RSIS COMMENTARY

25 February 2026

Policing the Feed: AI-Generated Sexual Content on Social Media and Its Impacts on the Vulnerable

Ysa Marie Cayabyab

RSiS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

RSiS 30

RSIS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

RSIS COMMENTARY

www.rsis.edu.sg No. 031 – 25 February 2026

Policing the Feed: AI-Generated Sexual Content on Social Media and Its Impacts on the Vulnerable

By Ysa Marie Cayabyab

SYNOPSIS

AI-generated sexual content and platform-driven amplification are intensifying online exploitation, disproportionately harming women and children across Southeast Asia. Existing moderation approaches remain largely reactive, leaving systemic risks unresolved. Stronger governance frameworks, platform accountability, and safety-by-design measures can help prevent harm. Through regional cooperation, ASEAN is well-positioned to strengthen coordinated safeguards, protect vulnerable users, and establish shared standards for the responsible and ethical use of AI across the region.

COMMENTARY

Over the past few months, the AI-powered chatbot Grok, hosted on the social media platform X (formerly Twitter), has faced widespread criticism. Grok could be prompted to generate non-consensual "undressing" images and other sexualised outputs of women and children, raising serious concerns about both product design and platform governance.

However, rather than disabling these capabilities outright, X has largely relied on reactive measures such as geoblocking image generation in jurisdictions where such content is explicitly illegal, restricting image creation and editing to paid subscribers, and removing offending posts or accounts after they surface. These fragmented interventions seem to focus more on damage control than prevention, effectively shifting responsibility onto users and regulators while leaving the underlying risks largely unresolved.

The exploitation of AI to generate and circulate sexualised imagery on social media is not a new crisis, but one that evolves over time. As early as 2020, investigations revealed that Telegram, an encrypted messaging platform, hosted AI-powered chatbots capable of "nudifying" photos of women submitted anonymously. Most recently, at least 150 Telegram channels were identified operating internationally to facilitate the creation and sale of deepfake sexual content.

These channels also double as information-sharing sources, where users exchange technical tips to bypass existing safeguards. This pattern suggests the emergence of a coordinated, cross-platform ecosystem that facilitates and normalises non-consensual exploitation. Telegram has since stated that it employs proactive monitoring and customised AI tools to enforce its policies, claiming to have removed more than 952,000 pieces of offending material last year.

Yet new channels routinely emerge after takedowns, exposing persistent enforcement gaps and underscoring the limitations of reactive moderation in addressing deeply embedded, technologically enabled harms.

The Root of the Problem

These recent events illustrate the critical shortcomings of existing technical safeguards in preventing the creation and dissemination of non-consensual sexually explicit material (NCSEM).

Fundamentally, the persistence of NCSEM stems from the design of most major social media platforms. Their recommendation algorithms are engineered to prioritise engagement and virality, often amplifying sensational, emotionally charged, or polarising content to capture user attention and maximise advertising revenue. This process usually occurs well before moderation systems can intervene, leaving little opportunity to prevent harm. These dynamics place clear responsibility on the platforms themselves and highlight the need for governance that tackles the systemic failure of content moderation.

While reactive takedowns are necessary, they ultimately serve as a temporary solution for a structural problem. Addressing it effectively requires a shift in digital governance, moving beyond voluntary platform guidelines to systemic frameworks that tackle algorithmic amplification, product deployment, and accountability. Without confronting these underlying mechanics, efforts to curb large-scale sexual exploitation and deepfake distribution will remain fundamentally inadequate.

A Heightened Issue for the Vulnerable

The proliferation of NCSEM on social media not only circulates harmful content, it also normalises digital harassment and institutionalises image-based abuse. The misuse of deepfake technology, in particular, inflicts profound psychological and social trauma to victims.

The highly realistic nature of AI-generated sexual content violates a victim's sense of self, often resulting in a perceived loss of body autonomy and leading to distress,

anxiety, shame, and lasting emotional harm. Research also indicates that the trauma from the non-consensual distribution of such images is comparable to that experienced by survivors of physical sexual violence, further highlighting how technology has evolved to inflict very real, tangible harm.

