

*SG60 Policy Simulation*

Vision White Paper

Team Code: 19

## I. Preamble

The rapid advancement of AI-generated media, particularly deepfake technology, is increasing the risk of malicious content targeting youths. While Singapore has strong online safety initiatives, there is a lack of recovery-focused measures for victims, especially for youths whose reputations, mental health, and future opportunities may be severely harmed.

Recent local incidents especially highlight the urgency of this threat. For instance, the police investigated a case involving deepfake nude images of Singapore Sports School Students created and shared by their peers, with both female students and teachers targeted [1]. The school lodged a police report, provided counselling to victims, and implemented disciplinary action. However, parents expressed concern over the long-term emotional and reputational harm, as well as whether perpetrators should remain in the same school environment.

Beyond targeting sexual imagery, the detection of AI-generated content amongst Singaporeans in itself poses a challenge. A 2024 Cyber Security Agency (CSA) survey [2] found that only one in four Singaporeans can accurately identify deepfake videos, despite 78% feeling confident that they could do so. This gap between confidence and ability raises pressing concerns about the misinformation, fraud, and harassment risks in the coming decade.

Singapore's online safety frameworks have evolved from cyber wellness education in the 2000s, to the Protection from Harassment Act in 2014 [3], and to the Online Safety Bill in 2022 [4]. These were largely reactive measures to cyberbullying and explicit content circulation. However, deepfake capabilities have outpaced current safeguards, becoming increasingly convincing and accessible, especially with cases already appearing in schools globally.

While prevention and detection are prioritised, post-incident recovery (such as rapid takedowns, digital footprint repair, legal recourse for victims, etc) remains underdeveloped. This ends up creating long-term harm that can manifest in several ways amongst the affected youth, such as but not limited to a lack of engagement in both personal and academic pursuits, the emergence of depressive symptoms such

as suicidal ideation, and increased social isolation, leading to a detrimental cycle that exacerbates their challenges [5].

While cyber safety education and reactive legal tools were effective against earlier threats like phishing and harassment, they do not anticipate synthetic media at scale. Consequently, assistance for these attacks remains lacking, leaving victims of these cyberattacks to endure the harsh repercussions of the perpetrators' actions [6]. By the years 2026 to 2035, advancements in AI deepfake technology are expected to lead to even more realistic and easily accessible tools [7]. This increased accessibility may result in a significant rise in malicious uses, highlighting the urgent need for awareness and protective measures against potential abuses. Thus, early intervention is crucial to building resilient legal, social, and technical systems for both prevention and recovery.

The growing concerns about the dangers of AI deepfakes are crucial when safeguarding Singapore's goals of security, fairness, and recovery. This is through ensuring the security and safety of people when facing online threats, making sure that all youths, regardless of their background, fairly receive the aid required for their recovery, and building systems to help victims of these cyberattacks integrate back into society. Through these actions, we can continue to show our commitment to supporting Singapore's goal of fostering a caring society that safeguards its citizens and an inclusive society where everyone can confidently navigate online content with ease, security, and satisfaction.

## **II. Actions by the Government**

In response to the growing threat of AI-generated malicious content such as deepfakes, the Singaporean government has taken a multi-faceted initiative involving legislation, regulation, education, and community engagement to promote digital security, especially among vulnerable groups such as youths [8]. These efforts are based on a strong legal framework, with statutes such as the Protection from Harassment Act (POHA) criminalising the circulation of intimate images, including deepfake pornography, without consent [8], and the Online Criminal Harms Act (OCHA), responding to online offenses such as scams and harmful AI-generated

content [9]. The Online Safety Commission, which will begin operations in 2026, will provide victims with access to civil redress while holding perpetrators accountable [10]. National efforts like Digital for Life and the Digital Defence pillar of Total Defence also encourage responsible use of tech, public awareness of online harm, and community engagement in fostering a secure digital space[11].

These legislations are supplemented by coordinated actions from agencies like the Cyber Security Agency of Singapore (CSA) and the Infocomm Media Development Authority (IMDA)[12]. The Ministry of Education (MOE) and Ministry of Home Affairs (MHA) also supplement these legislations by playing essential roles in education within schools, counselling, and enforcement [13].

Singapore's strength lies in its swift, top-down responses, clear legislation, and effective inter-agency coordination, with it being especially highlighted during the Covid-19 pandemic [14]. However, current measures focus more on immediate containment rather than long-term recovery, leaving victims, particularly youths, without comprehensive frameworks to rebuild trust, confidence, and digital reputations. Although some civil and common law remedies [15] exist, they remain limited, and gaps persist in detection and public literacy, with CSA's 2024 findings [2] highlighting a mismatch between policy ambition and societal resilience. Ultimately, legal tools and rapid responses are not substitutes for the slower but essential cultural and educational shifts needed to ensure lasting safety in an AI-saturated media landscape.

