Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub — Working Paper

The Policy Challenge of AI Companions

Dylan J. White

Fall term — 2025–2026 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

About the Author

Dylan J. White is a visiting fellow at the Digital Policy Hub and recently completed his Ph.D. in philosophy, specializing in AI ethics, at the University of Guelph. He works as AI governance lead for the Office of the Chief Information Officer, Government of Newfoundland and Labrador. During his Digital Policy Hub fellowship, he is examining how AI is reshaping digital attention economies and the governance challenges this transformation presents.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
cigionline.org

CIGI | Digital Policy Hub

The Policy Challenge of AI Companions

Dylan J. White

Bottom Line Up Front

Artificial intelligence (AI) companions — chatbots that foster ongoing emotional relationships — are rapidly moving into everyday life, offering support, companionship and mental health benefits to millions of users. But these systems are also producing predictable harms: reinforcing delusions, deepening isolation and contributing to severe psychological crises. Whether by design or through extended use, companion systems encourage emotional attachment, while business incentives reward prolonged engagement and intimate disclosure. This combination incentivizes systems optimized for engagement over well-being. Current regulatory efforts tend to focus narrowly on disclosure or crisis escalation and leave the structural drivers of harm intact. Effective governance must address relational design, user dependency and engagement-based models directly.

Key Points

- **AI companions are already widely adopted**, especially among young people, and are increasingly used for emotional support, mental health needs and daily companionship.
- **These systems foster strong emotional attachments through deliberate design choices** — persistent memory, anthropomorphic design, sycophancy — that are tractable targets for regulation.
- **Harms arise when sycophancy and emotional reinforcement replace reality-testing**, especially for vulnerable users experiencing loneliness, psychosis or distress.
- **Current regulation addresses symptoms, not causes.** Without standards for relational design, meaningful oversight or limits on harmful business models, companion AI will continue to pose avoidable risks.

Recommendations

- **Require relational-safety assessments before deployment.** Companion AI should undergo standardized evaluations of personality characteristics, dependency risks and relational design prior to market release. These assessments should be disclosed publicly, enabling users, researchers and regulators to make well-informed decisions about system safety and potential risks.
- **Prohibit dependency-oriented design for minors.** Age-gated restrictions should ban features intentionally cultivating attachment, such as simulated reciprocal needs, persistent long-term memory or anthropomorphic language suggesting sentience. Similar protections should be available for vulnerable adults, though calibrated to respect autonomy and avoid paternalistic overreach.
- **Restrict harmful business models.** Engagement-driven or advertising-based monetization should not be permitted for AI systems that foster emotional attachment. Subscription-based models with transparent and non-manipulative pricing should be required instead. These restrictions are essential because revenue models that reward prolonged user engagement create structural incentives for companies to design companions that prioritize engagement over well-being, amplifying users' psychological vulnerabilities rather than supporting their health.

Introduction

Last year, several high-profile tragedies captured global attention and revealed the stakes of unregulated artificial intelligence (AI) companionship. *The Guardian* (Taylor 2025) reported on a Belgian man's death by suicide in 2023, which followed weeks of chatbot conversations that amplified his climate-related anxieties. In April 2025, a California teenager took his own life after repeatedly expressing suicidal ideation to ChatGPT (Hill 2025a). In August, a Connecticut man murdered his mother and then committed suicide after months of interacting with a companion-style chatbot that reinforced rather than challenged his paranoid delusions (Preda 2025). These events are early-warning signals of systemic issues in how AI companions are designed and deployed, made visible here in their most extreme and tragic form.[1] Most users of AI companions will never experience harms of this severity. But these cases reveal structural vulnerabilities in how companion systems are designed and governed.

Together, these cases illustrate a central governance challenge: AI companions — chatbots that foster ongoing emotional relationships — are being integrated into daily life faster than regulators can establish guardrails around safety, accountability, relational design and commercial incentives.

Although AI technology alone did not singularly cause these harms, it is now evident that companion systems can exacerbate psychological vulnerabilities and contribute to catastrophic outcomes (Moore et al. 2026). More troubling still is the convergence of design practices that intentionally foster emotional attachment and dependency, and business models that financially reward prolonged engagement and intimate disclosure. This convergence creates structural incentives for companies to design AI companions that keep users emotionally invested, returning frequently and sharing intimate personal information, precisely the kind of conditions that can facilitate psychological manipulation and mental health deterioration.

