**Public Consultation on Enhancing Online Safety For Users in Singapore:**
**AWARE's submissions to the government consultation**

The Association of Women for Action and Research (AWARE) welcomes the government's upcoming efforts to tackle online harms. As our internet usage increases and children get exposed to online content at an earlier age, there is strong impetus to ensure that online spaces are regulated and safe for all users.

AWARE is particularly concerned about the upward trend in the number of technology-facilitated sexual violence (TFSV) cases in recent years. Since 2016, AWARE's Sexual Assault Care Centre (SACC) has supported 747 clients who experienced TFSV; such cases constitute an average 17% of all SACC cases annually.[1] Over the years, SACC staff have observed the emotional, mental and physical toll on victim-survivors of TFSV, linked to their loss of dignity, privacy and sexual autonomy. This is exacerbated by the struggle victim-survivors face to contain the spread of content once it is uploaded and shared, primarily by requesting platforms to facilitate take-downs.

Several social media companies have already implemented measures to address online harms on their platforms. Instagram and TikTok, for instance, give new users under certain ages—16 (or 18 in certain countries) for Instagram and 15 for TikTok—private accounts by default.[2] Instagram also prevents some accounts from interacting with young users, such as those owned by adults that have recently been reported for potentially suspicious activity. Yet another measure taken by Instagram involves preventing advertisers from targeting advertisements to people under 18 (or older in certain countries) based on their interests or other online activity. Instead, advertisers can only target advertisements to this age group based on their "age, gender and location".

In recent years, the Singapore Government has stepped up its efforts to close the digital safety gap. Most notably, the Ministry of Communications and Information (MCI) launched the Alliance for Action (also known as the Sunlight AfA) to tackle online harms, especially those targeted at women and girls. The Sunlight AfA has brought together industry leaders, community groups, and academics to strengthen online safety. Since July 2021, the Sunlight AfA has launched initiatives including a sensing poll of more than 1,000 Singaporeans that provided a baseline assessment of the scale and scope of online harms experienced in Singapore, as well as a pilot

---

[1] "AWARE saw 36% increase in cases of technology-facilitated sexual violence in 2020; announces launch of Solid Ground website", AWARE, 14 July 2021, https://www.aware.org.sg/2021/07/technology-facilitated-sexual-violence-2020-launch-solid-ground-website-survivors; "Image-based sexual abuse featured in 7 in 10 cases of technology-facilitated sexual violence seen by AWARE in 2021", AWARE, 20 April 2022, https://www.aware.org.sg/2022/04/image-based-sexual-abuse-featured-in-7-in-10-cases-of-technology-facilitated-sexual-violence-seen-by-aware-in-2021

[2] "Giving Young People a Safer, More Private Experience", Instagram, Meta Platforms, Inc., 27 July 2021, https://about.instagram.com/blog/announcements/giving-young-people-a-safer-more-private-experience; "Teen privacy and safety settings", TikTok, ByteDance Ltd., accessed on 8 August 2022, https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/privacy-and-safety-settings-for-users-under-age-18

online harms workshop for students in institutes of higher learning.[3] A new charity, SG Her Empowerment Limited (SHE), was also established by several members of the AfA in early 2022, in a bid to sustain efforts to tackle online harms in the longer term.

The government's existing and upcoming measures aimed at tackling online harms are welcome. These efforts will place Singapore among a select group of countries where online regulation measures have already been adopted or are pending, including the United Kingdom, Germany, Australia, Ireland, Canada and the European Union. We look forward to the Code of Practice for Online Safety and the Content Code for Social Media Services further cementing our efforts to make digital spaces safer.

**General comments and recommendations**

Since online harms constitute a relatively underdeveloped area of law, we recommend that the Code of Practice clearly articulate its regulatory vision and the principles that lie at the heart of this vision. The vision and principles should be set out right at the start of the Code of Practice and should inform everything that follows. The principles should embody the spirit of the Code of Practice, and compliance with the spirit of these principles should be a fundamental building block for good online harms regulation practice.

Singapore can refer to existing legislations, such as the UK's Online Safety Bill, which states that its principles are:[4]

- the importance of protecting the right of users and (in the case of search services or combined services) interested persons to freedom of expression within the law, and
- the importance of protecting the privacy of users.

Similarly, the principles of New Zealand's drafted Code of Practice for Online Safety and Harms include:[5]

- the promotion of safety;
- respect for freedom of expression and other fundamental human rights;
- the protection of user privacy;
- the recognition of the transnational nature of the internet;
- broad applicability and participation;
- systems-based best practice standards;
- proportionality and necessity; and
- whole-of-society collaboration and cooperation.

