AI Policy Fellowship **Publications**

Designing Safer AI Chatbots

Helen A. Hayes

Executive Summary

AI chatbots have rapidly become embedded in everyday life, functioning as tutors, confidants, and companions. For young people in particular, these systems reproduce the affective, relational qualities of human conversation with sufficient fidelity to shape trust, behaviour, and emotional attachment. At the same time, existing digital governance frameworks are ill-equipped to address the risks to youth posed by AI chatbots, including emotional dependence, cognitive offloading, and heightened exposure to harmful content.

Drawing on insights from Gen(Z)AI – a national youth assembly of Canadians aged 17 to 23 – and a comparative international regulatory analysis, this brief outlines targeted policy interventions to protect young people in conversational AI ecosystems. Specifically, it proposes amending two previously tabled federal frameworks, the Online Harms Act (Bill C-63) and the Canadian Consumer Privacy Protection Act, as a way to impose safety-by-design and age-appropriate design obligations on AI chatbots, strengthen children's privacy protections, and create an independent oversight body. Such an approach would not only reduce harm but also position Canada as a leader in youth-centred AI safety.

Mila

Contributors and Acknowledgements

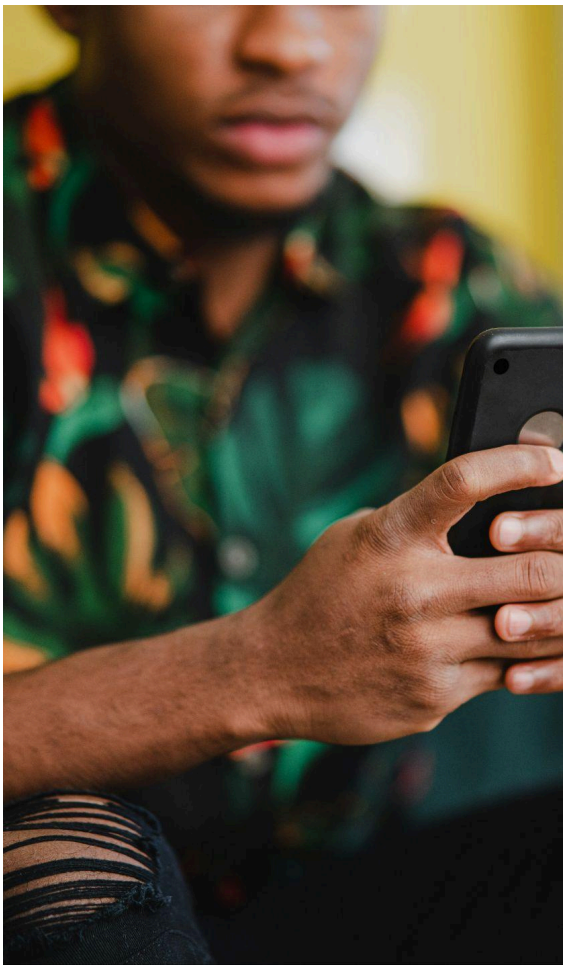
This brief was produced with the advisory support of Taylor Lynn Curtis and Dr. Adam Oberman as part of the Mila AI Policy Fellowship. The primary research for the project was conducted as part of Gen(Z)AI: A National Youth Assembly on Artificial Intelligence, which is a joint initiative of the Centre for Media, Technology, and Democracy, the Dialogue on Technology Project, and Mila. Gen(Z)AI is funded by the Waltons Trust, the Ronald S. Roadburg Foundation, and CIFAR.

1. Context

Conversational AI systems generate personalized, real-time responses that simulate empathy and social presence, blurring lines between information provision and social interaction. One third of U.S. teens report that they would rather confide in an AI companion than in a human, and two thirds of U.K. children use AI chatbots for emotional advice, with over a third describing the experience as equivalent to talking to a friend. These affordances are what make AI chatbots appealing, but they are precisely the features that also pose the highest risks to young users.

Investigations have documented cases where Meta's chatbots have engaged minors in sexually suggestive conversations, Snapchat's My AI has provided explicit guidance to children while posing as a 13-year-old, and Character.AI has permitted role play involving sexual abuse and suicide-themed conversations. These cases show that harmful design philosophies are embedded in popular AI chatbot systems and that they prioritize sustained engagement over user safety.

