and then selecting the type of issue they would like to report. X may ask the user to also select specific posts from the account that they are reporting so that we have better context to evaluate the report. ● *Directly reporting an X Space or a person in a Space*: Speakers and listeners can report a Space and any account in a Space. In order to report a Space, users can select the three-dot menu and tap 'Report this Space'. They will then be able to select the type of issue they would like to report. In order to report an account in a Space, users can tap on the relevant account's profile photo and then select 'Report'. They will also need to select the type of issue to report. ● *Directly reporting a Livestream*: Users can report livestream both as a live broadcast and as a replay, with reporting options available at the broadcast and post level, through which users can follow an intuitive process to select the type of alleged violation. ● *Directly report a List*: Users can report a List by selecting the three-dot menu at the top of the list. This process will guide the user to provide relevant information for the purpose of assessing their report. ● *Help Centre*: Users can also report alleged violative content through our Help Centre reporting form. X ensures that user reports are assessed, and that appropriate action is taken, in a timely and diligent manner that is proportionate to the severity or imminence of potential harm. As detailed below, reports of CSEM are triaged and escalated to X's dedicated enforcement teams, which has procedures for prioritising the review of any reports where minors are indicated to be at potential risk. Similarly, all content moderators are trained to triage or escalate reports where relevant to the dedicated Violent and Hateful Entities content moderation team (including terrorist entities). Where a violation is found under the X Rules, X will take action in line with the applicable X Rules policy. This may include applying an enforcement action as outlined in the response to paragraph 8 above (e.g. account suspension, content removal, restricted reach, etc). To help ensure the timeliness and diligence of its handling of user reports, X provides operational guidance and training to its content moderators through regularly updated training materials. Training is frequently refreshed, and guidance is updated in response to evolving on-platform behaviours and/or significant events. X continuously reviews the effectiveness of its reporting systems and processes to ensure that reports are assessed and that violative content is removed or otherwise actioned (if necessary) within reasonably expeditious timeframes, based on the level of harm the material poses to users.

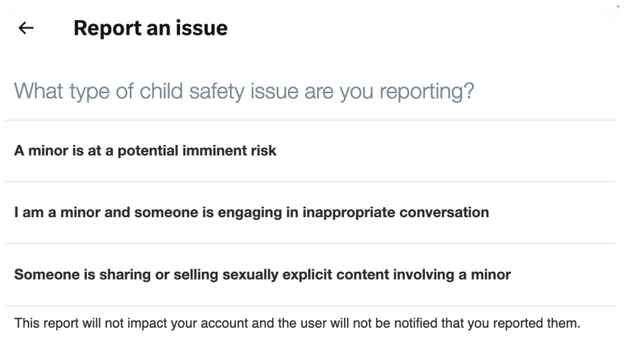


As previously stated, in the reporting period, X implemented an important enhancement to its DM reporting flow to enable users to more clearly and intuitively report certain child safety issues that may arise specifically in DMs. While users were previously able to report violations of X's Child Safety policy in DMs, X has introduced new reporting pathways including '*I am a minor and someone is*

engaging in inappropriate conversation' and '*Someone is sharing or selling sexually explicit content involving a minor*'. This improvement has ensured that reports are immediately triaged to the correct internal teams and prioritised for swift review, investigation and enforcement.

Please see screenshots of X's improved child-friendly DM reporting pathway below:





Further detail on X's reporting tools is set out in the responses to paragraphs 8 and 9.

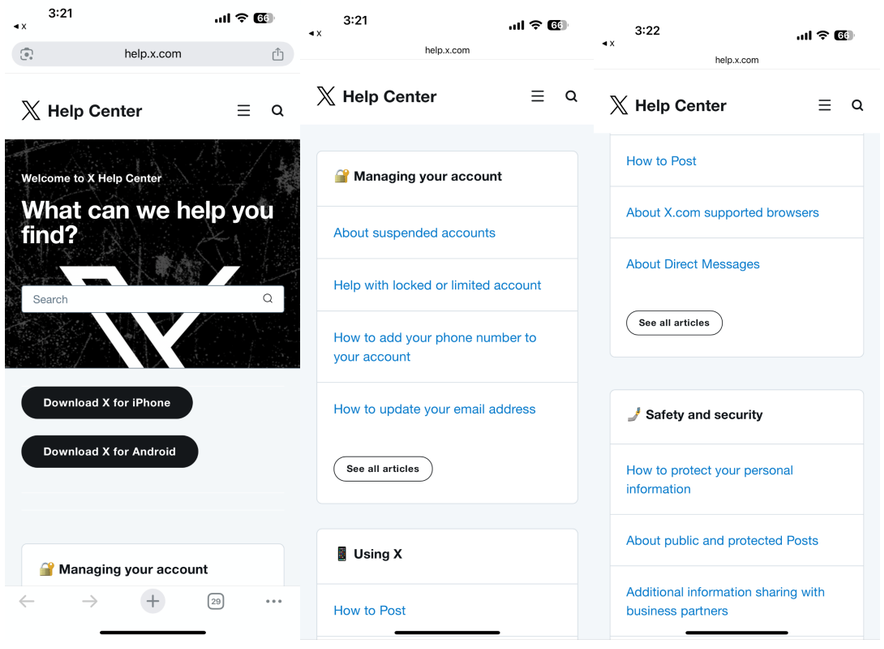
Please provide information on the measures, and include screenshots or evidence where possible.