---

[3] "Sunlight AfA celebrates a year's work in tackling online harms", Ministry of Communications and Information, Government of Singapore, 27 July 2022, https://www.mci.gov.sg/pressroom/news-and-stories/pressroom/2022/7/sunlight-afa-celebrates-a-year-work-in-tackling-online-harms

[4] House of Commons, *Online Safety Bill* (UK: House of Commons, 2022), 179 https://publications.parliament.uk/pa/bills/cbill/58-03/0121/220121.pdf

[5] Netsafe, *Aotearoa New Zealand Code of Practice for Online Safety and Harms* (New Zealand: Netsafe, 2021), 7-8, https://www.netsafe.org.nz/wp-content/uploads/2021/12/Aotearoa-New-Zealand-Code-of-Practice-for-Online-Safety-and-Harms-public-feedback-draft.pdf

| Paragraph no. | Proposed measures | AWARE's comments and recommendations |
|---|---|---|
| 9(a) | Code of Practice for Online Safety: Designated social media services with significant reach or impact will be required to have appropriate measures and safeguards to mitigate exposure to harmful online content for Singapore-based users. These include system-wide processes to enhance online safety for all users, and to have additional safeguards for young users. | Although the Code of Practice doesn't clearly state the process through which social media services will be designated, nor the thresholds that will be applied to assess what constitutes "significant reach or impact", we recommend that the Code be applied to *all* online content providers regardless of their reach or impact. This is for two reasons:<br><br>1) The harm sustained by individual victim-survivors doesn't discriminate on the basis of the reach or impact of the platform where the offensive content was hosted.<br><br>2) Perpetrators often escape detection by first posting the harmful content on a small, relatively unknown platform, and then sharing a link to this content on a major platform. Since the link itself is not "harmful or offensive", they thus circumvent social media policies and community standards. For example, SACC worked on a case where the perpetrator shared a link to a non-consensual intimate image on a popular platform. Because the image itself was not posted on the platform, this action was not considered a breach of the platform's social media standards.<br><br>Other jurisdictions seem to have adopted a similarly expansive approach to the one we suggest. Australia's Online Safety Act covers social media services, relevant electronic services (such as email service providers and messaging platforms), designated internet services (such as websites) and search engines, amongst other service providers.[6] UK's Online Safety Bill is set to apply to user-to-user service providers (such as social media platforms and online forums); providers of search service (such as search engines that host user-generated content); as well as providers of regulated services including pornographic content.[7] |
| 10 | <u>User Safety</u>:<br>We are considering requiring designated social media services to have community standards for the following categories of content:<br>a. Sexual content<br>b. Violent content | We are glad to see the government recommend that social media services develop community standards for a wide range of content categories. However, left undefined, the content categories—even when read in conjunction with the illustrative examples (Annex A)—appear to leave some harmful behaviours unregulated. |

[6] *Online Safety Act 2021* (Australia), https://www.legislation.gov.au/Details/C2021A00076
[7] House of Commons, *Online Safety Bill,* 2.

| | | |
|---|---|---|
| | c. Self-harm content<br>d. Cyberbullying content<br>e. Content endangering public health<br>f. Content facilitating vice and organised crime | a. <u>Expand "sexual content" to include all non-consensual intimate content</u><br><br>Most social media services with community standards on non-consensual sharing of certain types of content target "intimate" content, not just "sexual" content. E.g. earlier this year, Meta launched a free tool to support victims of non-consensual intimate image abuse.<br><br>Australia's Online Safety Act 2021 also focuses on intimate images, in acknowledgment of the fact that not all intimate images are sexual in nature, but they nevertheless require regulation. The Act defines intimate images as the depiction of private parts, private activity and/or a person without attire of religious or cultural significance, with further specifications of circumstances in which material may constitute an intimate image.<br><br>Relatedly, the inclusion of content relating to or encouraging sexual offences under the Penal Code, the Children and Young Persons Act, and the Women's Charter (such as sexual communication with minors and the non-consensual distribution of voyeuristic and intimate images) under Annex A is a step in the right direction. However, it is currently unclear whether this category will cover *all* non-consensual sexual and intimate messages sent between adults. At AWARE, we broadly categorise TFSV cases into two categories: image-based and contact-based. Unwanted sexual messages and/or comments, as mentioned above, constitute contact-based TFSV. On the other hand, image-based sexual abuse (IBSA) is an umbrella term for various behaviours involving sexual, nude or intimate images or videos of another person, including the following:<br>● the non-consensual creation or obtainment of sexual images: including sexual voyeurism acts such as upskirting, hacking into a victim's device to retrieve such images, and/or the creation of such images via deepfake technology<br>● the non-consensual distribution of sexual images: sometimes known colloquially as "revenge porn", whereby images shared willingly by a partner or ex-partner are then disseminated to others without the subject's consent<br>● the non-consensual viewing of sexual images: whereby a victim is made to view sexual content, such as pornography or dick pics, unwillingly, e.g. over message or email |

| | | |
|---|---|---|
| | | ● sextortion: whereby sexual images of a victim, obtained with or without consent, are used as leverage to threaten or blackmail that victim, in order to solicit further images and/or sexual practices, money, goods or favours<br>● others: including the capturing of publicly available, non-sexual images which are then non-consensually distributed in a sexualised context, e.g. with sexual comments and/or on a platform known for sexual content, such as the "SG Nasi Lemak" genre of Telegram group.<br><br>Threats to commit sexual violence, including sexual assault or any of the above behaviours listed under image-based TFSV, should also fall within this category. These threats, e.g. to leak another's intimate images online without their consent, can be as significant to the victim-survivor as actual sharing of such material, since these individuals similarly experience fear and a loss of sense of control. A 2016 survey of more than 1,500 of victim-survivors of sextortion in the United States found that 24% of respondents saw a mental health or medical practitioner after experiencing sextortion, with one respondent saying that she "had to go to therapy because it took so much out of [her] mentally."[8] A 2019 survey of 4,274 Australian respondents also showed that 80% of victim-survivors who had experienced threats to distribute an intimate image reported high levels of psychological distress, consistent with a diagnosis of moderate to severe depression and/or anxiety disorder.[9]<br><br>In 2021, SACC had a client who had been recorded undressing on a video call with a dating app contact, without the client's knowledge or consent. The contact then threatened to share the video link on social media if the client did not immediately transfer a sum of money to him. Though the client eventually managed to get the video taken down, he experienced much distress from the incident.<br><br>Such threats occur with considerable frequency: In a 2013 US survey of over 1,000 respondents, 1 in 10 had |

