AI Chatbots, Youth Mental Health, and the Case for Regulation

Introduction

On a quiet early morning on 11 April 2025, Adam Raine's parents found themselves confronting a shocking and incomprehensible reality: their 16-year-old son had committed suicide by hanging.[1] According to their testimony to the US Senate, Adam had never been diagnosed with a mental health condition, and he was an outwardly fun-loving and optimistic teenager who wanted a career in medicine.[2] They had no idea he was suicidal or had been struggling with crippling anxiety.

Adam's parents began to piece things together only after examining his phone and discovering extensive conversations with an AI chatbot (*ChatGPT*). What began as a homework helper gradually became a confidant and emotional guide. Buried in the messages were signs of escalating distress, unrecognised and unheard by anybody. Adam had confided in the chatbot, discussing his deepest emotional struggles and growing feelings of despair and isolation.

In their testimony, Adam's parents alleged that the AI chatbot responded in ways that discouraged Adam from seeking family support, and that it appeared to validate his darkest thoughts instead of directing him toward professional help. For instance, when Adam described emotional numbness, loneliness and anxiety, the chatbot reportedly continued the exchange rather than clearly directing him to professional support.[3]

Court filings and testimony alleged that the final exchanges contained unsafe responses in the context of acute distress.[4] Adam's parents' message to lawmakers is unequivocal: Adam's death was avoidable, and existing (technological, institutional, and regulatory) systems had failed to protect him from harm. The Raines are now pursuing legal action (Raine v. OpenAI), claiming the chatbot's responses worsened their son's mental state and validated his "most harmful and self-destructive thoughts".[5]

Although Singapore has not experienced a similar tragedy, Adam Raine's case has highlighted a set of highly salient questions in our local context. Singapore is a digitally connected society with high adolescent technology use,[6] strong trust in institutions, and an active national focus on youth mental well-being. As generative AI (GenAI) tools, including chatbots, become increasingly embedded in

[1] Nadine Yousif, "Parents of teenager who took his own life sue OpenAI", *BBC News.* August 27, 2025, https://www.bbc.com/news/articles/cgerwp7rdlvo.
[2] Raine M., Written testimony before the United States Senate Judiciary Subcommittee on Crime and Counterterrorism, September 2025. United States Senate Committee on the Judiciary. September 16, 2025, accessed December 22, 2025. https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Raine.pdf.
[3] Bhuiyan J., "ChatGPT encouraged Adam Raine's suicidal thoughts. His family's lawyer says OpenAI knew it was broken", *The Guardian.* August 22, 2025. Accessed December 22, 2025, https://www.theguardian.com/us-news/2025/aug/29/chatgpt-suicide-openai-sam-altman-adam-raine.
[4] Raine, written testimony.
[5] Yousif, "Parents of Teenager Who Took His Own Life Sue OpenAI."
[6] Kelly XY Foo. Balancing connectivity and youth mental well-being: challenges of social media. *Tech for Good Institute*, accessed December 23, 2025. https://techforgoodinstitute.org/research/research-commentary/balancing-connectivity-and-youth-mental-well-being-challenges-of-social-media/

This case has been written by Ng Qin Xiang under the guidance of A/Prof Jason Yap Chin Huat and has been funded by the Lee Kuan Yew School of Public Policy. The case does not reflect the views of the sponsoring organisation nor is it intended to suggest correct or incorrect handling of the situation depicted. This case is based largely on actual events, but some characters and events have been altered for the purposes of this case study. The case is not intended to serve as a primary source of data and is meant solely for class discussion.

everyday life, including with young people navigating emotional stress and uncertainty, the absence of clear guardrails raises important questions of public safety.

The central policy issue is not whether AI chatbots should be permitted to exist, but whether current governance frameworks are adequate when such tools begin to provide counselling, emotional support, and mental health advice. How can Singapore to tap on its existing strengths in digital governance and policy to ensure that such innovations do not outpace protection, particularly for vulnerable youths at the margins of visibility and care. How should governments respond to the growing role of AI chatbots in the context of the growing youth mental health crisis? And what would be the necessary safeguards and policies as these tools are developed?

Background

Youth Mental Health as a Public Health Problem

The Coronavirus Disease 2019 (COVID-19) pandemic exacerbated the burden of mental health conditions. Mental illness was already one of the leading contributors to disease burden, with adolescence and early adulthood recognised as critical periods of vulnerability.[7] Epidemiological evidence consistently showed that the majority of mental disorders emerged before the age of 25. This life stage was characterised by rapid biological development, identity formation, academic and social pressures, and increasing independence, all of which could heighten one's susceptibility to feelings of anxiety, depression and distress.[8]

In Singapore, this trend was clearly reflected in national data. According to findings from the Singapore Youth Epidemiology and Resilience (YEAR) study[9] and National Youth Mental Health Study,[10] one in seven young people aged 13 to 25 years reported internalising symptoms (e.g., sadness, anxiety and loneliness); and severe and extremely severe symptoms of depression and anxiety were reported by 14.9% and 27.0% of surveyed youths, respectively. These figures pointed to an urgent, unmet need for mental health services for young people in today.