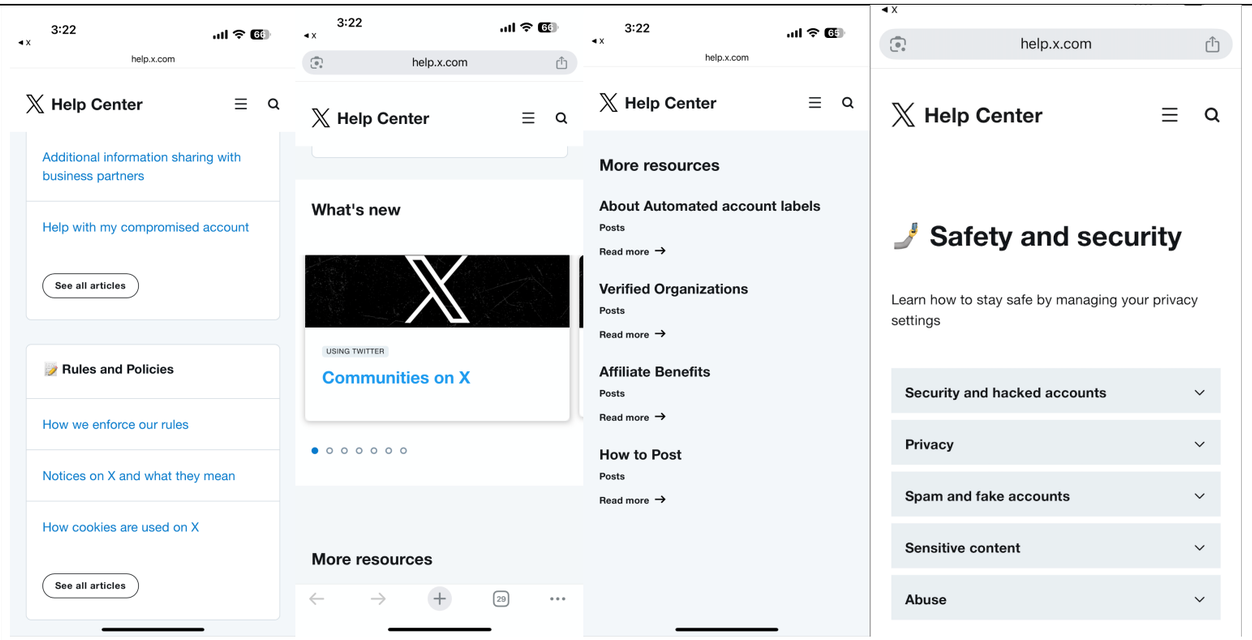
All End-users in Singapore:

Paragraph 23(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 report 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.

Following standard process, user notices are sent with the outcome of the report once the review process is complete and the appropriate remediation is applied. The Report(er) is sent a notice via in-app notification or an email that the content they reported was found 'violative' or 'not violative'. If the content/account is found 'violative', the report(ed) user receives notice of violation via in-app notification or email with an invitation to appeal (if eligible) and the reported user can submit an appeal from the in-app appeal form or the Help Center webform. User notices are sent with the outcome of the appeal once the review process is complete and the appropriate remediation is applied.