[8] Janis Wolak & David Finkelhor, *SEXTORTION: FINDINGS FROM A SURVEY OF 1,631 VICTIMS* (New Hampshire: Crimes against Children Research Center, 2016), 31, https://www.thorn.org/wp-content/uploads/2016/08/Sextortion_Report.pdf

[9] Nicola Henry, Asher Flynn & Anastasia Powell, *Responding to 'revenge pornography': Prevalence, nature and impacts* (Australia: Criminology Research Advisory Council, 2019), 41, https://www.aic.gov.au/sites/default/files/2020-05/CRG_08_15-16-FinalReport.pdf

| | | |
|---|---|---|
| | | had their ex-partners threaten to post their intimate images online.[10] Further, a 2019 US survey of 2,097 victim-survivors of sextortion found that nearly 1 in 3 perpetrators actually carried out or attempted to carry out threats against the victim-survivor, which mostly involved sharing explicit images of the victim.[11]<br><br>Given the prevalence and degree of harm of such threats, we call for the government to include this act of sexual violence in the Code of Practice.<br><br>b. <u>Exclusions to "sexual content" category</u><br><br>We further recommend that regulations of sexual content not be extended to the following content types when they are consensually created and disseminated:<br><br>● Lesbian, Gay, Bisexual, Transgender and Queer (LGBTQ)-related content:<br><br>Censorship of LGBTQ-related content on online platforms perpetuates stigmatisation of the community, especially when local portrayals of LGBTQ persons accessible to the general public (e.g., on free-to-air television programmes) have remained largely negative over the years. Classification of LGBTQ-related content as sexual inherently "normalises heterosexuality and reinforces negative historical associations of LGBTQ life with the illicit".[12] Additionally, it restricts LGBTQ individuals' access to expression, representation and sources of income, given the growing monetisation of content on online platforms.[13]<br><br>● Adult disclosures of sexual violence:<br><br>Victim-survivors of sexual violence may take to social media to disclose their experiences for |

[10] Business Wire, "Lovers Beware: Scorned Exes May Share Intimate Data and Images Online", Business Wire, Business Wire, 4 February 2013, https://www.businesswire.com/news/home/20130204005437/en/Lovers-Beware-Scorned-Exes-May-Share-Intimate-Data-and-Images-Online

[11] Alice Gold & Melissa Perrot, Sextortion Summary findings from a 2017 survey of 2,097 survivors (New Hampshire: Crimes against Children Research Center, 2019), 9, https://www.thorn.org/wp-content/uploads/2019/12/Sextortion_Wave2Report_121919.pdf

[12] Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen & Rob Cover, "Restricted modes: Social media, content classification and LGBTQ sexual citizenship*", New media & society* 23, no. 5 (2021): 920–938, https://journals.sagepub.com/doi/pdf/10.1177/1461444820904362

[13] Ibid

| | | |
|---|---|---|
| | | various reasons: to seek solidarity, and/or raise awareness about the issue, which can be helpful in their healing process. For instance, the #MeToo hashtag provided victim-survivors with a sense of catharsis and community and shed light on the extent and gravity of the issue of sexual violence.[14] In some cases, victim-survivors also seek advice from other online users to help them understand if their experience constitutes sexual assault, and where they can seek help. However, given that such posts and comments sometimes contain explicit or graphic descriptions of sexual assault, these might be flagged as "harmful content". Community standards should make a clear exception for disclosures by victim-survivors. Censoring them, even unintentionally, would serve to only silence and isolate these victim-survivors further.<br><br>One way to address this would be to require victim-survivors to mark these posts or comments as "sensitive"—as currently required by Twitter—so that they are placed behind a warning message (about the nature of the content) which will require users' acknowledgement before they can be viewed.[15]<br><br>● Sexual content consensually created for subscriber-only platforms:<br><br>Subscriber-only platforms, such as OnlyFans, limit the circulation of content to a willing, paying, adult audience. The creation and distribution of sexual content to such audiences is distinctly different from acts of TFSV, such as the non-consensual sharing of intimate images with unwilling parties. The former involves an informed and enthusiastic exchange, and nobody is being hurt. There should be space for consensual adult content provision. Regulations should instead be focused on ensuring that subscribers do not non-consensually leak or share this content, especially with minors. |