To prevent more failures, Singapore would need to go beyond swift enforcement and focus on building long-term resilience. This would mean developing comprehensive victim recovery frameworks that offer sustained psychological support, reputation restoration, and guidance for re-entering safe online spaces, especially for youths. Detection systems must be strengthened through more advanced monitoring tools and closer collaboration with tech platforms to identify and remove harmful content before it spreads. Public digital literacy focusing on artificial intelligence needs a deeper push, with it being not just one-off campaigns, but sustained education from early schooling onwards that teaches critical thinking, consent, and the realities of AI manipulation. Policies should also adapt quickly to new technological threats, with continuous reviews to close loopholes before they can be exploited.

### **III. Board/Agency Decisions**

The Online Criminal Harms Act (OCHA) empowers the Singapore Government to direct online platforms to prevent the spread of potential scam-related content, including deepfakes, to local users [9]. Complementing this, initiatives such as the National Library Board's S.U.R.E. programme aim to equip the public with skills to critically evaluate online information [17]. These programmes address areas such as digital media literacy, cybersecurity, and scam awareness, ultimately building a safer digital society. In responding to rising deepfake AI cases, statutory boards such as IMDA [12] could adopt operational frameworks focused on victim protection, rapid content removal, and platform accountability. These might involve close coordination with social media companies, the Singapore Police Force, and victim support services. Clear protocols for evidence preservation, investigation, and follow-up would also be vital. Targeted outreach programmes geared towards youth could also be trialled to test the effectiveness of such measures before they are scaled up.

Managing trade-offs, such as balancing swift takedowns with thorough evidence gathering, or protecting victim privacy while enabling legal action, would require well-defined operational guidelines. Board-level planning should also consider downstream impacts, including the psychological harm to victims, the strain on investigative resources, and the potential erosion of public trust in digital spaces. Mitigation strategies could involve establishing a centralised reporting hub, expanding public awareness campaigns about the legal consequences of creating or sharing non-consensual deepfake content, and building cross-agency detection and response teams. By embedding these safeguards, the response to the misuse of AI is sure to be both swift and victim-centred.

### **IV. Financial Spending**

Our policy is to extend the operation of the Centre for Advanced Technologies in Online Safety (CATOS) until 2036. Currently, CATOS has been established and in operation since 2023 with funds from the Online Trust and Safety (OTS) Research Programme under the Smart Nation and Digital Economy domain of Singapore's

Research, Innovation and Enterprise (RIE) 2025 plan. CATOS is focused on developing advanced detection systems for harmful digital content, with its funding ending in 2026. As of now, CATOS has not prioritised research into deepfake AI technology, with little to no ground broken on its prevention, early detection, and mitigation [16]. Since AI deepfakes are a rising threat, we believe that with a projected further funding of \$20 million from the RIE from 2026 to 2036 and a new focus on AI, CATOS will allow Singaporeans better protection from AI threats. This will be because policies for protection against AI threats will be based on thorough research that is applicable to our society and its cultural context.

The projected budget will cover the infrastructure and technology maintenance for the centre's laboratories and research tools, its research and development efforts, including hiring and training experts in AI safety and cybersecurity, and necessary collaboration and outreach initiatives to share findings with industry partners and the public. The primary cost drivers are expected to be the hiring of specialised researchers, which can be costly, especially in a competitive global market. Hence, a higher percentage of the available budget will be allocated to it in order to ensure CATOS is run with maximised efficiency.

An emergency fund of \$5 million will be earmarked within the initiative to address unforeseen threats, rapid technology shifts, or urgent operational needs. This reserve ensures flexibility in responding to sudden spikes in malicious online activity or technological vulnerabilities without disrupting planned timelines.

Barring major economic or geopolitical crises, we believe that CATOS will be able to effectively and sustainably utilise the funds allocated to it over its extended five-year timeline, given its structured expenditure plan, alignment with national strategy, and oversight under government accountability mechanisms.

High maintenance costs, talent retention challenges, and the need to demonstrate measurable public impact in a rapidly evolving digital threat environment pose a financial risk to the project because they could lead to overspending, mass talent attrition, and discontinued funding due to a lack of measurable expected outcomes. Since these factors are interdependent, their financial risk is exponentially greater as well.

Due to high implementation urgency, we have chosen to deploy funds in a single concentrated injection rather than staggering them on a needs-based approach. This approach would ensure immediate research momentum in the face of fast-emerging AI threats. However, it also front-loads costs and assumes that early breakthroughs will sustain impact beyond the initial funding period.

## **V. Evolving Key Issues**

The core issue remains that deepfake technology advances faster than detection systems. Even with CATOS receiving extended funding until 2036, creators of malicious content consistently stay ahead of protective measures. Additionally, while Singapore can quickly remove harmful content domestically, it cannot prevent the same material from spreading on international platforms beyond local regulatory control.

Current systems also excel at immediate response but fail at long-term victim recovery. As seen in the Sports School incident [1], while schools can provide immediate counselling and discipline, there are no systematic programs to help victims rebuild their digital reputation or safely return to online spaces. The government's approach remains reactive - responding to incidents rather than preventing new types of threats before they emerge.

Youth from lower-income families struggle to access specialised legal help and counselling services needed for recovery. International students face additional barriers due to unfamiliarity with Singapore's support systems and potential concerns about engaging with authorities.