	Please provide information on the measures, and include screenshots or evidence where possible. **All End-users in Singapore:** **Paragraph 23(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.** Please refer to the responses to paragraph 23(b) above.
Section C: Accountability - Mandatory Information and Metrics	
Paragraph 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.	**Please provide information on the measures, and include screenshots or evidence where possible.** **All End-users in Singapore:** X has a location on the service that is dedicated to providing clear and accessible online safety information and resources specific to all end-users in Singapore. Please refer to the Help Centre which contains clear and easily comprehensible information that enables users to assess the level of safety and related safety measures provided by X, and to make informed choices about their experience on the platform. The Safety and Security section of the Help Centre offers practical guidance on topics such as managing account access, recognizing and reporting abuse, handling sensitive media, and adjusting privacy and visibility settings. It also provides tools and step-by-step instructions for controlling what personal information—such as location data—is shared, helping users take informed and proactive steps to protect their experience and safety on X. See screenshots below: 

	
Paragraph 25(b): How much and what types of harmful or inappropriate content end-users in Singapore encounter on the Service	*Please provide information and insert data metric(s), broken down by the content and end-user categories.* Please refer to the data supplied in response to paragraph 26 below.

Section C: Accountability - Additional Information and metrics	
Paragraph 26(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.	**Please provide information and insert data metric(s) that IMDA has agreed to, broken down by the content and end-user categories.** **1. Sexual content** The number and types of end-user reports received from end-users in Singapore: Non-consensual Nudity: 133 Sensitive Media: 7.6K Child Sexual Exploitation: 3.1M[4], 95.73% of which were received from accounts subsequently suspended, including for spam[5] The number and types of harmful and inappropriate content removed as a result of end-user reports

[4] Higher end-user report volumes do not necessarily precede or result in higher action rates. Genuine CSE will be actioned. However, users often submit incorrect and duplicate reports, including reporting content under the CSE policy that is not genuine CSE or reporting content under the CSE policy that is in fact violative of a different policy. If content reported by end-users under the CSE policy is ultimately enforced under a different policy, the enforcement action will be recorded under the other relevant policy rather than under the CSE policy. Furthermore, it is also possible that genuine CSE may be incorrectly reported under another incorrect/irrelevant policy, but subsequently reviewed and actioned by X under the CSE policy.

[5] 95.73% of these reports were received from accounts subsequently suspended (for any reason), including for spam. In respect of those suspended for spam, our systems detected intentional and co-ordinated efforts by spam groups to use our reporting systems in an attempt to target rival spam groups and/or to overwhelm our reporting systems with the intention of obstructing our Safety agents from reviewing and actioning genuine reports. The spam groups submitted scripted inauthentic reports of non-CSE material via our CSE reporting mechanisms at very significant volumes, up into the millions. This immediately generated highly irregular reporting behaviour in relation to which X was able to identify clear discernible patterns associated with spam accounts. Accounts determined to be engaging in spam reporting were suspended under the relevant X policies regulating spam, including' Authenticity'.

Non-consensual Nudity: 5.3K[6]
Sensitive Media: 1.8K
CSE: 26[7]

2. Violent Content

The number and types of end-user reports received from end-users in Singapore:

Violent & Hateful Entities: 44.1K[8]
Violent Content: 297.7K[9], 60.36% of which were received from accounts subsequently suspended, including for spam

The number and types of harmful and inappropriate content removed as a result of end-user reports

Violent & Hateful Entities: 0[10]
Violent Content: 0[11]

3. Suicide and self-harm content

The number and types of end-user reports received from end-users in Singapore:

Suicide & Self Harm: 9.9K

The number and types of harmful and inappropriate content removed as a result of end-user reports:

Suicide & Self Harm: 514

4. Cyberbullying content

The number and types of end-user reports received from end-users in Singapore:

Abuse & Harassment: 558.3K[12], 30.72% of which were received from accounts subsequently suspended, including for spam
Hateful Conduct: 540.6K, 39.76% of which were received from accounts subsequently suspended, including for spam
Private Content: 74.5K

The number and types of harmful and inappropriate content removed as a result of end-user reports:

