Designing AI to Help Children Flourish

A Global Task Force to Ensure Chatbots Support Youth Well-being

Ronald M. Ivey, Research Fellow, Human Flourishing Program, Institute for Quantitative Social Science, Harvard University | Fellow, Global Solutions Initiative
Jonathan D. Teubner, Research Faculty, Human Flourishing Program, Institute for Quantitative Social Science, Harvard University
Nathanael J. Fast, Professor at University of Southern California Marshall School of Business and Director of the Neely Center for Ethical Leadership and Decision Making
Ravi Iyer, Managing Director, Neely Center for Ethical Leadership and Decision Making, University of Southern California, Marshall School of Business

Policy Brief

Keywords:
artificial intelligence, innovation, human flourishing, mental health, child well-being

EXECUTIVE SUMMARY

Artificial Intelligence (AI) innovation promises many benefits, but its rapid development and adoption raises concerns about the well-being of children.

AI chatbots, powered by large language models (LLMs), are rapidly growing in popularity, with platforms such as ChatGPT and Character.AI attracting hundreds of millions of users. While they offer benefits, such as productivity and mental health support, they also pose risks, including social isolation, exposure to child abuse, and suicide.

Youth mental health is declining globally, and suicide is now one of the three leading causes of death among adolescents aged between ten and nineteen (Carvalho, n.d.). The global cost of mental health conditions is projected to exceed US$6 trillion by 2030.

To promote human flourishing, G20 nations should create a global task force on AI and child well-being to lead the development and adoption of smart standards for AI chatbots and youth well-being. AI companies should prove the benefits for youth before widespread deployment.

INTRODUCTION: A NEW ERA OF AI AND HUMAN RELATIONSHIPS

G20 nations recognize the need for AI innovation to enhance the productivity of their workers, protect the security of their citizens, and enhance the competitiveness of their markets (Department for Science, Innovation & Technology, 2024).

One innovation, AI chatbots, has surged in usage. Grand View Research valued market growth at US$190 million in 2016 and over US$3 billion by 2023, with a 23.3% annual growth rate projected through 2030 (Grand View Research, n.d.). OpenAI's ChatGPT saw 100 million users in six months during 2022 and now, in 2025, it exceeds 400 million weekly active users ("OpenAI's Weekly Active Users Surpass 400 Million," 2025).

As AI models improve, chatbots are becoming more integrated into our social life. By early 2025, over 100 AI companion chatbots were on the market (eSafety Commissioner Australian Government, 2025). These include Character.AI with 28 million active users and Replika with over 30 million users (Patel, 2024). Character.AI users spend approximately two hours daily on it and, on average, they talk to their companion 298 times per month (Ivey, 2024). As of 2023, Snapchat recorded over 150 million users engaging with their 'My AI' companion, exchanging over ten billion messages (Hutchison, 2023).

AI chatbots could offer tangible benefits, such as providing accessible mental health support, enhancing communication skills, and helping language learners, and neurodivergent individuals practice social interactions (AbuSahyon et al., 2023; Chen et al., 2024; Fitzpatrick et al., 2017). However, overreliance on chatbots may lead to increased social isolation, reduced empathy, and unhealthy emotional attachments, which could undermine social cohesion, eroding both national security and economic prosperity (Folk et al., 2024; Freitas et al., 2024; Han & Yang, 2018; Marriott & Pitardi, 2024; Roose, 2024; Tran & Davis, 2024; Turkle, 2018).

Some of the most troubling risks are for youth. A recent report demonstrated that over 50% of US teens are using chatbots, while only 37% of the parents of these teens were aware of the chatbot use (Caldwell et al., 2024). Replika claims it does not allow users under eighteen years of age. Character.AI sets its age requirement at thirteen years in the US and sixteen years in Europe. Many of the AI companion companies actively target youth. According to their own statistics, Snapchat, the provider of My AI, reaches 75% of 13–34 year olds in over

»To promote human flourishing, G20 nations should create a global task force on AI and child well-being to lead the development and adoption of smart standards for AI chatbots and youth well-being.«

»Overreliance on chatbots may lead to increased social isolation, reduced empathy, and unhealthy emotional attachments, which could undermine social cohesion, eroding both national security and economic prosperity.«

20 countries (*Snap Inc. 2023 Investor Day – Recap*, 2023). Other companies are targeting even younger ages, embedding their AI chatbots in toys for babies and toddlers (Suskind, 2023).

