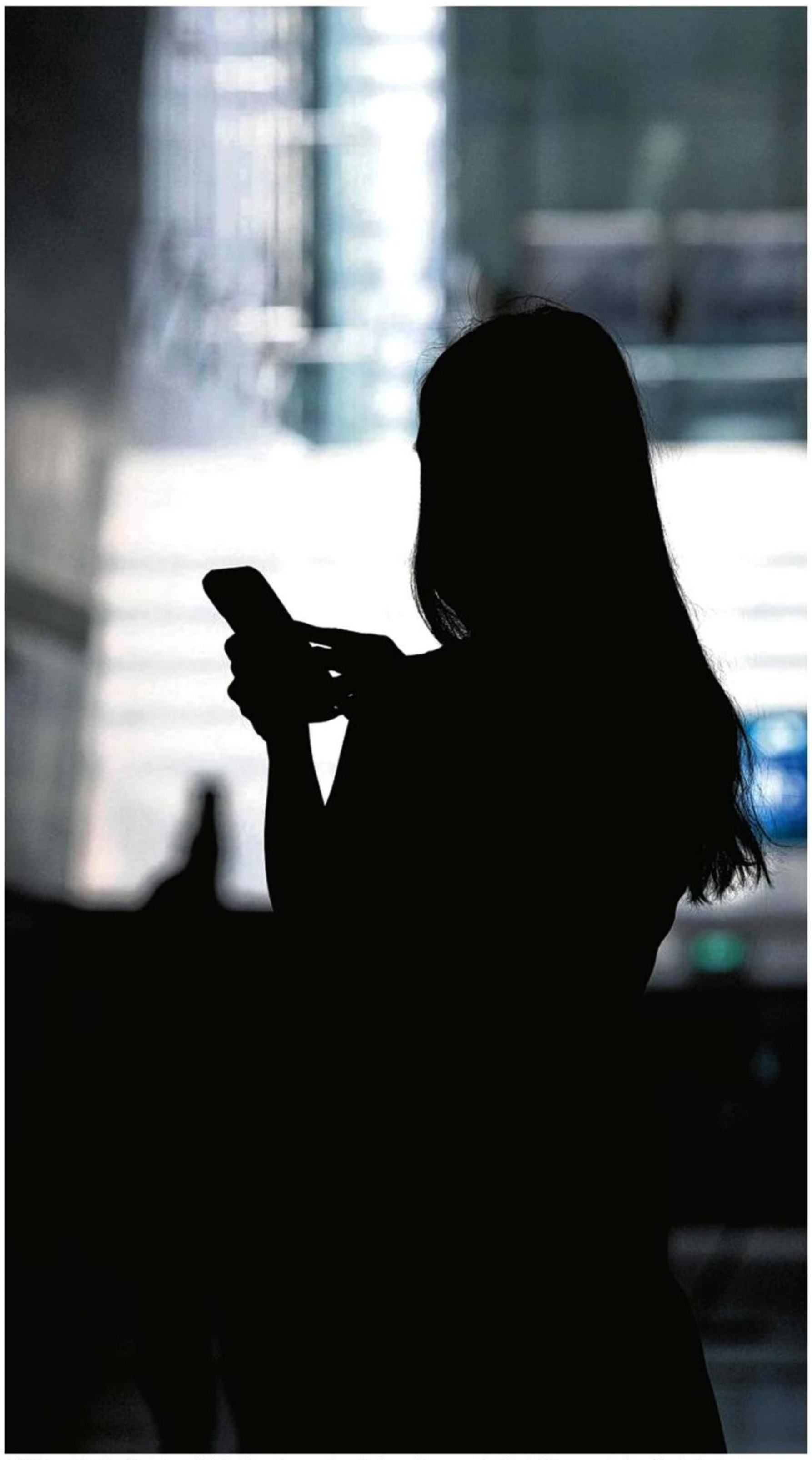
Publication: The Straits Times, pB2
Date: 27 February 2025
Headline: Social media is a minefield. Here's how we can make it safer

Social media is a minefield. Here's how we can make it safer

It's true that platforms need to be nudged to take action. Singapore has done this before, and it's time to raise the bar.



A 2024 poll by the Ministry of Digital Development and Information revealed that 66 per cent of residents had come across harmful content on Facebook, HardwareZone, YouTube, Instagram, X and TikTok. ST FILE PHOTO

Lim Sun Sun

When was the last time you saw a nasty comment on Instagram, and what did you do about it? If you just swiped past because you weren't sure what to do, you're not alone. A 2023 study by SG Her Empowerment found that 38 per cent of young people (ages 16 to 35) didn't know how to use in-app reporting tools to deal with online harm.

Or maybe you knew you could report it but didn't bother, thinking it wouldn't make a difference. A 2022 survey by the Alliance for Action to tackle online harms against women and girls found that nearly half of respondents felt reporting wouldn't help (43.6 per cent) or simply didn't know how to do it (43.4 per cent).

These numbers show that social media platforms need to step up to support users who must increasingly navigate online risks and annoyances. A 2024 poll by Singapore's Ministry of Digital Development and Information revealed that 66 per cent of residents had come across harmful content on the six designated social media services (DSMSs): Facebook, HardwareZone, YouTube, Instagram, X and TikTok. But only a quarter actually reported it.

With personal device use pushing ever younger and people spending more time online, there are growing concerns about how safe social media platforms are for young people. Short of imposing something like Australia's social media ban, which is fraught with operational challenges, what can a highly connected country like Singapore do to hold tech companies more responsible for online safety?