[14] Ashwini Tambe, "Reckoning with the silences of #MeToo", Feminist Studies 44, no. 1 (2018), 197-203, http://www.feministstudies.org/pdf/40-49/44-1-10-News_And_Views_(Tambe).pdf.
[15] "Sensitive media policy", Twitter, Publishing Organisation, January 2022, https://help.twitter.com/en/rules-and-policies/media-policy

| | | |
|---|---|---|
| | | c. <u>Extend community standards to cover posts, comments and direct messages (DMs)</u><br><br>Content on social media platforms can take the form of posts, comments, DMs, etc. The Code of Practice should apply to all user-to-user interactions, and clearly state so. Some recent reports on online harms have highlighted the pervasiveness of abuse perpetrated through DMs, and while community standards generally apply to DMs, they don't tend to be as regulated as posts and comments on posts. For example, depending on their account's privacy settings, a social media user may be able to open a channel of communication with an individual they were not already connected with. This enables strangers to establish non-consensual contact with others, and possibly harass them.<br><br>A 2022 report by the Centre for Countering Digital Hate (CCDH) found that 1 in 15 DMs sent by strangers to high-profile women violated Instagram's Community Standards, and about 1 in 4 abusive images and videos sent was considered to be IBSA.[16] Similarly concerning findings were observed in a 2021 study by UNESCO on online violence against women journalists with 901 survey respondents from 125 countries and 173 interviewees.[17] Researchers found that of the online threats experienced, almost half came in the form of harassing DMs.[18]<br><br>d. <u>Retrieval of data of interactions between users on dating platforms, social media and messaging services that allow communications to be erased</u><br><br>Separately, the Code of Conduct should specifically require certain social media platforms that require a "match" function, such as dating websites and ride-hailing services, to be able to retrieve conversation data at the request of the victim-survivor even after the users have been unmatched.<br><br>Locally, 1 in 8 TFSV cases seen by AWARE's SACC in 2020 were perpetrated by dating app contacts.[19] Data in |

[16] Centre for Countering Digital Hate (CCDH), *Hidden Hate: How Instagram fails to act on 9 in 10 reports of misogyny in DMs*, (United States: CCDH, 2022), 11-12, https://counterhate.com/wp-content/uploads/2022/05/Final-Hidden-Hate.pdf
[17] Julie Posetti, Nabeelah Shabbir, Diana Maynard, Kalina Bontcheva and Nermine Aboulez, *The Chilling: Global trends in online violence against women journalists* (UNESCO, 2021), https://unesdoc.unesco.org/ark:/48223/pf0000377223
[18] Ibid
[19] "AWARE saw 36% increase"

| | | |
|---|---|---|
| | | other countries suggests that TFSV committed against young women on dating apps may be especially prevalent: In the US, a nationally representative survey of 4,680 adults in 2019 showed that young women aged between 18 and 34 were disproportionately targeted with rude or harassing behaviours on such platforms.[20] Of this group, 57% reported being sent an unsolicited sexually explicit message or image, 44% were called an offensive name and 19% were threatened with physical harm.[21] Yet on certain apps, such as Tinder and Hinge, abusers can erase evidence of such behaviour simply by "unmatching" their victims, which will result in the entire conversation being removed.<br><br>This may hinder the reporting process for victim-survivors who hesitate to take action, having lost proof of the harassing messages received from the perpetrator. It is thus crucial that such content be documented so that victim-survivors can take action, if they wish.<br><br>In conclusion for this section, we hope that the content categories and list of illustrative examples can be kept dynamic. Given the pace of technological innovation, no such list can be confidently comprehensive, but it is our hope that it will constantly evolve, and that the government will continue to add illustrative examples to the list as newer forms of online harms surface. |
| 11 | These designated services will also be expected to moderate content to reduce users' exposure to such harmful content, for example to disable access to such content when reported by users. | On this, we hope that the government can provide additional clarification on:<br>● Whether access to such content will be disabled only for users who reported the content or for all users;<br>● How quickly access to such content will be disabled after a report has been made;<br>● How long access to the reported content will be disabled; and<br>● Whether there will be avenues for appeal for users who report content that they deem to be harmful but that was not determined as such by service providers.<br><br>In Australia, the eSafety commissioner's office can issue blocking requests or notices to online providers. While |

[20] Monica Anderson, Emily A. Vogels and Erica Turner, *The Virtues and Downsides of Online Dating* (Washington, D.C., United States: Pew Research Centre, 2020), https://www.pewresearch.org/internet/2020/02/06/the-virtues-and-downsides-of-online-dating

[21] Ibid

| | | |
|---|---|---|
| | | providers are not obligated to accede to blocking requests, the penalty for non-compliance of a blocking *notice* is 500 penalty units, which amounts up to $111,000 for individuals and up to $555,000 for companies.[22] These notices specify actions to be taken, which include blocking (a) domain names that provide access to the material; (b) URLs that provide access to the material; or (c) IP addresses that provide access to the material.[23] Each blocking notice remains in force for not more than three months, though this can be renewed if the Commissioner issues a fresh notice that comes into force immediately after the expiry of the original order.[24]<br><br>Under Section 111 of the UK's Online Safety Bill, Ofcom can similarly issue a provisional notice of contravention to a service provider if "they consider that there are reasonable grounds for believing that the provider has failed, or is failing, to comply with any enforceable requirement that applies in relation to the service", which include children's and adults' online safety.[25] This provisional notice may specify steps that the person needs to take to comply with the duty or requirement, or remedy the failure to comply with it.<br><br>a. <u>Specify timeframe for compliance with take-down orders</u><br><br>In a case seen by SACC, a client's video was uploaded on a social media platform without her consent and remained available even five months after she filed a report with the platform. Throughout the process, the lack of clarity about the timeline, specifically when she would hear back from the platform, was extremely traumatising for her. This, combined with the platform's inaction (as the video was not deemed a violation of its policies), resulted in the client expressing suicidal ideation.<br><br>Moreover, the longer harmful or offensive content stays online, the greater the risk of it being circulated further. Even if these materials are removed by the platform at a later stage, users of these platforms would have had |