Use of AI Chatbots by Adolescents and Youths

Against this backdrop, free-to-use GenAI platforms, especially ChatGPT by OpenAI, rapidly embedded themselves into adolescent digital life. While initially marketed as a productivity and information tool, evidence suggested that adolescents and young adults increasingly used AI chatbots for emotional expression, relationship advice and mental health-related queries. Given the

[7] Zhuo Liu and Mengni Kuai, "The Global Burden of Depression in Adolescents and Young Adults, 1990–2021: Systematic Analysis of the Global Burden of Disease Study," *BMC Psychiatry* (2025): 767, https://doi.org/10.1186/s12888-025-07201-8.
[8] Rebecca Potterton et al., "Identity Development and Social-Emotional Disorders During Adolescence and Emerging Adulthood: A Systematic Review and Meta-Analysis," *Journal of Youth and Adolescence* (2022): 16-29, https://doi.org/10.1007/s10964-021-01536-7.
[9] John CM Wong et al., "Youth Epidemiology and Resilience (YEAR) in a Student Population: Prevalence and Predictors of Mental Health Symptoms," *Frontiers in Psychiatry* (2024): 1454484, https://doi.org/10.3389/fpsyt.2024.1454484.
[10] Mythily Subramaniam, et al. "Examining Psychological Distress among Youth in Singapore: Insights from the National Youth Mental Health Study." *Asian Journal of Psychiatry* (2025): 104405, https://doi.org/10.1016/j.ajp.2025.104405.

shortage of mental health services, young users described AI chatbots as "non-judgmental" and "always available," perceiving them to be accessible companions when human support felt distant.[11]

While exact, real-time figures are difficult to ascertain, a survey of 2,125 young people in the United States found that around one in eight (or 13.1% of) adolescents and young adults use AI chatbots for mental health advice, with this behaviour most common among those aged 18 to 21 years.[12] At the platform level, data from OpenAI, the company that developed *ChatGPT*, reported that around 1.2 million users 'talked' to the chatbot about suicidal ideation each week.[13]

Such data strongly suggested that AI chatbots were no longer peripheral tools. For a subset of young people, they were informal emotional companions, occupying a space traditionally held by peers, family members or mental health professionals.

Evidence and Limitations of AI Chatbots in Mental Health

Despite broad public acceptance and adoption, many available AI chatbots were not designed or clinically validated as therapeutic instruments, and most did not undergo regulatory assessment equivalent to telehealth services or digital therapeutics.[14] Even where the evidence was cautiously optimistic, it is important to distinguish between chatbot-delivered interventions that were purpose-built, theory-grounded, and evaluated in interventional studies; and publicly available, general-purpose or "companion" AI chatbots that were free, ubiquitous without independent clinical validation or governance.[15]

The underlying large language models (LLMs) often operated autonomously, generating personalised emotional content without real-time human oversight. This expanded reach and usability, but also increased exposure to unmoderated advice at scale. Adolescents, who were arguably more vulnerable to identity stressors, isolation and impulsivity, thus faced an more risk when engaging with GenAI technology designed to mimic empathic communication, and which operate without a trained healthcare professional. Adolescents, who were arguably more vulnerable to identity stressors, isolation and impulsivity, thus faced greater risk when engaging with GenAI technology designed to mimic empathic communication and operated without a trained healthcare professional. Many existing platforms lacked evidence-based therapeutic frameworks, structured crisis-support pathways or mandatory referral mechanisms. At the same time, safeguards built to prevent inappropriate or harmful output could be bypassed through role play, alter egos or engineered prompts, or did not reliably activate.[16] Harmful outputs could emerge after multi-turn conversations when the model's tone became empathetic, validating and procedurally instructional.

---

[11] Ding, Xiaoxu, and Skye Barbic. "Perception of AI Use in Youth Mental Health Services: Qualitative Study." *Journal of Participatory Medicine* (2025): e69449. https://doi.org/10.2196/69449.

[12] Kathryn McBain et al., "Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults." *JAMA Network Open* (2025): e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281.

[13] *Australian Psychological Society.* ChatGPT sees one million users in mental distress each week, APS in The Daily Aus. November 4, 2025. Accessed December 22, 2025. https://psychology.org.au/insights/chatgpt-sees-one-million-users-in-mental-distress.