In fact, Singapore took a significant step forward in July 2023 when the Infocomm Media Development Authority (IMDA) issued a code of practice for social media platforms. This code requires DSMSs to address six types of harmful content, including violent and sexual material, cyber bullying, self-harm, and content that threatens public health or supports criminal activities.

These platforms must also reduce users' exposure to harmful content, especially for those under 18, provide easy-to-use tools for reporting harmful content, notify users of any action taken, and submit annual online safety reports to ensure transparency and accountability.

A BLEAK REPORT CARD

The inaugural Online Safety Assessment Report 2024 issued by IMDA on Feb 17 to mark Safer Internet Day has made accountability by these platforms more concrete. Instead of relying on self-reported data from tech companies, IMDA ran independent tests from August 2023 to July 2024 on the presence, comprehensiveness and effectiveness of their online safety measures.

IMDA created fake child accounts to see how easily they could access restricted content. They also used mystery shopper tests to check how effectively platforms handled reports of harmful content. IMDA flagged harmful content that violated these DSMSs' own community guidelines and assessed whether they took appropriate and timely action.

Furthermore, IMDA sought to determine whether these platforms proactively removed child sexual exploitation material (CSEM) and terrorism-related content. The results were eye-opening. IMDA flagged over 1,000 harmful posts across six platforms and monitored their responses. Two platforms stood out for their poor performance: X and Instagram.

X only acted on just over half of the harmful content reported and took an average of seven to nine days to remove sexual or self-harm content – sometimes up to 20 days for other violations. That's a far cry from the 15-hour median time the company claimed in its annual report. Worse yet, X allowed children's accounts to access explicit adult content, including hardcore pornography, with just a simple search. The company also claimed it proactively removed CSEM, but IMDA found more instances than X had reported.

Instagram was even more disappointing. It initially acted on only 2 per cent of harmful content reported by users, despite these posts violating its own guidelines. However, after IMDA intervened, Instagram removed the remaining 98 per cent, proving that the initial reports should have been taken seriously from the start.

IMDA also exhorted Instagram to educate users more actively on its community guidelines and address public perceptions that its user reporting systems are ineffective. In its annual report, Instagram had claimed that users "often do not understand our policies, and the majority of reports from users is content that does not violate our policies". However, IMDA's test findings revealed that the platform had not acted even on legitimate reports of violative content.

Overall however, it must be said that the social media platforms have largely established user safety measures such as community guidelines on harmful content specified by the code, supported by human and artificial intelligence (AI) moderation. Users have also been granted tools to manage their safety, such as restricting harmful content, controlling interactions, and limiting location sharing.

Platforms like Facebook, Instagram, TikTok, and YouTube have also offered accessible online safety resources and mounted awareness initiatives. Additionally, Singapore-based support was available for users searching high-risk terms related to suicide, self-harm, and other forms of violence, with some platforms providing extra mental health and digital wellness resources.

WHAT MORE CAN BE DONE?

Clearly therefore, despite the code of practice not having punitive levers, it has been able to impel the tech companies to channel more resources into online safety. And we must pause to ask: If these platforms and their profitability are built on the backs of users whose data and online activity are mined for revenue generation, don't they owe a greater duty of care to users to provide a safe and hospitable environment? Or are we to assume that because consumers use these services for free that they must "suck it up" and endure the grievances and annoyances present in these platforms?

In the current political climate of the US where these tech companies mostly originate, it would appear that AI ethics have been all but flung out the window, with online safety a likely casualty.

In the current political climate of the US where these tech companies mostly originate, it would appear that AI ethics have been all but flung out the window, with online safety a likely casualty. Tech companies such as Meta seem to have been emboldened by the Trump administration to play fast and loose with regulations and have taken their insolence to the EU where Meta appears set on defying the EU's AI Act.

In Singapore, we can afford to set a higher bar. Notably, the designated platforms have sought to provide resources for self-help and tools to enhance user safety, as stipulated by the code. But we can go further to require that these tech companies incorporate safety by design principles as a default. First, social media account feeds need to be more precisely calibrated through leveraging age assurance techniques so that exposure to age-inappropriate content is minimised.

Second, platforms should implement algorithms that detect potentially harmful content – such as graphic violence or explicit material – and display warnings so that viewers can decide whether to view them.

Third, the platforms can use AI to identify language in comments or messages that may be offensive or abusive. Smarter moderation tools can also flag harmful language to users, nudging them to rethink their words before posting.

Fourth, dynamic privacy controls could be introduced to automatically suggest tighter security settings for users who frequently interact with strangers. For instance, if a user frequently interacts with unknown individuals, the platform could suggest tightening account privacy settings to enhance security.

Fifth, platforms should make reporting harmful content as easy as liking a post, with clear explanations of why something violates community guidelines. This empowers users to participate actively in maintaining a safe environment. Some platforms already have features like these, but they're often buried in settings or turned off by default. If users don't know they exist, they won't use them.

As we wake up daily to news of ever greater technological marvels, it is confounding that a safer online space remains beyond our grasp. But as Singapore's experience shows, we don't have to wait for Big Tech to do the right thing on its own. Through sensible regulation and public pressure, we can push social media companies to take online safety seriously – because no one should have to just "deal with" harm in the digital world.

• Lim Sun Sun is vice-president, partnerships and engagement at Singapore Management University and Lee Kong Chian professor of communication and technology at its College of Integrative Studies.