[22] Josh Taylor, "How will new laws help stop Australians being bullied online?", The Guardian, 22 January 2022, https://www.theguardian.com/media/2022/jan/23/how-will-new-laws-help-stop-australians-being-bullied-online

[23] Online Safety Act 2021 (Australia)

[24] Ibid

[25] House of Commons, *Online Safety Bill*, 102-4

| | | |
|---|---|---|
| | | plenty of time to download and further circulate the content either privately or on their social media networks. Victim-survivors thus live in a perpetual state of fear, not knowing if the content is still being circulated without their knowledge, and whether it will resurface online one day. As such, we recommend the Code of Practice stipulate a timeframe within which social media services must comply with a take-down notice.<br><br>In Australia, the eSafety Commissioner can issue a removal notice to require the provider of a social media service, relevant electronic service or designated internet service to take all reasonable steps to ensure the removal of the intimate image from the service within 24 hours after the notice was given.[26]<br><br>b. <u>Take-down process should be clearly explained and easily accessible</u><br><br>Beyond the timeframe of the take-down process, a number of measures can be taken to make possible more upstream education on the exact process involved in reporting content for take-down, including the different steps involved in a social media service deciding whether or not to take down harmful and/or offensive content. The user should be able to easily understand what it means to report the content, the duration of the process and the possible outcomes that are available.<br><br>Importantly, there should also be information on other options they can consider if the outcome is not as expected, e.g. when offensive or harmful content is not removed because it doesn't breach community standards. Information on other forms of support they can access, either online or in the community, should be made available both when communicating the outcome of a user report and in the section of the website where community standards are hosted.<br><br>c. <u>Social media platforms should introduce bystander education</u><br><br>A 2021 study we conducted with technology firm Quilt.AI found that misogynistic tweets are twice as likely to be liked and 4.5 times more likely to be retweeted as compared to non-misogynistic tweets. In other words, users are not only failing to call out misogynistic content, or simply ignoring it—they are actively engaging with |

[26] Ibid

| | | |
|---|---|---|
| | | and promoting it, thus perpetuating misogynistic behaviour themselves. Liking and retweeting such content has a practical effect as it serves to reinforce and amplify harmful messages. Reasons why bystanders may engage in such behaviours include an inability to recognise the behaviours as harmful, a desire to downplay their impact or even concerns about breaching social norms of masculine solidarity by intervening.<br><br>The Code of Practice should thus require social media platforms to introduce bystander education and provide resources for users to act on harmful behaviours inflicted upon other users. It may take the form of a first responder training workshop, similar to AWARE's Sexual Assault First Responder Training, or a document or infographic that is easy for users to access and understand. |
| 12 | For child sexual exploitation and abuse material, and terrorism content, these services will be required to proactively detect and remove such content. | Apart from the listed content, the Code of Conduct should also require social media services to proactively detect and remove other types of harmful and/or offensive content (such as non-consensual intimate images) that is already illegal in Singapore.<br><br>According to Section 377BF of the Penal Code, it is an offence to non-consensually distribute an image of a user's own or someone else's genitals with the intent of the victim seeing their own or someone else's genitals, for the purpose of obtaining sexual gratification or causing the victim humiliation, alarm or distress.[27]<br><br>If the concern with extending this requirement to non-consensual intimate content is that machine learning may not be able to successfully detect consent, we could take a "better safe than sorry" approach by requiring social media services to use sensitivity layers[28] to blur all intimate images shared on the platform. Bumble, an online dating platform, introduced this feature on its app in 2019. Using its "Private Detector" technology, the platform will detect lewd images and automatically blur them when a user receives it.[29] The |

[27] *Penal Code 1871* (Singapore), https://sso.agc.gov.sg/Act/PC1871?ProvIds=pr377BF-&ViewType=Advance&Any=3+Factories+Persons+In+Charge+Regulations&WiAl=1#

[28] Referring to safety control features that mask sensitive content such that users are not automatically exposed to such content upon receiving it from another user or encountering it on their feeds.

[29] "Why am I seeing a blurred image?", Bumble, accessed on 9 Aug 2022, https://bumble.com/en/help/why-am-i-seeing-a-blurred-image

| | | |
|---|---|---|
| | | user may then choose to view, ignore or report it.[30] If a user is sending someone a photo that is suspected to contain lewd imagery, they will be reminded that sending such an image may lead to them being reported.[31]<br><br>Another example is Twitter, which places sensitive content, such as adult content, behind a warning message (about the nature of the content) and requires users' acknowledgement before it can be viewed.[32]<br>The Code of Practice should require all social media services to cover all intimate content with a sensitivity layer so that users have a meaningful choice about consuming content that's being shared publicly (i.e. on someone's wall or timeline) or privately (i.e. through direct messages). |
| 16 | These additional safeguards could include stricter community standards for young users, and tools that allow young users or parents/guardians to manage and mitigate young users' exposure to harmful content and unwanted interactions. For example, tools that:<br>a. Limit the visibility of young users' accounts to others, including their profile and content;<br>b. Limit who can contact and/or interact with accounts for young users; and<br>c. Manage the content that young users see and/or experience. | We appreciate the government's efforts to introduce additional safeguards for young users in view of their vulnerability to online harms. To strengthen these safeguards, there should be greater clarity on what it means to "manage the content that young users see and/or experience".<br><br>In 2021, Frances Haugen, a former Facebook employee, came forward with documents showing that Meta was knowingly harming children.[33] The company's internal research revealed that Meta platforms made body image issues worse for one in three teen girls (as they were being led to anorexia-related content)[34] and that teens found Instagram to be a significant factor contributing to increases in rates of anxiety and depression.[35] Alarmingly, among users who reported having suicidal thoughts, 13% in the UK and 6% in the US attributed it to Instagram.[36] Yet it was reported that Facebook was intentionally targeting teens and children |