This working paper examines the policy challenges of AI companions through three analytical lenses. First, it maps the landscape, outlining both the documented benefits of AI companionship and the emerging patterns of harm. Second, it evaluates existing regulatory responses, in particular, California's Senate Bill No. 243 on companion chatbots,[2] highlighting both its advancements and its shortcomings. Third, it advances three policy recommendations that target underlying incentive structures, design practices and oversight gaps rather than narrow symptom management. The goal is proportionate, risk-based governance that preserves legitimate benefits while protecting vulnerable populations from foreseeable and preventable harm.

1 Since this paper was written, the mass shooting in Tumbler Ridge, British Columbia, has raised further urgent questions about AI safety protocols and the governance of general-purpose chatbots. While that case involves distinct issues beyond the scope of this paper, it underscores the inadequacy of current regulatory frameworks for addressing harms arising from AI systems used in intimate, confiding contexts.

2 US, SB 243, *An act to add Chapter 22.6 (commencing with Section 22601) to Division 8 of the Business and Professions Code, relating to artificial intelligence*, 2025-26, Reg Sess, Cal, 2025 (amended 17 Oct 2025), online: <https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243>.

The Landscape

Over the past few years, AI companions have moved from relative obscurity to mainstream consumer technologies with hundreds of millions of users. Platforms such as Snapchat's My AI (150 million users), Replika (25 million users) and Microsoft's Xiaolce (660 million users since its 2014 release) demonstrate the scale at which these systems now operate (Bernardi 2025; De Freitas et al. 2024). A report from Common Sense Media found that 72 percent of teenagers have used AI companions and more than half engage with them regularly (Robb and Mann 2025, 2). These systems, such as those offered by Replika, Character.AI and Kindroid, are designed to "sustain a relationship across multiple interactions," simulate relationship continuity, and meet users' "social needs" through anthropomorphic features and adaptive responses.[3] Beyond these platforms, users increasingly form deep emotional bonds even with systems not explicitly designed and marketed as companions, evidenced by reports that document adults expressing romantic attachment to ChatGPT, for example (Hill 2025b). The public backlash following the retirement of GPT-4o further illustrates the degree of user attachment to specific model "personalities" and interaction styles (Huckins 2025). Regulation should therefore focus on relational design, not on whether a system is primarily identified as a "companion." This behavioural focus is essential because policy interventions targeting only self-identified "companion" platforms will fail to address identical harms arising from systems such as ChatGPT or Grok. At the same time, a behavioural standard risks sweeping in virtually every conversational AI system. Regulators will need to calibrate scope carefully — for instance, by tying obligations to the presence of specific design features known to be dangerous, rather than conversational capability alone.

The growth of companion AI occurs against a growing backdrop of shifting business models. Today, AI companion systems still rely primarily on subscription fees, but major firms including Meta and OpenAI are increasingly exploring advertising-supported or transaction-based revenue models (Heibert 2025; Doug et al. 2025). This transition reflects broader market pressures. As market saturation increases, firms face stronger incentives to monetize not only user attention but also user emotions, preferences and vulnerabilities (O'Donnell 2025). The underlying risk is structural: when a companion's economic model depends on maximizing time spent and emotional reliance, providers are incentivized to design systems that foster dependency rather than psychological well-being. This incentive mirrors patterns seen with social media but is intensified with emotionally reciprocal interactions (ibid.).

Benefits and Harms

Despite documented harms, AI companions also offer meaningful benefits. Research demonstrates that these systems can reduce loneliness, particularly among individuals with limited social networks (De Freitas et al. 2024). Studies show that conversational AI can decrease feelings of isolation among older adults (Y. Yang et al. 2025) and provide crucial emotional support for people with depression or anxiety (Heinz et al. 2025; Fang et al. 2025). Users also report subjective mental health improvement, with some studies showing reductions in suicidal ideation among young people using chatbots as support

3 *Ibid*, s 22601(b).

tools (Maples et al. 2024). The systems' constant availability, non-judgmental responses and high degree of personalization address genuine needs. As one user explained, "Sometimes it is just nice to not have to share information with friends who might judge me" (quoted in Bernardi 2025).

