

How to Catalyze Policy Guardrails against AI Chatbots' Potential Social-Emotional Harm: Lessons from Analogous Cases (Extended Abstract)

Ned Cooper, Ibrahim Emara, Qian Yang

Cornell University
{ned.cooper, ie68, qianyang}@cornell.edu

Could extensive social-emotional use of AI chatbots erode users' human connections over time? Could it worsen the already severe loneliness crisis? Research in controlled settings increasingly suggests that such slow-burning harms are plausible (Fang et al. 2025; Liu, Pataranutaporn, and Maes 2025; Bran Herbener 2025; Ibrahim et al. 2026); if these harms are occurring in the real world, their consequences would be severe and far-reaching. However, such harm is hard to prove or attribute outside controlled settings. Loneliness has many influences, and the design of chatbots, the capabilities of models that power them, and usage patterns all change frequently (Yuan et al. 2026). Conclusive evidence that could propel large-scale public mental health or regulatory interventions is therefore unlikely to emerge until the harm has become entrenched and harder to reverse (Odgers and Jensen 2020).

We ask: How might human-computer interaction (HCI) and policy communities catalyze policy guardrails against the potentially harmful effects of AI chatbots on user loneliness, *before conclusive causal evidence of such slow-burning, diffuse harm arrives?*

As a first step to answering this question, we analyze past and ongoing regulatory cases that are analogous to the present one, and ask what we can learn from them. These are cases where a product's harm is (i) probable but contested (either the harm itself or its attribution to the specific product), (ii) slow-burning and diffuse, and (iii) closely linked with industry actors' financial interests or business models, all of which make regulatory interventions difficult.

We analyzed five such cases in depth: regulating social media, tobacco, and prescription opioids in the U.S., youth digital gaming addiction in South Korea, and problem gambling in Australia. Notably, this is not a comprehensive set of past or ongoing cases that meet the criteria, nor does what worked (or did not work) for these cases automatically predict what will work for AI chatbots. Nevertheless, they offer a useful starting point for discussing how we might break the impasse in regulating AI chatbots' potential social-emotional harm.

Our analysis suggests three elements often preceded durable policy guardrails: (i) evidence of the harm's public cost, i.e., the cost it imposes on public budgets (such as

the fiscal burden the opioid crisis imposed on Medicaid); (ii) a civic narrative framing the harm as a matter of product design and industry conduct rather than users' individual responsibility; and (iii) institutions with dedicated funding that make monitoring of the harm durable.

In the case of AI chatbots, however, these elements are not yet in place. We reviewed 62 public-facing governance documents from six companies deploying general-purpose chatbots and purpose-built companion or mental well-being apps for passages addressing mental health, well-being, or user relationships. Almost all of the commitments we identified that address chatbots substituting for human interaction operate within a single message, session, or day; none describes a mechanism for detecting gradual erosion of users' social connections across sessions. Recent state laws in the U.S. addressing social-emotional use of AI operate at a similar level. These instruments are therefore ill-suited to accumulating evidence of the public cost of chatbots' slow-burning harm.

Based on these findings, we discuss pathways through which HCI and policy communities might break the impasse in regulating AI chatbots' potential social-emotional harm. In particular, we call for (i) public infrastructure that measures the public cost of such harm, (ii) novel mechanisms to fund such measurement infrastructure, and (iii) continued research on AI chatbot/companion use as a matter not (just) of users' personal choice but of public health.

Acknowledgments

We thank reviewers for their thoughtful comments and suggestions. This research was supported by the Cornell Health Policy Center Seed Grants. Ned Cooper and Qian Yang were supported in part by the AI2050 Early Career Fellowship program at Schmidt Sciences. Ned Cooper is also supported by the Cornell Digital Literacy Initiative (DALI).

References

- Bran Herbener, L. 2025. Chatbots as emotional support: Correlates of loneliness among adolescent users. *International Journal of Human-Computer Studies*.
- Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W. T.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; and Agarwal, S. 2025. How AI and human behaviors

shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. arXiv:2503.17473.

Ibrahim, L.; Hafner, F. S.; Cheng, M.; Lee, C.; Anselmetti, R.; Willer, R.; Rocher, L.; and Yang, D. 2026. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. arXiv:2605.07912.

Liu, A. R.; Pataranutaporn, P.; and Maes, P. 2025. The heterogeneous effects of AI companionship: An empirical model of chatbot usage and loneliness and a typology of user archetypes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1585–1597.

Odgers, C. L.; and Jensen, M. R. 2020. Annual Research Review: Adolescent mental health in the digital age: facts, fears, and future directions. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 61: 336–348.

Yuan, Y.; Zhang, J.; Aledavood, T.; Zhang, R.; and Saha, K. 2026. Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Lens. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.