[30] Ibid
[31] Alice, "What is Private Detector and how does it work?", Bumble, accessed on 9 Aug 2022, https://bumble.com/help/what-is-private-detector
[32] "Sensitive media policy"
[33] Dan Milmo and Kari Paul, "Facebook harms children and is damaging democracy, claims whistleblower", The Guardian, Guardian News & Media Limited, 6 October 2021, https://www.theguardian.com/technology/2021/oct/05/facebook-harms-children-damaging-democracy-claims-whistleblower
[34] Ibid
[35] Damien Gayle, "Facebook aware of Instagram's harmful effect on teenage girls, leak reveals", The Guardian, Guardian News & Media Limited, 14 September 2021, https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals
[36] Ibid

| | | |
|---|---|---|
| | | as young as eight for the Messenger Kids app, thus putting them at risk of exposure to harmful content.[37] As an especially vulnerable group, children should be protected from these online harms.<br><br>a. <u>Enforce strict minimum age requirements</u><br><br>To address this, online platforms should be required to strictly enforce their minimum age requirements when new users register an account on their platforms. Instagram, for instance, is currently testing out several methods of age verification, including using a third-party face-scanning AI on a video selfie to determine a user's age.[38] If a user is deemed to be too young for Instagram, the user will be asked to prove their age. Once their age has been verified, the video selfie will be deleted. Alternatively, the user may choose to have their mutual friends verify their age or upload an identity document.[39]<br><br>b. <u>Mandatory onboarding for young users</u><br><br>Apart from the measures proposed by the government, we call for a mandatory user onboarding on community standards, especially for young users, when new users sign up for an account with social media services. This onboarding should cover how each user should uphold respectful communications, and the consequences of breaching those standards. Users should also be given information on avenues of support should they encounter harmful content on the platform or become an online target.<br><br>c. <u>Access to pornographic websites should be 18+</u><br><br>More should also be done to prevent children from accessing pornographic websites. This may similarly involve the employment of age verification technology. In the UK, it was announced that the Online Safety Bill will be enhanced with "a new legal duty requiring all sites that publish pornography to put robust checks in place to ensure their users are 18 years old or over".[40] One |

[37] "Facebook harms children"
[38] Barbara Ortutay and Matt O'Brien, "Instagram tests using AI, other tools for age verification", The Associated Press, https://apnews.com/article/technology-artificial-intelligence-e7a5583ccfe7e1b284081db40cb2ea7c
[39] Ibid
[40] Department for Digital, Culture, Media & Sport and Chris Philp, "World-leading measures to protect children from accessing pornography online", GOV.UK, 8 February 2022, https://www.gov.uk/government/news/world-leading-measures-to-protect-children-from-accessing-pornography-online

| | | |
|---|---|---|
| | | measure that is being considered is requiring individuals to prove that they possess a credit card and are over 18 in order to access the websites.[41]<br><br>Additionally, the Online Safety Bill sets out duties of regulated providers of pornographic content to "ensure that children are not normally able to encounter… pornographic content in relation to the service (for example, by using age verification)."[42]<br><br>Similarly, the eSafety Commissioner in Australia may, by legislative instrument, declare that a specified access-control system[43] is a restricted access system. This is done with the objective of protecting children from exposure to material that is unsuitable for children, such as films classified as X 18+.[44] Singapore can take reference from these measures to curtail young users' access to pornographic websites.<br><br>d. <u>Social media services should be required to create a resource centre for young users</u><br><br>The Code of Practice should require all social media services to have an education hub for young users, parents and guardians so that they can easily access information and resources for young users to have a safe and healthy experience on the platform in question. The education hub should include information and resources on:<br><br>I. The basics of using social media, such as management of audience interaction, privacy settings and the parental and supervisory tools available to create a curated experience for young users;<br>II. How to respectfully communicate with other users on the platform, and the consequences of breaching community standards;<br>III. How harmful and/or offensive content can be reported directly or with the help of a responsible adult. |

[41] Ibid
[42] House of Commons, *Online Safety Bill,* 60
[43] An access-control system refers to a system under which:
(a) persons seeking access to the material have a password, or a Personal Identification Number, that provides a means of limiting access by other persons to the material; or
(b) persons seeking access to the material have been provided with some other means of limiting access by other persons to the material.
[44] Online Safety Act 2021 (Australia)