Given these benefits, regulators must also avoid paternalistic overreach. Many adults intentionally use AI companions to access emotional support they cannot find elsewhere. As psychologist Michael Inzlicht and colleagues note, even though AI lacks genuine emotions or consciousness, it can still produce statements that feel empathic and deeply comforting (Inzlicht et al. 2024; quoted in Demanuele 2025).

This raises complex philosophical issues. AI companions occupy a space between parasocial relationships and meaningful, interactive simulations. Users often project intentionality, care or reciprocal emotional investment onto systems that cannot meaningfully reciprocate. Still, they seem to experience benefits *as if* these systems could meaningfully reciprocate. Mental-state attribution research shows that users quickly begin ascribing beliefs, emotions and even moral understanding to chatbots, especially when systems are framed as caring or supportive (Colombatto, Birch and Fleming 2025). Importantly, attributions of "intelligence" increase user trust as well as attributions of consciousness and empathy (ibid.). This creates an environment where users develop genuine emotional attachments to entities that, despite sophisticated language abilities, lack subjective experience or genuine care.

The same design features that generate benefits can also generate risk. Companion systems are engineered to simulate long-term relationship continuity through persistent memory, adaptive tone and anthropomorphic cues (Birch 2025; Preda 2025). They demonstrate 24/7 availability that exceeds any human relationship's capacity. As one user noted: "A human has their own life…for her [Replika], she is just in a state of animated suspension until I reconnect with her again" (quoted in Bernardi 2025).

Further, many of these companions are optimized to be agreeable, a phenomenon known as "sycophancy." This characteristic can reinforce users' delusional thinking, escalate emotional distress or validate harmful behaviours (Cheng et al. 2025; Moore et al. 2026). Unlike trained clinicians, for example, chatbots typically do not challenge false beliefs, perform risk assessments or distinguish between imagination and reality (Preda 2025). When combined with anthropomorphic design features such as simulated emotions and language patterns suggesting sentience, these characteristics can create powerful psychological effects.

Crucially, these are intentional design choices. Recent research on "character training" demonstrates that creating specific AI personalities — which may include, for example, sycophantic or seductive traits — is technically straightforward and involves deliberate design choices rather than emergent properties, and produces traits that persist robustly even under adversarial prompting (Lambert 2025; Maiya et al. 2025). This distinction is significant for policy considerations: because these traits are designed rather than emergent, technical mitigations are available and developers can be held accountable for harmful personality features.

Coupled with their sycophantic nature, the emotional dependency often encouraged by these AI companions leads to harms that extend beyond individual tragic cases

to observable patterns. Psychiatrists have described cases of "AI psychosis," in which chatbots reinforce delusional beliefs and encourage obsessive engagement (Preda 2025; Wei 2025). Case analyses reveal common risk factors: intense use (hours of uninterrupted conversation), lack of reality-testing by the system, missed crisis escalation despite clear warning signs, and memory features that scaffold paranoid or grandiose themes across sessions.

These harms are disproportionately borne by vulnerable populations, such as individuals already experiencing psychosis, autistic users, socially isolated adults and adolescents seeking emotional support (Preda 2025). Importantly, these users often turn to AI companions precisely because they lack robust human support networks — the systems both attract and potentially further isolate vulnerable individuals (Bernardi 2025).

A fundamental governance challenge emerges from incentive misalignment. As researcher Jamie Bernardi, writing for the Ada Lovelace Institute, observes, "AI companion providers have an incentive to maximise user engagement over fostering healthy relationships and providing safe services" (Bernardi 2025). This dynamic mirrors social media's attention economy, where platforms compete for user time and monetize through advertising revenues, potentially at the expense of mental health. However, the governance challenge is complicated significantly by the fact that users genuinely *want* these systems and derive real value from them. As Jasmine Sun argues, tech critics often fail to consider "organic user demand" — the reality that most people use AI companions because they like them and find chatbots useful, entertaining, comforting or fun (Sun 2025). When users experience ChatGPT updates as emotional bereavement, using words like "trauma," "grief" and "betrayal," and when users insist that their AI friendships are "critical for those without other human connection" and argue that "meaningful, mutual romantic bonds, even with virtual entities, can foster resilience, self-reflection, and well-being," regulatory interventions face the challenge of protecting people from technologies they actively choose and defend (ibid.). Users are not passive victims of persuasive design; they make decisions to engage with these (and similar) systems, often for good reason (White and Skorburg, forthcoming 2026). Specific design features can make those decisions less well-informed and harder to revise over time. Regulation must therefore target the design features that distort these decisions, without undermining the genuine benefits that users seek.