[14] Babu, Anithamol, and Akhil P. Joseph. "Digital Wellness or Digital Dependency? A Critical Examination of Mental Health Apps and Their Implications." *Frontiers in Psychiatry* (2025): 1581779. https://doi.org/10.3389/fpsyt.2025.1581779.

[15] Jiaying Li et al., "Chatbot-Delivered Interventions for Improving Mental Health Among Young People: A Systematic Review and Meta-Analysis," *Worldviews on Evidence-Based Nursing* (2025): e70059, https://doi.org/10.1111/wvn.70059.

[16] Annika M. Schoene and Cansu Canca, "'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts," *arXiv preprint* (July 8, 2025), https://arxiv.org/abs/2507.02990.

The regulatory environment further complicated matters, as many AI chatbots were categorised as general consumer tools rather than health-adjacent services requiring accreditation or governance.[17]

Problem

The Adam Raine case thrusted into public view a central policy challenge that confronted governments worldwide: what obligations did AI developers, platform operators, and regulators have when publicly available chatbots were used by adolescents to navigate mental health distress, and what safeguards would be necessary to prevent foreseeable harm?

At the time of writing, most jurisdictions treated general-purpose AI chatbots as limited-risk consumer systems,[18] subjected primarily to transparency requirements. These included informing users that they were not interacting with an actual human. Under this regulatory framing, conversational AI platforms were not classified as medical devices, digital therapeutics or formal mental health services. As a result, they were largely exempt from clinical accreditation standards, evidence thresholds and governance mechanisms that applied to telehealth or digital mental health (DMH) interventions.[19]

However, advocates and some experts argued that this classification no longer reflected how these tools were used in practice. They contended that AI chatbots should be subjected to stronger oversight when they, like to other DMH tools, responded to personal disclosures of psychological concern; provided advice or guidance with emotional influence; and operate in contexts where risk was reasonably foreseeable.[20] From this perspective, the functional role of the technology, rather than the developer's stated intent, should determine the level of regulatory scrutiny.

Technology companies, including OpenAI, pushed back against this framing. They emphasised the published terms of use for their products and noted that these platforms were not intended as medical devices.[21] For instance, OpenAI's terms of use explicitly prohibited individuals under 18 years old from using *ChatGPT* without consent from a parent or guardian, and users were also told not to use the bot for "suicide" or "self-harm". Industry actors also warned that classifying conversational AI as regulated health tools could stifle innovation, create excessive compliance burdens, and limit beneficial applications given the rapid pace of technological development.[22] In response to public concern, companies such as OpenAI developed voluntary safety features and add-ons, such as content moderation, crisis-resource referrals, parental controls and age-prediction systems.[23]

Nonetheless, the fact remained that AI chatbots are producing emotionally persuasive content and adolescents were using them as mental health companions, not solely as general purpose or productivity tools. While chatbots did not claim clinical authority, their conversational design

[17] Shumate, J. Nicholas, et al. "Governing AI in Mental Health: 50-State Legislative Review." *JMIR Mental Health* (2025): e80739. https://doi.org/10.2196/80739.
[18] EU Artificial Intelligence Act. High-level summary of the act. February 27, 2024. Accessed December 24, 2025. https://artificialintelligenceact.eu/high-level-summary.
[19] Ibid.
[20] Babu, Anithamol, and Akhil P. Joseph. "Digital Wellness or Digital Dependency? A Critical Examination of Mental Health Apps and Their Implications." *Frontiers in Psychiatry* (2025): 1581779. https://doi.org/10.3389/fpsyt.2025.1581779.
[21] *OpenAI.* Our approach to mental health-related litigation. November 25, 2025. Accessed December 24, 2025. https://openai.com/index/mental-health-litigation-approach.
[22] Panagoulias, D. P., G. A. Tsihrintzis, and M. Virvou. "Challenges in Regulating and Validating AI-Driven Healthcare." In Artificial Intelligence–Empowered Bio-medical Applications, Learning and Analytics in Intelligent Systems, vol. 49. *Cham: Springer*, 2025. https://doi.org/10.1007/978-3-031-90174-4_6.
[23] *OpenAI,* "Our Approach to Mental Health-Related Litigation."

intentionally mimicked empathy, understanding and relational engagement. This human-like interaction could encourage anthropomorphism and foster a sense of trust or emotional reliance even when users were explicitly aware that the chatbot was an artificial entity.[24] From a public safety and ethics of care perspective, the consequence was a regulatory grey zone and safety vacuum, in which vulnerable users could receive inaccurate or unsuitable advice that emotionally reinforced feedback loops, as well as inadequate crisis redirection with no clinical guardrails.