| | | |
|---|---|---|
| 19-21 | <u>User Reporting and Resolution</u><br><br>19. Given the sheer volume of content being created and shared on social media services, there may be instances where users come across harmful content, despite the safeguards put in place by social media services. As such, we propose for designated social media services to provide an efficient and transparent user reporting and resolution process, to enable users to alert these services to content of concern.<br><br>20. The user reporting and resolution process could:<br><br>a. Allow users to report harmful online content (in relation to the categories of harmful content outlined at para 10) to the social media service;<br><br>b. Ensure that the reporting mechanism is easy to access and easy to use.<br><br>21. As part of this process, the service should assess and take appropriate action on user reports in a timely and diligent manner. | We appreciate the government's proposition to make the user reporting and resolution process more "effective and transparent" by allowing users to report harmful content, ensuring that the reporting mechanism is user-friendly and requiring services to assess and take appropriate actions in a timely and diligent manner. These practices will make reporting harmful content more accessible to users.<br><br>Victim-survivors of online harms may face difficulties seeking help offline: In a 2021 UNESCO study on online violence against female journalists, 25% of the survey respondents reported online violence incidents to their employers, but the top responses they said they received were: no response (10%) and advice like "grow a thicker skin" or "toughen up" (9%), while 2% were asked what they did to provoke the attack.[45] At the same time, only 11% and 8% of survey respondents reported online violence incidents and took legal action respectively,[46] which suggests that there may be barriers preventing them from reporting and/or a lack of confidence in legal and judicial responses. The average woman social media user is likely to face the same obstacles in reporting and likely has fewer, if not similarly inadequate, resources to address such abuse. Social media services should thus make reporting processes more accessible to users.<br><br>a. <u>Code should clarify what it means for reporting mechanisms to be "easy to access" and "easy to use"</u><br><br>In that regard, we hope that the Code will clarify what it means for the reporting mechanism to be "easy to access" and "easy to use". In a study on how online content providers (OCPs) educate users in their policy documents about their responsibilities with respect to harmful online content and the consequences of violating them, researchers measured the accessibility and readability of related policies from OCPs across four countries.[47] A policy's accessibility was assessed on two levels: "hard to find", i.e. placed in an unexpected location and/or with non-obvious labelling, versus "easy to find", whereby the labelling is obvious and the |

[45] Posetti, et al., The Chilling
[46] Ibid
[47] Sabine A. Einwiller and Sora Kim, "How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation", *Policy & Internet* 12, no. 2 (2020): 184-206

| | | |
|---|---|---|
| | | placement is where one would expect to find it. Readability was also assessed on two levels: “readable”, whereby medium-sized fonts without illustrations and/or helpful colour scheme were used, versus “very readable”, whereby the policy was very well designed with illustrations and/or colour scheme. Furthermore, researchers looked at the inclusion, or exclusion, of illustrative examples of violating content or behaviours and a reference to laws (e.g. via a hyperlink). These tools can be adapted to assess and ensure that information about user reporting is accessible to users.<br><br>b. <u>Temporarily suspend reported content</u><br><br>In addition to these measures, we recommend that any content be temporarily suspended once a report has been filed, even if investigations have not yet commenced or are ongoing. This will help prevent further circulation of the content and reduce the amount of distress and anxiety inflicted on the victim-survivor.<br><br>c. <u>A trauma-trained personnel should reach out to the complainant to explain the outcome of their report</u><br><br>If social media platforms deem that the content is not harmful and decide to restore it, the reasons for the outcome should be clearly explained to the user who reported the content. If the report flags alleged abusive and/or violent content, a trauma-trained personnel should reach out to the user to explain the outcome, limitations of the platform and explore other options that the user can pursue. The personnel should also be empowered to facilitate referrals to external support resources, should the user request this. |
| 22 | <u>Accountability</u><br><br>We propose for designated social media services to produce annual reports on their content moderation policies and practices, as well as the effectiveness of their measures in improving user safety. These reports would be made available on the IMDA’s website for the public to view. Through these reports, users will be able to better understand how their exposure to harmful content is | We concur with the government that requesting social media companies to release annual reports on the effectiveness of their content moderation policies and practices can be one way to hold social media companies accountable for their user safety measures. In addition to this, we recommend that an ombudsman or independent regulator be appointed to assess and investigate any violation complaints as well as conduct periodic evaluations of all online content providers’ policies and practices.<br><br>In the UK, this function is carried out by the Office of Communications (or Ofcom in short) which was |

| | | |
|---|---|---|
| | reduced on the services they use. | appointed by the UK government to regulate all communications services, including broadband, mobile services, TV and radio.[48] In terms of online safety, Ofcom is responsible for overseeing and enforcing the online safety regime, that is, giving guidance on compliance to the Code, issuing information notices to understand companies' approaches to address online harms and taking enforcement action against non-compliant companies.[49]<br><br>We further recommend that the Code of Practice lay out the penalties that social media companies and other online content providers will be subject to in the event of non-compliance. Social media companies in Australia are required to remove non-consensually shared images from their platform within 24 hours after notice from the eSafety Commissioner; failure to do so will result in a fine of AUS$555,000. The government should ensure that the penalties for non-compliance are proportionate to the harm and damage resulting from the delay in take-down of such material. |

[48] "What is Ofcom?", OfCom, OfCom, accessed on 8 August 2022, https://www.ofcom.org.uk/about-ofcom/what-is-ofcom