The confidence gap identified in CSA's survey [2] - where most Singaporeans think they can spot deepfakes but cannot - creates false security that increases vulnerability. Coordination between multiple agencies (MOE, MHA, CSA, IMDA) [12][13], while generally effective, can create delays when incidents require both immediate action and long-term support.

Cultural stigma around being a victim of digital harassment discourages reporting and help-seeking, while AI literacy education struggles to keep pace with rapidly

changing technology. Policy review cycles are too slow for the exponential advancement of AI capabilities expected through 2035 [7].

To address these evolving challenges, Singapore must shift from reactive enforcement to proactive prevention, establish comprehensive victim recovery programs, and create faster policy adaptation mechanisms that can keep pace with rapidly advancing AI threats.

## VI. List of References

[1] G. Tan, "Police investigating deepfake nude photos of Singapore Sports School students," The Straits Times, p. 1, Nov. 12, 2024 [Online].

<https://www.straitstimes.com/singapore/police-investigating-deepfake-nude-photos-of-singapore-sports-school-students>

[2] Channel News Asia, "Only one in four Singaporeans can accurately identify deepfake videos: CSA Survey" Channel News Asia, p. 1, Jul. 02, 2025 [Online].

<https://www.channelnewsasia.com/watch/only-one-in-four-singaporeans-can-accurately-identify-deepfake-videos-csa-survey-5216656>

[3] Singapore Statutes Online, "Protection from Harassment Act 2014",

<https://sso.agc.gov.sg/>, Aug. 12, 2025 [Online]. Available:

<https://sso.agc.gov.sg/Act/PHA2014?ProvIds=pr3-> [Accessed: Aug. 12, 2025].

[4] "First Reading of Online Safety (Miscellaneous Amendments) Bill", Oct. 3, 2022

[Online]. [https://isomer-user-content.by.gov.sg/38/2d30dd91-977f-43ea-997d-280a0d534614/media%20factsheet%20-%20online%20safety%20\(miscellaneous%20amendments\)%20bill%20\(3%20oct\\_final\).pdf](https://isomer-user-content.by.gov.sg/38/2d30dd91-977f-43ea-997d-280a0d534614/media%20factsheet%20-%20online%20safety%20(miscellaneous%20amendments)%20bill%20(3%20oct_final).pdf)

[5] S. Bansal et al, "Cyberbullying and mental health: past, present and future",

Jan. 15, 2024 [Online]. <https://doi.org/10.3389/fpsyg.2023.1279234>

[6] Rice Media, “A Cyber Wellness Expert's Thoughts on Online Bullying | Shem”, Oct. 28, 2024 [Online]. <https://www.youtube.com/watch?v=7yeNwiegAJU&t=3s>

[7] D. Klepper, “Creating realistic deepfakes is getting easier than ever. Fighting back may take even more AI”, Jul. 28, 2025 [Online].  
<https://www.ap.org/news-highlights/spotlights/2025/creating-realistic-deepfakes-is-getting-easier-than-ever-fighting-back-may-take-even-more-ai/>

[8] K. W. Thin, “CNA Explains: What are the legal options for victims of deepfake porn?”, Nov. 18, 2024 [Online].  
<https://www.channelnewsasia.com/singapore/deepfake-porn-legal-action-cna-explains-4749306>

[9] “Online Criminal Harms Act 2023”, Aug. 11, 2023 [Online]. Available:  
<https://sso.agc.gov.sg/Act/OCHA2023> (accessed Aug. 13, 2025)

[10] A. Leck *et al*, “Singapore: Update on government agency intended to tackle online harms”, May. 9, 2025 [Online].  
<https://connectontech.bakermckenzie.com/singapore-update-on-government-agency-intended-to-tackle-online-harms/>

[11] “Digital Defence: Being secure, alert, and responsible online”, Jul. 14, 2025 [Online]. Available:  
<https://www.totaldefence.gov.sg/about/the-six-pillars-of-total-defence/digital-defence>

[12] “Summary of Initiatives to Help Organisations Enhance Cybersecurity”, Oct. 17, 2024 [Online].  
<https://www.csa.gov.sg/news-events/press-releases/summary-of-initiatives-to-help-organisations-enhance-cybersecurity>

[13] “Training of educators/counsellors as well as concerns about teachers’ mental health”, Sep. 14, 2021 [Online]. Available:  
<https://www.moe.gov.sg/news/parliamentary-replies/20210914-training-of-educators-counsellors-as-well-as-concerns-about-teachers-mental-health>

[14] S. Asthana *et al*, “Singapore's COVID-19 crisis decision-making through centralization, legitimacy, and agility: an empirical analysis”, Jul. 8, 2024 [Online].  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC11279669/>

[15] “LEGAL REMEDIES FOR VICTIMS OF HARASSMENT”, Feb, 2021 [Online].  
<https://www.mlaw.gov.sg/files/news/press-releases/2021/02/POHA-AnnexA.pdf>

[16] “Online Safety is a Foundation to a Thriving Digital Society”, 2015.  
<https://www.catos.sg/>

[17] National Library Board, “Be S.U.R.E. Before You Share,” Sept. 15, 2023.  
<https://www.youtube.com/watch?v=Z2cPIIVOD8w>