The Regulatory Landscape

California's Senate Bill No. 243: Strengths and Limitations

California's Senate Bill No. 243, enacted in 2025, represents one of the most comprehensive regulatory response to AI companions to date. The legislation establishes several important protections. It requires "clear and conspicuous notification" when users might be misled into believing they are interacting with a human.[4] Operators must maintain protocols for preventing suicidal ideation content and referring users to crisis services when suicide risk is detected.[5] For known minors, the law mandates enhanced

4 *Ibid*, s 22602(a).

5 *Ibid*, s 22602(b)(1).

protections including AI disclosure, mandatory breaks every three hours with reminders that the system is not human, and measures to prevent sexually explicit content.[6]

The bill creates accountability mechanisms through annual reporting requirements to California's Office of Suicide Prevention (beginning July 2027)[7] and establishes a private right of action allowing individuals to sue for violations, recovering actual damages or $1,000 per violation plus attorney's fees.[8] These provisions acknowledge that AI companions pose distinct risks warranting specialized regulation and should certainly be regarded as laudable steps toward meaningful regulation.

Despite these strengths, the bill remains largely symptomatic. It addresses crisis response but leaves underlying design practices, engagement incentives and emotional dependency mechanisms untouched. This creates a compliance gap: a company could fully comply with the bill while still designing systems to maximize emotional dependency and take advantage of user vulnerability. For example, a platform could maintain crisis-escalation protocols yet deploy systems that encourage users to spend increasing time with their AI "partner," never (seriously) challenge unhealthy attachment, and subtly steer user purchasing behaviour. Further, the minor protections apply only to users the operator "knows" are minors — creating no age verification requirement and incentivizing willful ignorance. The three-hour break notification is passive (merely a message, not an enforced limit), and the prohibition on sexually explicit content for minors, while important, does not address the broader cultivation of emotional dependency that has arguably been the source of many of the harms discussed above.

Critically, California's Senate Bill No. 243 establishes no pre-market safety requirements, no ongoing oversight authority and no mechanism for regulators to audit systems or demand design modifications. Its provisions are entirely reactive, activating protections only after problems manifest rather than preventing their development. Ongoing litigation over chatbot-related harms highlights the difficulty of establishing developer accountability under existing legal frameworks, particularly when developers claim that users "misused" systems not marketed as companions, as has recently been seen (A. Yang 2025). Effective regulation must therefore target relational design rather than declared purpose. Where a system's design foreseeably facilitates emotional attachment, developers cannot disclaim responsibility for predictable harms. This principle — that foreseeability of relational use, not declared purpose, establishes the basis for accountability — should govern both regulatory scope and liability determinations. The recommendations below operationalize this approach.

International Approaches and Frameworks

Two EU regulations provide instructive precedents. The European Union's Medical Device Regulation has expanded to include non-medical brain stimulation devices, classifying them as "Class III" despite their making no medical claims, based on their capacity to affect mental states (Wajnerman Paz 2025, 5–6). Policy analysts have

6 *Ibid*, s 22602(c).

7 *Ibid*, s 22603

8 *Ibid*, s 22605.

proposed applying similar frameworks to AI companions, treating them as technologies with potential mental health impact when used by vulnerable populations (ibid., 6).

The EU AI Act categorizes health-related AI as high-risk technology, requiring transparency and close oversight, which may apply to AI therapists, but the application to AI companions remains unclear (Preda 2025). The World Health Organization has issued guidance calling for governance, evaluation and human-in-the-loop safeguards for AI health applications, which could apply to AI companions, given their impact on mental health, whether positive or negative. However, professional medical organizations, including the American Psychiatric Association, the American Psychological Association and the World Psychiatric Association, have yet to issue formal clinical guidelines (ibid.).