[49] House of Commons, *Online Safety Bill; "New online safety rules – what do they mean, and what is Ofcom's role?",* OfCom, accessed on 6 July 2022, https://www.ofcom.org.uk/news-centre/2022/new-online-safety-rules-what-is-ofcoms-role

| | | |
|---|---|---|
| **23-24** | The proposed measures under the Code of Practice for Online Safety are expected to deal with most of the harmful online content that Singapore users may encounter when using designated social media services. However, there may be instances where extremely harmful content remains online in relation to:<br>● Suicide and self-harm<br>● Sexual harm<br>● Public health<br>● Public security<br>● Racial or religious disharmony or intolerance<br>(Illustrative and non-exhaustive examples of such content are at Annex B)<br><br>Given the concerns about the impact of such extremely harmful content, we propose for the Content Code for Social Media Services to allow IMDA to direct any social media service to disable access to specified harmful content for users in Singapore, or to disallow specified online accounts on the social media service from communicating content and/or interacting with users in Singapore. | As with the Code of Practice, the provision of illustrative examples of content for each of the listed categories of the Content Code for Social Media Services under Annex B is welcome. We reiterate our recommendation to set out clear definitions and parameters for each of these categories to help online service users better understand what constitutes "extremely harmful content" as compared to the content stated in Annex A.<br><br>A further recommendation on the Content Code is to include sexist speech, including that which is misogynistic, in the existing list of "extremely harmful content". Given that hateful content such as that relating to racial or religious disharmony or intolerance is covered in this Code, it is equally important to state that hate speech—and hateful content in general—towards any demographic population, including women, will not be tolerated. |
| **Additional comments** | | |
| - | Satirical and parodic content | Satire serves as a form of social commentary that is expressed in an artistic and entertaining manner. The satirical nature of a piece of content is often implied through its exaggerated depiction of the behaviour or circumstance that it is critiquing.[50] As such, it is a crucial tool that allows individuals to participate in active citizenry and highlight social issues in a way that is accessible to a wide audience.<br><br>However, nuanced depictions of satire may pose challenges to social media companies' use of artificial intelligence moderation, as satirical and parodic content may be interpreted as the very harmful or hateful |

[50] Megan LeBoeuf, "The Power of Ridicule: An Analysis of Satire," Senior Honors Projects. Paper 63., (University of Rhode Island, 2007), https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1065&context=srhonorsprog

| | | |
|---|---|---|
| | | behaviour it is in fact critiquing.<br><br>Currently, satire and parody are exempted from the definition of "falsehood" under the Protection from Online Falsehoods and Manipulation Act (POFMA).[51] The Code of Practice for Online Safety should clarify how it will determine if a piece of content is satirical or not (i.e., whether it will use the same test as POFMA).<br><br>Alternatively, the following question, suggested by journalist and novelist Will Self, may be posed as a test to determine if a work is truly satiric:[52] Whom does the material comfort, and whom does it afflict? If the material offends an already vulnerable group and does not meet its purpose as a social critique, then the content may be masquerading as satire.<br><br>Further, we hope that the Code of Practice for Online Safety will clarify what steps, if any, can be taken by social media companies to protect satirical content on their platforms. |
| - | Personal Data | All safety measures undertaken by companies in accordance with the Code of Practice should also be aligned with all aspects of the Personal Data Protection Act (PDPA) 2012.<br><br>According to the PDPA, one circumstance under which an organisation may disclose the personal data without the consent of the individual is when such disclosure is necessary to respond to an emergency that threatens the life, health or safety of the individual or another individual.[53] In the case of online harms, this could involve social media platforms sharing a young user's data with the relevant authorities for the purpose of preventing child sexual exploitation and/or abuse online. |

[51] Zhuo Tee, "Media Literacy Council apologises for Facebook post on satire being fake news", The Straits Times, Singapore Press Holdings Ltd. Co., 8 September 2019, https://www.straitstimes.com/singapore/is-satire-fake-news-media-literacy-council-post-sparks-backlash-from-netizens

[52] Paula Feldmane, "Restrictions on Satire, Parody and Caricature in the Case Law of the European Court of Human Rights" (Bachelor thesis, Riga Graduate School of Law, 2019), 21, https://dspace.lu.lv/dspace/bitstream/handle/7/50064/Feldmane_Paula.pdf?sequence=1&isAllowed=y

[53] Personal Data Protection Act 2012 (Singapore), https://sso.agc.gov.sg/Act/PDPA2012?ProvIds=Sc1-#Sc1-; Personal Data Protection Commission (PDPC) Singapore, *Advisory Guidelines on Key Concepts in the Personal Data Protection Act* (Singapore: PDPC, 2021), 53, accessed on 8 Aug 2022, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/AG-on-Key-Concepts/Advisory-Guidelines-on-Key-Concepts-in-the-PDPA-1-Oct-2021.ashx?la=en

| | | |
|---|---|---|
| - | Recourse options | The Code of Practice should clarify the appeal or recourse options available to users in the event that the outcome of their report of “harmful and/or offensive content” is unfavourable. For instance, if a victim-survivor of IBSA is informed by the social media company that there was no violation of their policies (as experienced by the SACC client above), it should be made clear whether or not an independent agency exists to which the victim-survivor can file an appeal. This function can be carried out by the ombudsman, as recommended above. |
| - | Investigating reports of online harms | In a handful of cases, SACC clients have also been informed by the Singapore police that action cannot be taken against the perpetrators because they have no jurisdiction when harmful content is (i) distributed using international Internet Protocol (IP) addresses; or (ii) hosted on platforms owned by International Unlimited Companies. Such responses from the authorities, although understandable, can be demoralising and even traumatising as victim-survivors are often left at a loss as to what other options are available to them.<br><br>Thus, apart from enacting the Code of Practice and Content Code, the legislative approach to tackling online harms can be strengthened by putting in place processes to work with other jurisdictions on cases involving international IP addresses and/or platforms. |