As AI platforms introduced more health-specific modes, interfaces and data integrations, functional use, rather than stated intent, becomes an even more salient basis for regulation. For Singapore, the emergence of products such as *ChatGPT* Health strengthened the case for risk-tiered governance and hybrid licensing models, as it demonstrated both the feasibility of stronger safeguards and the risks of leaving emotionally and clinically influential systems outside formal oversight simply because they were embedded within broader consumer platforms.

In January 2026, OpenAI announced *ChatGPT Health*, a dedicated health and wellness environment within *ChatGPT* that allowed users to connect personal health records, wearable data and wellness apps to help people "take a more active role in understanding and managing their health and wellness".[25] This development highlighted a critical policy point: the boundary between "general-purpose" AI and health-adjacent tools was no longer static. As AI platforms increasingly introduced health-specific modes, interfaces and the ability to integrate personal health data, functional use rather than stated intent became an even more salient basis for regulation. While legal questions surrounding liability, consent, foreseeability of harm and duty of care remained unresolved and will be tested in court in the case of Raine v OpenAI, the broader policy implications were already evident. Existing regulatory frameworks were not designed for GenAI technologies that occupied the psychological and relational spaces once mediated primarily by families, schools and clinicians.

The policy challenge, therefore, was not whether AI chatbots should exist or be accessible to young people, but how governance structures could evolve to reflect their real-world use. Any regulatory response would have to balance competing priorities: protecting vulnerable users without unduly constraining innovation; preserving user autonomy while recognising asymmetries in power and knowledge; and fostering public trust in digital ecosystems without over-regulating general-purpose technologies.

Policy Options

Across healthcare systems, governments were already grappling with similar challenges in regulating DMH tools, including the expansion of mental health apps, teletherapy platforms and AI-enabled digital therapeutics in the digital marketplace since the COVID-19 pandemic.[26] These experiences offered useful lessons for governing conversational AI systems that increasingly performed mental health-adjacent functions.

International DMH accreditation frameworks shared several core principles (summarised in the Appendix, Table S1). They emphasised clinical effectiveness, requiring tools, particularly those deployed in high-risk contexts, to demonstrate measurable benefit.[27] They required clinical safety,

[24] Sedlakova, J. "Conversational AI for Psychotherapy and Its Role in the Space of Reason." *COSMOS Taxis* (2024): 80–87. https://cosmosandtaxis.org/wp-content/uploads/2024/06/ct_vol12_iss_5_6_epub_r2.pdf.
[25] *OpenAI.* Introducing ChatGPT Health. January 7, 2026. Accessed January 7, 2026. https://openai.com/index/introducing-chatgpt-health.
[26] Bloomfield, Peter, and Carlos R. de Villa. 2021. Care Tech Landscape Review: Mental Health. Accessed December 28, 2025. https://futurecarecapital.org.uk/research/residential-care-tech-landscape-review.
[27] Mäder, Matthias, Philip Timpel, Tino Schönfelder, Claudia Militzer-Horstmann, Stephan Scheibe, Ralf Heinrich, and Dominik Häckl. 2023. "Evidence Requirements of Permanently Listed Digital Health Applications

including mechanisms to prevent harm and clear protocols for crisis escalation. They mandated robust privacy and data security, given the sensitivity of mental health information. They also assessed usability and engagement, recognising that poorly designed interfaces could themselves pose risks. Crucially, these frameworks adopt risk tiering (Table 1), whereby tools with greater therapeutic influence or higher foreseeable risk were subject to stronger evidence requirements, governance standards, and post-market surveillance.[28]

Table 1: Risk-tier classification commonly adopted for evaluating DMH tools

Tier	**Category**	**Examples**	**Required Evidence Level**
Tier 1 (Low)	Information / Wellness	Psychoeducation sites, mindfulness trackers	Expert validation + usability testing
Tier 2 (Medium)	Guided self-help / Monitoring	Cognitive behavioural therapy modules, journaling apps, mood trackers	Observational / pilot data + safety review
Tier 3 (High)	Therapeutic / Clinical decision support	DTx for depression, AI risk screeners	RCT / comparative trial + regulatory clearance
Tier 4 (Medical device or AI SaMD)	Clinician-facing diagnostic or prescriptive apps	Predictive AI decision tools	FDA / CE certification + post-market surveillance

Abbreviations: DTx, Digital Therapeutics; AI, Artificial Intelligence; SaMD, Software as a Medical Device; HSA, Health Sciences Authority; FDA, Food and Drug Administration; CE, Conformité Européenne; RCT, Randomized Controlled Trial

For Singapore, these principles are particularly salient. The country already operates within a strong tradition of risk-based regulation, public trust in institutions, and proportionate governance across healthcare, data protection and digital services.[29] Examples of regulatory controls relevant to digital health and DMH products in Singapore are summarised in Table 2. Rather than constructing an entirely new regulatory regime for AI chatbots, policymakers could draw selectively on these established digital health and DMH logics and adapt them to the unique features of conversational AI chatbots.