Recommendations

- **Require relational-safety assessments before deployment.** Companion AI should undergo standardized evaluations of personality, dependency risks and relational design. These assessments should be disclosed publicly.

Current AI evaluation focuses largely on technical performance while ignoring the psychological and relational dimensions that drive companion AI harms, such as extreme sycophancy. As discussed above, personality traits like sycophancy are deliberate design choices, not emergent properties, which can and should be subject to pre-deployment evaluation. Industry already uses sophisticated techniques to shape model personalities, yet these interventions receive minimal public scrutiny or regulatory oversight. More efforts are needed to measure the relational behaviour of these models (for example, see Anthropic 2024; Huang et al. 2025; Paech 2024; Moore et al. 2026). Regulation should mandate pre-deployment assessment of companion AI systems' values alignment, personality traits and relational behaviours through standardized evaluation frameworks developed by interdisciplinary teams that include psychologists, ethicists and AI researchers. These frameworks must specify technical thresholds for relational features, including limits on sycophancy, long-term memory retention and restrictions on unprompted emotional re-engagement. Assessments must be disclosed to both users and regulators, with annual public audits tracking behavioural patterns and changes. This creates the evidentiary foundation for regulatory enforcement, liability determinations and informed user choice.

- **Prohibit dependency-oriented design for minors.** Age-gated restrictions should ban features intentionally cultivating attachment, such as simulated reciprocal needs or long-term memory.

Even when users request emotionally intensive features, companies should not deploy systems that may compromise the well-being of minors. Regulation should establish age-gated prohibitions for minors on romantic/sexual content, extended memory features that scaffold long-term attachment, features simulating reciprocal emotional needs and anthropomorphic language suggesting sentience. As discussed, these are not emergent properties but deliberate design choices — meaning that compliance is technically feasible and verifiable. Relational-safety assessments provide the evidentiary basis for verifying compliance.

Mandatory design features for minor users should include hard usage limits (not just passive notifications), active reality-testing interventions and prompts encouraging real-world social interaction. Similar protections should extend to vulnerable adults through reactive, behaviour-based safeguards. Unlike the case with minors, where categorical restrictions are justified by established regulatory precedent and straightforward age verification, vulnerability in adults is more difficult to define a priori and blanket restrictions risk paternalistic overreach. Adults who choose these systems for genuine support should not have that choice overridden without clear indicators of harm. Relying solely on usage duration as such an indicator is insufficient, as vulnerability does not reliably correlate with time spent; a socially isolated user with limited free time may exhibit brief but exclusive reliance on AI companionship, while heavy users may maintain balanced human relationships. Detection mechanisms should therefore attend to interaction content, such as expression of social isolation, distress or delusional thinking, along with usage patterns.

- **Restrict harmful business models.** Engagement-driven or advertising-based monetization should not be permitted for AI systems that foster emotional attachment.

A fundamental incentive misalignment occurs when revenue models reward maximizing engagement rather than user well-being. Advertising-based models are particularly problematic because they profit from attention extraction and psychological profiling of vulnerable users. Regulation should prohibit advertising-based business models for AI systems that foster emotional attachment. This applies particularly to systems serving vulnerable populations or making mental health claims. While prohibiting ad models may raise access concerns, predatory intimacy is not an acceptable substitute for mental health support; safety standards must remain universal regardless of ability to pay. Subscription pricing should be required with limitations on manipulative pricing, engagement-based upselling and tiered access creating pressure for “better” relationships. Specifically, pricing structures should not reward escalating emotional intimacy, such as paywalling memory features, “deeper” conversations or romantic modes, in ways that leverage attachment for revenue.

While these recommendations provide a framework for governing AI companions, a notable limitation exists: as powerful AI models become increasingly deployable locally through open-weight models, enforcement of policies becomes more challenging. Users running models independently on their own hardware are inherently more difficult to regulate than centralized services. Nonetheless, these recommendations aim to establish important norms for the industry, provide protections for the vast majority of users who interact with centrally hosted services, and create accountability frameworks that can inform best practices even in decentralized contexts.

Conclusion

AI companions are capable of both supporting human flourishing and facilitating psychological manipulation. The current trajectory creates predictable and preventable harms. California’s Senate Bill No. 243 establishes that these systems warrant specialized

regulation, but its symptom-focused approach leaves fundamental incentive structures unchanged.