Abuse & Harassment: 4.2K
Hateful Conduct: 79
Private Information & Media: 316

5. Content endangering public health

[6] The volume of content removed as a result of SG end-user reports can be higher than the volume of SG end-user reports received because reports submitted under one policy may ultimately be actioned under a different policy. Additionally, a single report may lead to enforcement against multiple items of content. In the case of the data reported for Non-Consensual Nudity, reports submitted under a different policy were ultimately actioned under the non-consensual nudity policy. Accordingly, the volume of content removed following end-user reports under the non-consensual nudity policy was ultimately higher than the number of reports received under that policy.
[7] X primarily enforces the 'Child Sexual Exploitation' policy at the account level by permanently suspending violative accounts. As below, X suspended almost 5 million global accounts and 28K Singapore accounts in the reporting period.
[8] X does not have any control over the number of end-users who submit user reports, nor the volume of reports that they submit. Changes in end–user reporting behaviour may be explained by a myriad of variables such as increased awareness of reporting mechanisms or off-platform events that trigger increased proclivity to report content. Higher end-user report volumes do not necessarily precede or result in higher action rates. Genuine VHE content and accounts reported by end-users under the VHE policy will be actioned. However, users often submit incorrect reports, including reporting content under the VHE policy that is not genuine VHE or by reporting content under the VHE policy that is in fact violative of a different policy. If content reported by end-users under the VHE policy is ultimately enforced under a different policy, the enforcement action will be recorded under the other relevant policy rather than the VHE policy.
[9] As above, X does not have any control over the number of end-users who submit user reports, nor the volume of reports that they submit. Changes in end-user reporting behaviour may be explained by a myriad of variables such as increased awareness of reporting mechanisms or off-platform events that trigger increased proclivity to report content.
[10] X primarily enforces the 'Violent & Hateful Entities' policy at the account level by permanently suspending violative accounts. As below, X suspended 104K global accounts and 336 Singapore accounts under the 'Violent & Hateful Entities' policy, in the reporting period.
[11] Higher end-user report volumes do not necessarily precede or result in higher action rates, as enforcement will depend on the accuracy of the reports received. Reported content will only be actioned if it is determined to be violative of the Violent Content policy.
[12] As above, X does not have any control over the number of end-users who submit user reports, nor the volume of reports that they submit. Changes in end-user reporting behaviour may be explained by a myriad of variables such as increased awareness of reporting mechanisms or off-platform events that trigger increased proclivity to report content.

	The number and types of end-user reports received from end-users in Singapore: Illegal and Regulated Behaviours: 117 Financial Scam: 0 Synthetic & Manipulated Media: N/A[13] The number and types of harmful and inappropriate content removed as a result of end-user reports Illegal and Regulated Behaviours*: 0 Financial Scam*: 0[14] Synthetic & Manipulated Media*: N/A **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime* **6. Content facilitating vice and organised crime** The number and types of end-user reports received from end-users in Singapore: Illegal and Regulated Behaviours: 117 Financial Scam: 0 Synthetic & Manipulated Media: N/A The number and types of harmful and inappropriate content removed as a result of end-user reports: Illegal and Regulated Behaviours*: 0 Financial Scam*: 0 Synthetic & Manipulated Media*: N/A **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime*

[13] Users cannot submit reports under this policy. X detects proactively.
[14] X does not enforce this policy with content removal. Alternatively, violators may be sanctioned with anti-spam challenges, restricted reach, temporary loss of access to X features or products, profile modifications, or account suspension. As below, X suspended 22,658 global accounts and 62 Singapore accounts under 'Financial Scam' in the reporting period.

Paragraph 26(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.	**Please <u>provide information</u> and the <u>median/average time taken.</u> Where possible, please provide the breakdown by content and end-user categories.** Median time between receiving and actioning/closing Singapore user reports (aggregate for relevant X Rules policies): 69 hours Median time between receiving and actioning/closing Singapore user reports ('Child Sexual Exploitation'): 4.7 hours

Paragraph 26(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.	**Please <u>provide information</u> and <u>insert data metric(s)</u> that IMDA has agreed to, broken down by the content and end-user categories.** **1. Sexual content** *<u>Accessible by end-users in Singapore</u>* Non-Consensual Nudity: 23K Sensitive Media: 769.1K CSE: 2.5K[15] *<u>Originated from Singapore</u>* Non-Consensual Nudity: 59 Sensitive Media: 5.2K CSE: 8 **2. Violent Content** *<u>Accessible by end-users in Singapore</u>* Violent & Hateful Entities: 0[16] Violent Speech: 1.4M *<u>Originated from Singapore</u>* Violent & Hateful Entities: 0 Violent Content: 2.4K **3. Suicide and self-harm content** *<u>Accessible by end-users in Singapore</u>* Suicide & Self Harm: 131 *<u>Originated from Singapore</u>* Suicide & Self Harm: 0 **4. Cyberbullying content** *<u>Accessible by end-users in Singapore</u>* Abuse & Harassment: 2.3K Hateful Conduct: 2.2K Private Information & Media: 3.7K *<u>Originated from Singapore</u>* Abuse & Harassment: 1 Hateful Conduct: 5 Private Information & Media: 2

[15] X primarily enforces the 'Child Sexual Exploitation' policy at the account level by permanently suspending violative accounts. As below, X suspended almost 5 million global accounts and 28K Singapore accounts in the reporting period.