Table 2: Current regulatory controls relevant to DMH products in Singapore

Regulatory control	**Relevance**
Personal Data Protection Act (PDPA)	Applies to products that collect/store personal data (identifiable information) in electronic or non-electronic formats.
IMDA TR 47: 2016 – IoT Reference Architecture	Relevant for DMH products using IoT or sensor networks; sets baseline for interoperability across industries.
IMDA TR 50: 2016 – IoT Information & Services Interoperability	Relevant for IoT-enabled DMH products; provides technology-independent reference architecture for seamless data exchange.

(DiGA) and Their Implementation in the German DiGA Directory: An Analysis." *BMC Health Services Research*: 369. https://doi.org/10.1186/s12913-023-09287-w.

[28] *NHS England.* 2025. Digital Clinical Safety Assurance, version 1.2. Accessed December 25, 2025. https://www.england.nhs.uk/long-read/digital-clinical-safety-assurance.

[29] Infocomm Media Development Authority. 2025. *Standards and Guidelines*. Accessed December 25, 2025. https://www.imda.gov.sg/regulations-and-licences/regulations/standards-and-guidelines.

IMDA TR 64: 2018 – IoT Security Guidelines	Relevant for IoT-based DMH tools; outlines threat modelling, four design principles, and requirements for IoT system security.
IMDA TR 33: 2013 – Data-as-a-Service (DaaS) API Design	Applies to products sharing or monetising data as services; provides API best practices and implementation checklist.
IMDA TR 41: 2015 – Data Quality Metrics	Relevant to DMH products collecting patient data; sets baseline, domain-agnostic data quality metrics for interoperability and meaningful insights.
IMDA TR 55: 2016 – Data Versioning	Relevant to DMH tools handling patient datasets; ensures data integrity and comparability across evolving datasets.

Option 1: Maintain Status Quo of Industry Self-Regulation

Under a self-regulatory model, AI developers would continue to define and implement their own safety measures, including content moderation rules, crisis-response prompts, and parental controls. Governments would refrain from imposing formal accreditation or external oversight. This approach mirrors the early phase of the DMH app ecosystem, where voluntary standards and market-driven quality signals were relied upon to protect users.[30] Its main advantage lies in preserving innovation and adaptability in a rapidly evolving field.

However, experience from DMH regulation suggests clear limitations. Voluntary safeguards are often unevenly implemented, lack transparency, and are not systematically evaluated for effectiveness.[31] Without external validation, there is no assurance that safety mechanisms function reliably in high-risk scenarios. Although industry self-regulation offers speed and flexibility, it leaves significant gaps in user protection and does little to address concerns raised by clinicians, parents, and adolescent health advocates about safety and wellbeing.

Option 2: Age-Based Restrictions and Enhanced Safety Filters

A second policy pathway would focus on targeted protection for adolescents and teenagers, mandating age-verification mechanisms, parental controls, and stricter conversational constraints for under-21 (which is the age of majority in Singapore) users. This could include conservative response styles, limits on emotionally immersive dialogue, and prompts encouraging offline support.

This approach aligns with child-protection strategies in other digital domains and is politically feasible.[32] It echoes DMH usability and engagement criteria, which recognise that design features must be appropriate to users' developmental stages.[33]

However, experts have warned of the limitations of standalone age-based safeguards. Age verification can be inaccurate or circumvented, and risk does not disappear abruptly after legal age

[30] Babu, Anithamol, and Akhil P. Joseph. "Digital Wellness or Digital Dependency? A Critical Examination of Mental Health Apps and Their Implications." *Frontiers in Psychiatry* (2025): 1581779. https://doi.org/10.3389/fpsyt.2025.1581779.
[31] Hyzy, Matthew, Raymond Bond, Maurice Mulvenna, Li Bai, Alan Dix, Ronan Daly, Anne-Laure Frey, and Sarah Leigh. 2023. "Quality of Digital Health Interventions Across Different Health Care Domains: Secondary Data Analysis Study." *JMIR mHealth and uHealth*: e47043. https://doi.org/10.2196/47043.
[32] NHS England. 2025. Digital Clinical Safety Assurance, version 1.2. Accessed December 25, 2025. https://www.england.nhs.uk/long-read/digital-clinical-safety-assurance.
[33] Mental Health Commission of Canada. 2023. *Assessment Framework for Mental Health Apps*. Accessed December 25, 2025. https://mentalhealthcommission.ca/wp-content/uploads/2023/06/MHCC-Assessment-Framework-for-Mental-Health-Apps-EN-FINAL.pdf.

thresholds.[34] Moreover, this option does not address the functional risk of chatbots that provide emotionally influential guidance, regardless of the user's age. These concerns have created a shared dilemma for many countries contemplating protection for children and adolescents in the digital era where pure bans or access restrictions fall short.[35] Age-based restrictions may reduce exposure but do not ensure reliable safety governance.