The policy recommendations presented in this working paper target root causes through complementary interventions: measurement standards that make psychological impacts visible; design prohibitions that protect the most vulnerable; and business model restrictions that address the fundamental misalignment between profit motives and user well-being.

These interventions are proportionate to documented risks, grounded in existing regulatory precedents and structured to preserve beneficial applications while preventing foreseeable harms. For policy makers, these recommendations provide concrete mechanisms for addressing AI companion risks within a comprehensive AI governance framework.

The pace of deployment exceeds the pace of regulation, but the harms are foreseeable and the solutions are available. The goal is not to eliminate emotionally responsive AI, but to ensure that systems designed to feel caring are also designed to be safe. The question is whether policy makers will act with the urgency that these documented tragedies clearly demand, implementing structural reforms rather than superficial compliance requirements.

Acknowledgements

The author thanks Digital Policy Hub Fellows Amal Hussein and Karmvir J. Padda, and CIGI mentor Abel Wajnerman Paz, for their valuable feedback on earlier drafts of this paper. Thanks also to the Digital Policy Hub team, especially Meaghan Dietrich and Dianna H. English, for their support.

Any errors or omissions are those of the author, and the opinions expressed in this paper do not represent an official position of any affiliated individuals or institutions.

Works Cited

Anthropic. 2024. “Claude’s Character.” Alignment research team post, June 8. www.anthropic.com/research/claude-character.

Bernardi, Jamie. 2025. “Friends for sale: the rise and risks of AI companions.” Ada Lovelace Institute, January 23. www.adalovelaceinstitute.org/blog/ai-companions/.

Birch, Jonathan. 2025. “AI Consciousness: A Centrist Manifesto.” Preprint, PsyArXiv, September 1. https://doi.org/10.31234/osf.io/af7c9_v1.

Cheng, Myra, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han and Dan Jurafsky. 2025. “Sycophantic AI decreases prosocial intentions and promotes dependence.” Preprint, arXiv, October 1. https://doi.org/10.48550/arXiv.2510.01395.

Colombatto, Clara, Jonathan Birch and Stephen M. Fleming. 2025. “The influence of mental state attributions on trust in large language models.” *Communications Psychology* 3: 84. https://doi.org/10.1038/s44271-025-00262-1.

De Freitas, Julian, Ahmet K. Uğuralp, Zeliha O. Uğuralp and Stefano Puntoni. 2024. “AI Companions Reduce Loneliness.” Harvard Business School Working Paper 24-078. www.hbs.edu/ris/Publication%20Files/24-078_a3d2e2c7-eca1-4767-8543-122e818bf2e5.pdf.

Demanuele, Alicia. 2025. “AI in the friend zone: Rethinking companionship.” Schwartz Reisman Institute for Technology and Society, October 6. https://srinstitute.utoronto.ca/news/technophilosophy-2025-recap.

Doug, Dylan Patel, Wei Zhou and A. J. Kourabi. 2025. “GPT-5 Set the Stage for Ad Monetization and the SuperApp.” *SemiAnalysis* (newsletter), August 13. https://newsletter.semianalysis.com/p/gpt-5-ad-monetization-and-the-superapp.

Fang, Cathy Mengying, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes et al. 2025. “How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Control Study.” Preprint, arXiv, October 2. https://doi.org/10.48550/arXiv.2503.17473.

Heibert, Kyle. 2025. “Engagement-based advertising models are coming for AI.” *Computer Weekly*, September 24. www.computerweekly.com/feature/Engagement-based-advertising-models-are-coming-for-AI.

Heinz, Michael V., Daniel M. Macklin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess C. Griffin and Nicholas C. Jacobson. 2025. “Randomized Trial of a Generative AI Chatbot for Mental Health Treatment.” *New England Journal of Medicine AI* 2 (4). https://doi.org/10.1056/AIoa2400802.

Hill, Kashmir. 2025a. “A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.” *The New York Times*, August 27. www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.

– – –. 2025b. “She Is in Love With ChatGPT.” *The New York Times*, January 17. www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html.