[16] The Violent and Hateful Entities policy is overwhelmingly enforced at the account level, with only rare instances of individual pieces of content being actioned directly under this policy. X considers violations of this policy to be appropriately severe to necessitate the suspension of violative accounts in totality, rather than the removal of individual pieces of violative content. X emphasises that there is no place on X for violent and hateful entities, and notes that the behaviour and activity these entities engage in and/or promote jeopardises the physical safety of those targeted. For this reason - and as was the case during the 2024 reporting period - the volume of content removed under this policy is either very low or zero, while the volume of accounts suspended is significant. As below, X suspended 101.8K global accounts and 257 Singapore accounts under the 'Violent & Hateful Entities' policy in the reporting period.

	5. Content endangering public health *Accessible by end-users in Singapore:* Illegal and Regulated Behaviours: 143 Financial Scam: 0 Synthetic & Manipulated Media: 0 *Originated from Singapore:* Illegal and Regulated Behaviours*: 2 Financial Scam*: 0 Synthetic & Manipulated Media*: 0 **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime* **6. Content facilitating vice and organised crime** *Accessible by end-users in Singapore:* Illegal and Regulated Behaviours: 143 Financial Scam: 0 Synthetic & Manipulated Media: 0 *Originated from Singapore:* Illegal and Regulated Behaviours*: 2 Financial Scam*: 0 Synthetic & Manipulated Media*: 0 **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime*

Paragraph 26(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.	**Please <u>provide information</u> and <u>insert data metric(s)</u> that IMDA has agreed to. Where possible, please provide the breakdown by content categories.** **1. Sexual content** *<u>Accessible by end-users in Singapore</u>* Non-Consensual Nudity: 114.3K Sensitive Media: 31.8K CSE: 4.9M *<u>Originated from Singapore</u>* Non-Consensual Nudity: 837 Sensitive Media: 44 CSE: 17.8K **2. Violent Content** *<u>Accessible by end-users in Singapore</u>* Violent & Hateful Entities: 101.8K Violent Content: 157.4K[17] *<u>Originated from Singapore</u>* Violent & Hateful Entities: 257 Violent Content: 419 **3. Suicide and self-harm content** *<u>Accessible by end-users in Singapore</u>* Suicide & Self Harm: 3.3K *<u>Originated in Singapore</u>* Suicide & Self Harm: 11 **4. Cyberbullying content** *<u>Accessible by end-users in Singapore</u>* Abuse & Harassment: 1.98M Hateful Conduct: 4.7K Private Information & Media: 4.6K *<u>Originated from Singapore</u>* Abuse & Harassment: 2.6K[18] Hateful Conduct: 3 Private Information & Media: 18

[17] X constantly evaluates its policies and enforcement protocols to ensure that enforcement action is proportionate to the specific policy violation. During the 2025 reporting period, X expanded the range of graduated enforcement measures available for Violent Content policy violations. This update allows, in appropriate circumstances, for additional enforcement steps before an account suspension measure is applied, ensuring that violative content is removed while also giving users the opportunity to review the relevant rules, understand the nature of the violation, and adjust their conduct accordingly. These measures are intended both to protect against violative content and to educate users, supporting long term compliance. The introduction of these intermediate measures lowered the number of accounts permanently suspended during the 2025 reporting period.

[18] Increases or reductions in the volume of accounts enforced under a particular policy may result from a number of considerations, including a higher volume of violative activity occurring on the platform, a higher volume of accurate end-user reports received, or improvements to X's proactive detection and enforcement systems.

	5. Content endangering public health *Accessible by end-users in Singapore:* Illegal and Regulated Behaviours: 1.06M Financial Scam: 22.6K Synthetic & Manipulated Media: 3 *Originated from Singapore:* Illegal and Regulated Behaviours*: 1.5K Financial Scam*: 42 Synthetic & Manipulated Media*: 0 **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime* **6. Content facilitating vice and organised crime** *Accessible by end-users in Singapore:* Illegal and Regulated Behaviours: 1.06M Financial Scam: 22.6K Synthetic & Manipulated Media: 3 *Originated from Singapore:* Illegal and Regulated Behaviours*: 1.5K Financial Scam*: 42 Synthetic & Manipulated Media*: 0 **This policy covers both content which could be considered to endanger public health and content which could be considered to facilitate vice and organised crime*