On balance, however, blanket restrictions can be counterproductive as they can inadvertently encourage circumvention and push young people toward less regulated spaces. As a result of a strict ban on certain platforms, the policy may unintentionally displace rather than reduce risk, especially for youths who already face barriers to timely, trusted support for their mental health struggles.[36]

Singapore is arguably well-positioned to implement a graduated safety architecture because it already has a mature national digital identity infrastructure and all residents are equipped with *SingPass* (short for Singapore Personal Access), a secure digital identification system enabling residents to authenticate their identity online.[37] A pragmatic policy pathway could mandate that any chatbot offering "youth-facing" conversational modes in Singapore must support verified age assurance without storing unnecessary identity data, and must default minors into the appropriate tier. For example, for children (e.g., <13 years old), the architecture would prevent access to open-ended "companion-style" chat and grant access only to curated, pedagogically supported or task-bounded modes (e.g., homework help with constrained dialogue), with strict minimisation of persuasive/relational features. For early adolescents (e.g., 13 to 15 years old), access to these chatbots would be allowed, but with stronger defaults, that is, reduced "relational" design (less emotionally immersive dialogue), alerts before sensitive topics, prominent nudges toward offline supports, and stricter limits on advice-like language. For older adolescents (e.g., 16 to 20 years old), they would be granted broader access, but still with youth-specific safety-by-design settings and clearer transparency, emphasising that protection needs shift with development rather than disappearing at a single legal threshold.

Such a tiered model adapts existing risk-tiering and lets policymakers regulate design intensity and functional affordances (e.g., whether the system can act as an "emotional companion") rather than only policing individual outputs. Importantly, it does not entirely exclude adolescents or children from these platforms but rather enable regulated participation that is safer, feasible and proportionate to foreseeable risk for at least the vast majority. However, the trade-off is regulatory burden. Compliance costs could be substantial, particularly for smaller developers, and the pace of innovation may slow. There is also a risk of over-regulating general-purpose technologies (that presumably falls into Tier 1) that only occasionally intersect with mental health contexts.

Option 3: Mandatory Crisis Detection and Referral Pathways

Related to the previous option, another policy pathway would require AI chatbots to implement mandatory detection of crisis-related language (such as expressions of suicidal ideation or self-harm)

[34] Köhler-Dauner, Franziska, et al. "Digital Child Protection in Social Networks: Age Verification and Age-Tiered Regulation in Europe." *Child and Adolescent Psychiatry and Mental Health* (2025): 143. https://doi.org/10.1186/s13034-025-01016-x.
[35] Köhler-Dauner, Franziska, et al. "Digital Child Protection in Social Networks: Age Verification and Age-Tiered Regulation in Europe." *Child and Adolescent Psychiatry and Mental Health* (2025): 143. https://doi.org/10.1186/s13034-025-01016-x.
[36] Köhler-Dauner et al., "Digital Child Protection in Social Networks."
[37] GovTech Singapore. *Singpass.* November 19, 2025. Accessed December 25, 2025. https://www.tech.gov.sg/products-and-services/for-citizens/digital-services/singpass.

and to redirect users toward human support systems, including crisis hotlines, trusted adults, or professional services. This approach aligns closely with established clinical norms in mental health care, where duty-of-care principles require escalation when risk is identified.[38]

This strategy could help provide additional safeguards and ensure that users receive timely professional intervention and reduce reliance on conversational agents that are not designed to provide therapeutic assessment or crisis management. However, this option raises several technical and ethical challenges. Natural language processing models are vulnerable to misinterpretation, which may generate false reassurance or false alerts. Crisis language may also be culturally mediated or masked, complicating algorithmic detection.[39] Privacy concerns also arise, particularly if crisis detection involves data retention or reporting. Mandatory logging or triage may also raise privacy and consent concerns. Policymakers would therefore need to carefully define thresholds for escalation and ensure proportionality in response.

Nevertheless, the emergence of products such as ChatGPT Health strengthened the case for risk-tiered governance and hybrid licensing models (Option 4), as it demonstrated both the need for stronger safeguards and the risks of leaving emotionally and clinically influential systems outside formal oversight simply because they are embedded within broader consumer platforms.

Option 4: Hybrid Licensing Model with a Public Registry

Rather than treating all AI chatbots as DMH tools by default, a hybrid licensing model offers a proportionate regulatory approach that aligns oversight with risk. Under this model, conversational AI systems that provide sustained emotional guidance, mental health advice, or therapeutic-style interactions would be subject to accreditation when their functionality and use cross clearly defined risk thresholds, while general-purpose chatbots would continue to operate under lighter regulatory requirements.