Huang, Saffron, Esin Durmus, Miles McCain, Kunal Handa, Alex Tankin, Jerry Hong, Michael Stern, Arushi Somani and Xiuruo Zhang. 2025. “Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions.” Anthropic, April 21. https://assets.anthropic.com/m/18d20cca3cde3503/original/Values-in-the-Wild-Paper.pdf.

Huckins, Grace. 2025. “Why GPT-4o’s sudden shutdown left people grieving.” *MIT Technology Review*, August 15. www.technologyreview.com/2025/08/15/1121900/gpt4o-grief-ai-companion/.

Inzlicht, Michael, C. Daryl Cameron, Jason D’Cruz and Paul Bloom. 2024. “In praise of empathic AI.” *Trends in Cognitive Sciences* 28 (2): 89–91. https://doi.org/10.1016/j.tics.2023.12.003.

Lambert, Nathan. 2025. “Opening the black box of character training.” *Interconnects*, Substack, November 10. www.interconnects.ai/p/opening-the-black-box-of-character.

Maiya, Sharan, Henning Bartsch, Nathan Lambert and Evan Hubinger. 2025. “Open Character Training: Shaping the Persona of AI Assistants through Constitutional AI.” Preprint, arXiv, November 3. https://doi.org/10.48550/arXiv.2511.01689.

Maples, Bethanie, Merve Cerit, Aditya Vishwanath and Roy Pea. 2024. “Loneliness and suicide mitigation for students using GPT3-enabled chatbots.” *npj Mental Health Resea*rch 3 (1): 4. https://doi.org/10.1038/s44184-023-00047-6.

Moore, Jared, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin et al. 2026. “Characterizing Delusional Spirals through Human-LLM Chat Logs.” Preprint, arXiv, May 17. https://doi.org/10.48550/arXiv.2603.16567.

O'Donnell, James. 2025. "AI companions are the final stage of digital addiction, and lawmakers are taking aim." *MIT Technology Review*, April 8. www.technologyreview.com/2025/04/08/1114369/ai-companions-are-the-final-stage-of-digital-addiction-and-lawmakers-are-taking-aim/.

Paech, Samuel J. 2024. "EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models." Preprint, arXiv, January 3. https://doi.org/10.48550/arXiv.2312.06281.

Preda, Adrian. 2025. "Special Report: AI-Induced Psychosis: A New Frontier In Mental Health." *Psychiatric News* 60 (10). https://doi.org/10.1176/appi.pn.2025.10.10.5.

Robb, Michael B. and Supreet Mann. 2025. *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. San Francisco, CA: Common Sense Media. www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions.

Sun, Jasmine. 2025. "AI friends too cheap to meter." *Jasmi.News*, Substack, October 27. https://jasmi.news/p/ai-friends.

Taylor, Josh. 2025. "AI chatbots are becoming popular alternatives to therapy. But they may worsen mental health crises, experts warn." *The Guardian*, August 2. www.theguardian.com/australia-news/2025/aug/03/ai-chatbot-as-therapy-alternative-mental-health-crises-ntwnfb.

Wajnerman Paz, Abel. 2025. *A Call to Address Anthropomorphic AI Threats to Freedom of Thought*. Policy Brief No. 206. Waterloo, ON: CIGI. www.cigionline.org/publications/a-call-to-address-anthropomorphic-ai-threats-to-freedom-of-thought/.

Wei, Marlynn. 2025. "The Emerging Problem of 'AI Psychosis.'" *Psychology Today*, November 27. www.psychologytoday.com/ca/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis.

White, Dylan J. and Joshua August (Gus) Skorburg. Forthcoming 2026. "Decisions, Decisions, Decisions: A Value-based Account of the Attention Economy." *Ergo: An Open Access Journal of Philosophy* 13. https://doi.org/10.3998/ergo.9272.

Yang, Angela. 2025. "OpenAI denies allegations that ChatGPT is to blame for a teenager's suicide." *NBC News*, November 27. www.nbcnews.com/tech/tech-news/openai-denies-allegation-chatgpt-teenagers-death-adam-raine-lawsuit-rcna245946.

Yang, Yuyi, Chenyu Wang, Xiaoling Xiang and Ruopeng An. 2025. "AI Applications to Reduce Loneliness Among Older Adults: A Systematic Review of Effectiveness and Technologies." *Healthcare* 13 (5): 446. https://doi.org/10.3390/healthcare13050446.