These harms, however, are disproportionately concentrated on women and children, who are often the primary targets. The victims face not only personal distress but also considerable social stigma, damage to their reputation, and reduced professional opportunities.

This crisis is especially problematic in Southeast Asia, where significant legal gaps intersect with cultural stigma and limited access to mental health resources. Women and children, in particular, are vulnerable to online trafficking, exploitation, and sexual abuse in the region. With limited legal protections, rights, and support systems, these groups remain exposed to abuse and grooming.

The unchecked proliferation of AI-generated sexual material exacerbates these vulnerabilities, producing tangible harms for victims while also amplifying broader geopolitical and reputational risks for the region. However, existing legal frameworks and platform governance mechanisms have so far proven inadequate to deter perpetrators or offer meaningful redress to victims, emphasising the need for proactive, preventive regulatory approaches rather than reactive enforcement.

What Now For ASEAN?

Amid recent events concerning platform regulation, ASEAN states have a timely opportunity to strengthen regional digital governance. The temporary bans by Malaysia and the Philippines on Grok, followed by X's implementation of additional safety measures, show that member states can influence platforms, uphold local norms, and protect users from harmful content.

The widespread proliferation of harmful content has also put platforms under intense scrutiny, amplifying legal and reputational risks and pressuring them to adopt robust, multi-layered safeguards. Building on this momentum, ASEAN can advance coordinated regional guidelines grounded in local values and reflecting international best practices.

Drawing on frameworks such as the EU's Digital Services Act and AI Act, ASEAN states can, provided they achieve meaningful regional coordination and regulatory alignment, require platforms to carry out comprehensive risk assessments, implement safety-by-design measures, and maintain ongoing human oversight alongside automated moderation.

Moreover, effective regional governance requires national regulatory frameworks developed in collaboration with major platforms. Shared definitions of prohibited content, interoperable enforcement mechanisms, and common transparency requirements on moderation can provide a foundation for coordinated action.

Multilateral and multistakeholder forums can also establish shared norms, while app stores and infrastructure providers should be included in accountability frameworks as gatekeepers with enforceable duties of care.

While broader digital governance may face political and regulatory differences across the region, member states can find common ground in tackling NCSEM on social media. In this clearly harmful content, local values and international norms largely align. Embedding protections for women and children within these shared standards can help ensure that regional cooperation results in meaningful safeguards for the most vulnerable users. Focusing on this shared priority could also serve as a starting point for building trust and testing cooperative mechanisms that can later extend to other digital risks.

While these proposed measures point to promising directions, significant challenges persist: regulatory capacity varies widely across the region, enforcement is often inconsistent, and harmonised frameworks with proactive measures remain lacking.

To address these gaps, it is necessary to establish regional centres of expertise, shared audit mechanisms, and capacity-building initiatives for regulators to ensure that algorithmic amplification does not systematically prioritise harmful or exploitative content. Preventive systems should also be mandated to hold platforms accountable. Likewise, safeguards must be implemented to protect regulatory sovereignty from geopolitical pressures, preventing diplomatic or commercial interests from undermining public-interest protections.

Lastly, policy responses should go beyond content takedowns to invest in digital and AI literacy, equipping individuals – especially young people – with the skills to recognise, report, and resist online sexual harms. Governments and platforms should also provide accessible resources for parents and caregivers, including education toolkits, reporting pathways, and trauma-informed support services.

Strengthening these preventive measures is vital for reducing vulnerability and building community-level resilience against image-based abuse and AI-enabled exploitation. By fostering a coordinated ASEAN-wide strategy, member states can enhance user protection, strengthen digital sovereignty, and promote the responsible and ethical deployment of AI across the region.

Ysa Marie Cayabyab is an Associate Research Fellow with the Future Issues and Technology (FIT) research cluster at the S. Rajaratnam School of International Studies (RSIS), Nanyang Technological University (NTU), Singapore.

Please share this publication with your friends. They can subscribe to RSIS publications by scanning the QR Code below.