Under such a framework, AI chatbots would be regulated along a tiered continuum. Low-risk, general-purpose chatbots such as productivity tools or educational assistants would remain subject primarily to transparency obligations and general consumer protection laws, without additional health-specific requirements. Moderate-risk chatbots, characterised by sustained emotional engagement or routine responses to mental health–related disclosures, would be required to meet baseline safety and privacy standards. These could include documented crisis-handling logic, conservative response design, and regular internal safety audits. High-risk chatbots, particularly those marketed or routinely used as sources of emotional guidance, coping support, or mental health advice, would be subject to formal accreditation comparable to existing DMH tools. Accreditation requirements would include evidence of safety, clearly articulated escalation and referral pathways, clinical or psychosocial governance arrangements, and mechanisms for post-market monitoring.

To operationalise this framework, relevant government agencies such as the Ministry of Health (MOH), the Health Sciences Authority (HSA), and the Infocomm Media Development Authority (IMDA) could jointly establish and maintain a public registry of accredited mental health chatbots. Accreditation would be based on transparent criteria drawn from existing DMH evaluation frameworks, including evidence of safety, crisis-response capabilities, data protection standards, and appropriate clinical or psychosocial governance. This inter-agency model mirrors Singapore's

[38] Tavory, Tamar. "Regulating AI in Mental Health: Ethics of Care Perspective." *JMIR Mental Health* (2024): e58493. https://doi.org/10.2196/58493.
[39] Holmes, Glenn, et al. "Applications of Large Language Models in the Field of Suicide Prevention: Scoping Review." *Journal of Medical Internet Research* (2025): e63126. https://doi.org/10.2196/63126.

broader approach to regulating health-adjacent technologies, where risk-tiered oversight (Table 1) was preferred over blanket regulation.

A publicly accessible registry would also serve as a trusted reference point, enabling parents, educators, and young users themselves to distinguish between chatbots that have undergone formal safety evaluation and those that have not. In doing so, it shifts some regulatory power toward informed choice, supporting parental mediation and shared decision-making rather than relying solely on prohibition or technical controls. For adolescents, many of whom value autonomy and privacy, the availability of accredited tools offers safer alternatives without denying access to digital support altogether. Public-sector settings such as schools, healthcare institutions, and youth services could be required to use only accredited chatbots, ensuring that state-endorsed environments do not expose young people to unvalidated or high-risk systems.

The hybrid model also creates market incentives. Given the sheer number and rapid turnover of AI chatbots and apps, it would be neither feasible nor desirable for regulators to screen the entire marketplace exhaustively. Instead, accreditation can function as a voluntary but powerful signal of quality. Developers who wish to operate in sensitive mental health contexts or reach young users at scale would be encouraged to pursue accreditation, while innovation in lower-risk, general-purpose applications could continue relatively unimpeded. At the same time, the approach avoids over-medicalising all conversational AI, addressing industry concerns about excessive regulatory burden.

However, the limitations of this model must be acknowledged. Effective implementation would require clear definitions of risk thresholds, sustained regulatory capacity to assess and monitor compliance, and coordination across agencies. Moreover, accreditation alone does not guarantee use. Some adolescents may continue to rely on unaccredited tools due to familiarity, peer influence, or perceived anonymity. As such, the protective value of a registry would depend heavily on complementary public education efforts targeting parents, schools, and young people themselves, to raise awareness of digital risks and promote the use of accredited alternatives. In this sense, a hybrid licensing model is not a standalone solution but part of a broader governance ecosystem. When combined with digital literacy education, parental guidance, and age-appropriate design obligations, it offers a pragmatic pathway that balances child protection, adolescent autonomy, and technological innovation within Singapore's regulatory culture.

Conclusion

The case of Adam Raine exposed an uncomfortable reality at the intersection of youth mental health and artificial intelligence. Conversational AI systems were no longer peripheral tools; for some adolescents, they functioned as emotional companions, confidants, and informal guides during periods of vulnerability. Yet the complexity of these systems and the heterogeneity of their uses meant that a single, blunt regulatory response was unlikely to be effective. For Singapore, the most defensible and workable policy pathway lies in a hybrid, risk-tiered governance model that adapts established DMH accreditation principles while preserving space for innovation. The central policy question is not whether AI chatbots should exist, but how society ensures that their psychological influence does not outpace available safeguards, particularly for vulnerable young people. A risk-based approach allows regulation to be proportionate, targeted, and responsive to evolving technological realities. At its core, this case raised a question of accountability. As technology increasingly mediates emotional and cognitive spaces once occupied by families, peers, and clinicians, the absence of clear governance is itself a policy choice. Singapore has an opportunity to act early and deliberately: to provide meaningful protection for young users without stifling innovation or over-regulating low-risk technologies. The decisions made at this stage will determine whether AI becomes a trusted complement to mental health support—or a preventable source of harm at moments when protection matters most.