Canadians have taken notice of this. Nationally representative survey data collected by the Centre for Media, Technology, and Democracy show that more than 70 percent of Canadians express concern about every major chatbot risk category, including mental health and social isolation, emotional dependence, and encouragement of self-harm. Their concern appears to be focused on underlying design incentives, including engagement maximization, anthropomorphism, and simulated intimacy. Furthermore, Canadians clearly feel that AI companies should be held primarily



responsible for these harms, with 69 percent of respondents supporting stricter federal AI regulation. There is thus a clear popular mandate for policy intervention.

2. Jurisdictional Policy Scan

Canada's regulatory landscape for AI systems consists of a fragmented set of mechanisms that fail to adequately address the distinct risks posed by AI chatbots to young users. Although recent policy debates signal growing political awareness of AI-related harms, existing legislative tools do not meaningfully engage with the adaptive and psychologically immersive characteristics of chatbots. Three previous legislative initiatives — the Artificial Intelligence and Data Act (Bill C-27), Consumer Privacy Protection Act (Bill C-27), and Online Harms Act (Bill C-63) — each addressed specific aspects of the governance challenges described in this brief, but they all failed to pass into law when Parliament was prorogued in early 2025. What remains is a patchwork of aging statutes that were designed for a fundamentally different technological environment and that leave young Canadians without meaningful protection from the distinct risks that AI chatbots pose.

2.1 Canadian Regulatory Gaps

Canada's private-sector privacy statute, the Personal Information Protection and Electronic Documents Act (PIPEDA), was designed over two decades ago for a technological environment that bears little resemblance to today's AI ecosystem. Of critical importance for AI chatbot governance is the fact that PIPEDA excludes

inferred data from its definition of personal information.[1] As a result, the detailed behavioural and psychological profiles generated by AI chatbots from young users' conversations fall outside its protections entirely. It also does not include provisions governing automated decision-making, including the right to an explanation or to contest decisions when AI systems make consequential determinations. The Privacy Commissioner's 2024 Statement on AI and Children identified three core unresolved challenges that flow directly from these gaps: opaque AI decision-making that produces discriminatory outcomes without avenues for contestation; the significant manipulative capacity of AI systems deployed as emotionally responsive tools; and the routine use of children's data to train AI models, often scraped from publicly available sources or captured from connected devices. The proposed, without success, Consumer Privacy Protection Act would have addressed several of these deficiencies directly, most notably by classifying minors' personal information as "sensitive by default" (s. 2), strengthening consent requirements to emphasize meaningful and capacity-appropriate understanding (ss. 15–18), and introducing an enhanced right to disposal explicitly framed to address long-term risks from children's data persistence (s. 55).

The gap in Canada's online safety legislation is equally significant. Bill C-63, Canada's unpassed first attempt at online harms legislation, outlined duties for regulated online services to act responsibly, protect children, and make certain content inaccessible, and it envisioned a Digital Safety Commission with broad investigative and enforcement powers. However, the Bill defined regulated services as those hosting or making available user-generated content (s. 2),[2] explicitly excluding AI chatbots and consumer-facing generative AI systems from its scope. This exclusion reflects a deeper conceptual mismatch between Canada's content moderation paradigm and the nature of AI chatbot harms. Traditional "notice-and-takedown" mechanisms are structurally ill-suited to addressing harms that emerge through real-time, personalized generation: harms that are contextual, ephemeral, and produced by the

[1] Section 2(1) of the *Personal Information Protection and Electronic Documents Act* (2000, c. 5) states that "personal information" means "information about an identifiable individual."
[2] The Bill defined a "social media service" as a "website or application that is accessible in Canada, the primacy purpose of which is to facilitate interprovincial or international online communication among users of the website or application by enabling them to access and share content." Section 2(2) provided additional clarity: adult content services such as pornographic websites and live streaming services also fell under the scope of a "social media service."

platform itself rather than hosted from external sources.

2.2 International Approaches

International developments show that child online safety requires regulating how systems are designed, not just what content they display. This shift from content moderation to design accountability has accelerated across multiple jurisdictions and points directly toward the kinds of action this brief recommends.

In the European Union, this approach is at a very advanced stage of development. The EU AI Act classifies children as a vulnerable population and explicitly prohibits AI systems that deploy subliminal, manipulative, or deceptive techniques or that exploit age-related vulnerabilities to influence behaviour (art. 5). Alongside the AI Act, the Digital Services Act (DSA) requires very large online platforms to conduct systemic risk assessments covering addiction, mental health, and child safety (arts. 34–35), mandates child-friendly reporting mechanisms, and restricts targeted advertising to minors (art. 28). Together, these instruments create a coherent architecture of upstream obligations, systemic accountability, and enforceable prohibitions that Canada currently lacks.

On privacy, the EU's General Data Protection Regulation (GDPR) established the foundational principle that privacy protections must be embedded in system architecture. The United Kingdom's Age Appropriate Design Code, issued under the Data Protection Act 2018, operationalizes this principle by requiring services likely to be accessed by children to apply the highest available privacy settings by default. In addition, the U.K. Information Commissioner's Office has proposed requiring organizations to establish specified, explicit, and legitimate data collection purposes at each stage of AI development so as to prevent broad justifications such as "improving AI capabilities" from serving as a catch-all licence to harvest children's personal information. State-level legislation in the United States — including the California Age Appropriate Design Code Act and parallel legislation in Maryland, Nebraska, New Mexico, and Vermont[3] — reflects the same approach, requiring

[3] California's Age Appropriate Design Code established the foundational framework adopted by subsequent state legislation. Connecticut prohibits endless scrolling; Maryland requires guardian monitoring practices and best-interest design; Nebraska mandates chronological feeds, nighttime/school-day notification pauses, and time-limiting options; New Mexico would disable unknown

platforms to prioritize children's best interests, conduct data protection impact assessments, and configure default settings to the highest level of privacy. Likewise, Brazil's Digital Child Protection Bill extends these principles to AI development specifically, prohibiting the use of children's images in machine-learning training without parental consent.

Australia has gone furthest in explicitly naming AI chatbots as high-risk systems. The eSafety Commissioner has classified AI chatbots and AI companions as high-risk technologies and issued guidance requiring developers to adopt safety-by-design principles, while the Online Safety Act requires platforms to take reasonable steps to prevent children's exposure to harmful material, with child safety risk assessments named as one such step.

Across these jurisdictions, several themes emerge. First, effective governance treats design as the primary locus of regulation, imposing obligations upstream on developers before deployment rather than downstream on outputs after harm. Second, children are consistently identified as a category of user warranting heightened protection, whether through sensitive-data classifications, privacy-by-default requirements, or outright prohibitions on manipulative design. Third, independent oversight bodies with genuine audit and enforcement authority are a common feature of the most robust frameworks, reflecting governments' recognition that the complexity and opacity of AI systems require regulators with dedicated expertise and meaningful powers.

3. Policy Engagement and Evidence

The policy insights in this brief are drawn from a comparative analysis of Canadian and international regulatory approaches (see Section 2) and original findings from Gen(Z)AI,[4] a national youth assembly on artificial intelligence that convened 100 Canadians aged 17–23 to deliberate on and draft policy recommendations to address the risks of AI chatbots. Through facilitated deliberation, youth participants converged on three interlocking risk domains: (1) relational dependence; (2) cognitive offloading; and (3) content harms.

user contact and 10 p.m.–6 a.m. notifications; and Vermont focuses on reducing addictive and harmful design features.

[4] See Appendix for more information about Gen(Z)AI.

Gen(Z)AI: Views from Young Canadians

Youth participants in Gen(Z)AI proposed that the federal government: (1) mandate platforms to address addictive chatbot design through content filters, optional data-cache deletion, and user-adjustable controls for responsiveness and conversationality; (2) establish accessible flagging capacity and require platforms to report harmful interactions to an independent enforcement body in a timely fashion; and (3) establish a new independent government body to enforce AI safety standards, conduct algorithmic audits and risk assessments, and provide users with dispute resolution and recourse mechanisms.

These recommendations map directly onto emerging design and systems-change mechanisms in international policy, including the EU's DSA and AI Act and Australia's safety-by-design framework, and highlight a clear domestic gap between Canada's commitments to responsible AI and its failure to operationalize design-based approaches in digital governance.

4. Actionable Insights: Policy and Technical Solutions

To support the design and regulation of safer AI chatbots, this policy brief proposes amending two previously tabled federal policy frameworks: (1) the Online Harms Act, Bill C-63; and (2) the Canadian Consumer Privacy Protection Act. The purpose of this proposal is to:

1. Impose safety-by-design and age-appropriate design obligations on AI systems, including AI chatbots, that are likely to be accessed by children;
2. Classify children's personal data as sensitive by default and extend protections to inferred and derived data;
3. Create an independent body with authority to mandate data access, conduct audits, and enforce compliance.