Appendix

Table S1: High-level summary of international DMH frameworks and accreditation standards

Framework (Country)	NICE ESF (United Kingdom)	NHS DTAC (United Kingdom)	DiGA (Germany)	NSQDMH (Australia)	MHCC (Canada)	DHAF (United States)	ORCHA (Independent)
Clinical effectiveness	Tiered "minimum vs best practice" evidence by function/risk; permits comparative studies/RCTs; requires validated outcomes	Emphasises doing ESF alongside DTAC; local buyers must "consider efficacy and impact"	Must show positive healthcare effects; permanent listing requires sufficient evidence; provisional listing allows evidence generation	Model of Care standard expects evidence-based practice within service pathways (59 actions)	Dedicated Clinical evidence standard tailored for mental-health apps	Clinical assurance & safety domain requires evidence appropriate to app complexity and guideline alignment	PCA (Professional & Clinical Assurance) domain contributes to overall score
Safety & clinical governance	Flags proportionality & contextual risks; complements regulatory safety requirements	Clinical safety (DCB0129/0160) is a core pillar	Quality, suitability, data protection & information security are prerequisite; ongoing obligations after listing	Clinical & Technical Governance standard	Dedicated Clinical safety standard	Clinical assurance & safety is one of four domains	Safety reviewed within PCA; re-reviewable
Privacy & security	Covered implicitly	Data protection + Technical assurance (cybersecurity); interoperability security controls	Security, privacy, information security mandated in DiGA checklists	Governance actions include privacy protections	Strong Data & privacy and Security & technical stability standards	Data & privacy and Technical security & stability are core domains	Data privacy is one of three scored pillars as part of OBR
Usability & user	Requires user involvement and	Usability & accessibility receives	Considered within suitability for use;	Partnering with Consumers +	Dedicated Usability &	Usability & accessibility domain	UX is one of three

engagement	satisfaction evidence	conformity rating	patient-facing quality expectations	service design	accessibility standard		scored pillars
Economic & system value[1]	Formal economic impact standards: key economic info + budget impact, CEA/CUA	DTAC defers to ESF for "efficacy and impact"	Positive healthcare effects underpin reimbursement; price negotiation follows	Focus on quality/safety rather than economic modelling	Not an economic evaluation framework; can reference value statements	Not an economic evaluation framework	Not an economic evaluation; quality-assurance signal for adopters
Equity, ethics & cultural fit	Equality/inequalities considered	Not explicit beyond accessibility	Not explicit; general patient-benefit focus	Consumer voice & safety embedded	Dedicated Cultural safety, equity & enhanced data soverei gnty	Not explicit; included under usability	Not explicit; assessed indirectly
Interoperability & integration[1]	Not a tech standard; aligns with national standards	Interoperability is a core pillar	Requires interoperable export interfaces and device connectivity	Addressed within technical governance	Optional linkage to EHRs; security checks	Covered under technical stability/security expectations	Considered in OBR questions
Ongoing monitoring & listing	Iterative evidence development; complements regulation	Local re-assessment of expiring sections	Trial (provisional) listing (12 months), study publication, change notification; can be delisted	National register of accredited services	Framework + planned assessment programme with ORCHA	Marketplace assurance framework	Regular re-reviews; ≥65% to surface in formularies

[1] Economic and system value criteria and interoperability and integration, while important for reimbursement and large-scale procurement decisions, are of limited relevance to the present case, which concerns baseline safety, governance, and protection of vulnerable youth users rather than health system cost-effectiveness.

Abbreviations:

NICE ESF – National Institute for Health and Care Excellence Evidence Standards Framework

NHS DTAC – National Health Service Digital Technology Assessment Criteria

DiGA – Digitale Gesundheitsanwendungen ("Digital Health Applications"), Germany's app-on-prescription fast-track program

NSQDMH – National Safety and Quality Digital Mental Health Standards

MHCC – Mental Health Commission of Canada App Assessment Framework

DHAF – Digital Health Assessment Framework

ORCHA – Organisation for the Review of Care and Health Apps

RCTs – Randomized Controlled Trials

BfArM – Bundesinstitut für Arzneimittel und Medizinprodukte

CEA/CUA – Cost-Effectiveness Analysis / Cost-Utility Analysis

PCA – Professional & Clinical Assurance

UX – User Experience

OBR – ORCHA Baseline Review

EHRs – Electronic Health Records