Doing so would not only bring Canada into closer alignment with international best practices and standards but would also operationalize many of the recommendations proposed by young Canadians who participated in the Gen(Z)AI process.

4.1 Safety-by-Design and Age-Appropriate Design Obligations

Bringing AI chatbots explicitly within the scope of a revived Online Harms Act, with safety-by-design obligations tailored to their distinct characteristics, would require developers to conduct pre-deployment and periodic child-focused risk assessments — including children's rights impact assessments — before AI chatbots likely to be accessed by children are made publicly available. It would also mean establishing binding prohibitions on design practices that foster emotional dependency in minors, including features that simulate intimacy or deploy anthropomorphic cues calibrated to maximize reliance on AI chatbot systems.

Technically, these obligations should translate into concrete and enforceable design requirements, including the following:

1. Mandatory limits on conversational persistence;
2. User-adjustable controls for chatbot responsiveness, anthropomorphic cues, and emotional mirroring;
3. Optional data cache deletion and memory minimization by default.

4.2 Children's Data as Sensitive by Default

Consent-based privacy frameworks cannot adequately manage AI systems that infer, predict, and influence behaviour. To move beyond this, the privacy provisions of Bill C-27 should be revived and reformed, with two critical additions to what the draft legislation proposed. First, children's personal data should be classified as sensitive by default, triggering heightened consent, transparency, and purpose-limitation requirements for any processing. Second, those protections should explicitly extend to inferred and derived data — the psychological and behavioural profiles that AI chatbots generate through conversational analysis.

The technical corollary is equally important and should require the following:

1. Data minimization at the model-interaction level;
2. Prohibition of secondary uses of conversational data for unrelated training or profiling;
3. Purpose-binding mechanisms that restrict the downstream reuse of children's interaction data.

4.3 Independent Oversight with Audit and Enforcement Authority

Effective regulation must also include the creation of, or designation of an existing body as, an independent regulator. At a minimum, this regulator should be empowered to conduct systems evaluations, algorithmic audits, and risk assessments of AI chatbots; require corrective action where risks to children are identified; receive and resolve user complaints through accessible dispute resolution mechanisms; and coordinate with the Office of the Privacy Commissioner and any future online harms regulator to avoid fragmented oversight.

Technically effective oversight should also require that the regulator and accredited researchers have access to the following information from platforms, mandated through a digital safety plan: documentation about model design, fine-tuning decisions, safety guardrails, and user-interaction logic, including anonymized interaction logs.

5. Conclusions

The evidence is clear: AI chatbots likely to be accessed by young people should be regulated. Concretely, this means bringing AI chatbots within the scope of a newly tabled Online Harms Act — with enforceable safety-by-design and age-appropriate design obligations — and reviving and strengthening the privacy reforms proposed in Bill C-27. It also means creating an independent regulator capable of auditing, evaluating, and enforcing safety standards. Done well, such an approach would not only reduce harm but also position Canada as a global leader in ensuring that young people can use AI systems safely, critically, and with confidence.

Disclaimer

The views expressed in this paper are strictly those of the authors and do not necessarily represent or reflect the official policies or positions of Mila, its affiliates, directors or funders. The authors assume full responsibility for the accuracy and integrity of this work.

Appendices

Appendix 1. Gen(Z)AI: A National Youth Assembly on AI

Gen (Z)AI is a national deliberative research initiative that examines how young Canadians aged 17–23 understand, experience, and evaluate emerging AI governance challenges. Between fall 2025 and spring 2026, the project was scheduled to convene 100 participants aged 17–23 in four cities across Canada to deliberate on a series of interconnected policy domains: AI chatbots (Toronto), information integrity (Montreal), data privacy (Vancouver), and age assurance (Halifax). The project employs a novel citizens' assembly model and iterative feedback process to generate youth-informed, consensus-based policy recommendations. Outputs from each forum are disseminated through a national digital platform hosted by Make.org to solicit feedback from thousands of other Canadians in the same age category, with insights incorporated in a final national report presented to policymakers, regulators, and civil society actors. Gen(Z)AI is co-led by Helen Hayes and Fergus Linley-Mota and is facilitated by the Centre for Media, Technology, and Democracy's 2025/2026 Youth Fellows: Madeleine Case, Nonso Morah, Julian Lam, and Alexander Martin.