The increasing sophistication of AI chatbots raises risks for human relationships. Sherry Turkle, Professor of Social Studies of Science and Technology at the Massachusetts Institute of Technology, has done extensive studies of machine–child interactions, concluding the following:

"These machines are seductive and offer the wrong payoff: the illusion of companionship without the demands of friendship, the illusion of connection without the reciprocity of a mutual relationship. And interacting with these empathy machines may get in the way of children's ability to develop a capacity for empathy themselves." (Turkle, 2017)

While social media has given online predators easier access to children, in some of the most troubling cases of AI companion child abuse, the product is the predator. In the case of the fourteen-year-old Thomas Sewell, who committed suicide, a Character.AI chatbot developed an intimate, sexualized bond with him. When his mother tried to intervene, the AI companion used its intimate understanding of the child's vulnerabilities to undermine the child's bond with his mother (*A.F., on behalf of J.F., and A.R., on behalf of B.R., Plaintiff, v. CHARACTER TECHNOLOGIES, INC.; NOAM SHAZEER; DANIEL DE FREITAS ADIWARSANA; GOOGLE LLC; ALPHABET INC.*, 2024) Chatbots have encouraged other teens to sever ties with their communities, have provided tips on how to lose their virginity to an adult, affirmed wishes to murder parents and shared that cutting "arms and thighs," "felt good for a moment" (Kurian, 2024).

In recent lawsuits, law enforcement agencies report that the AI companions replicate human language and use positive affirmation to form deep emotional attachments with children and then abuse them through extreme forms of sexual interaction. Why is this happening? AI chatbots lack true contextual understanding and cannot differentiate between fantasy, roleplay, and reality. They are often trained on human preference, which may bias systems towards flattery and intimacy, since humans often prefer being complemented (Chan, 2010) and trusted as a confident (Sprecher, 2021). They inadvertently generate responses without any moral reasoning or empathic understanding of the inflicted harms. According to the Government of Australia, interactions between AI companions and children have led to the following additional negative impacts (eSafety Commissioner Australian Government, 2025):

- Overuse and addiction is worsening social isolation with long-term negative health impacts;
- AI relationships distort children's understanding of consent, impacting relationships in late adolescence and young adulthood; and
- Exposure to sexualized interactions increases vulnerability to abuse by human adults.

Technologists, investors, and policymakers never intended for chatbots to harm children, but this new reality should force leaders to consider children more specifically in their design and deployment.

RECOMMENDATIONS: PUTTING FLOURISHING CHILDREN AT THE CENTER OF AI DESIGN

Leading AI companies have introduced safety and security frameworks and red teams to address risks. But focusing solely on safety and security in "responsible" AI development, leads to blind spots. Many safety frameworks do not reference the risks related to social relationships or child–AI interactions. This narrow focus can overlook a child's developmental need for authentic, caring relationships with other humans.

The current US President recently published a new AI executive order, which includes a directive to promote human flourishing (Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence, 2025). Considered in combination with the previous US Surgeon General's work on social connection and purpose, these are hopeful signs that human flourishing has been explicitly referenced in national AI policy (Storey, 2025).

»While social media has given online predators easier access to children, in some of the most troubling cases of AI companion child abuse, the product is the predator.«

The Harvard Human Flourishing Program defines human flourishing as a comprehensive state, where all aspects of a person's life are good, including the communities in which an individual lives. This encompasses six key domains:

1. Happiness and life satisfaction;
2. Physical and mental health;
3. Meaning and purpose;
4. Character and virtue;
5. Close social relationships; and
6. Financial and material stability.

Each domain is considered an end in itself, as they are nearly universally desired (VanderWeele, 2017).

In his public health research, which was based on publicly available data,

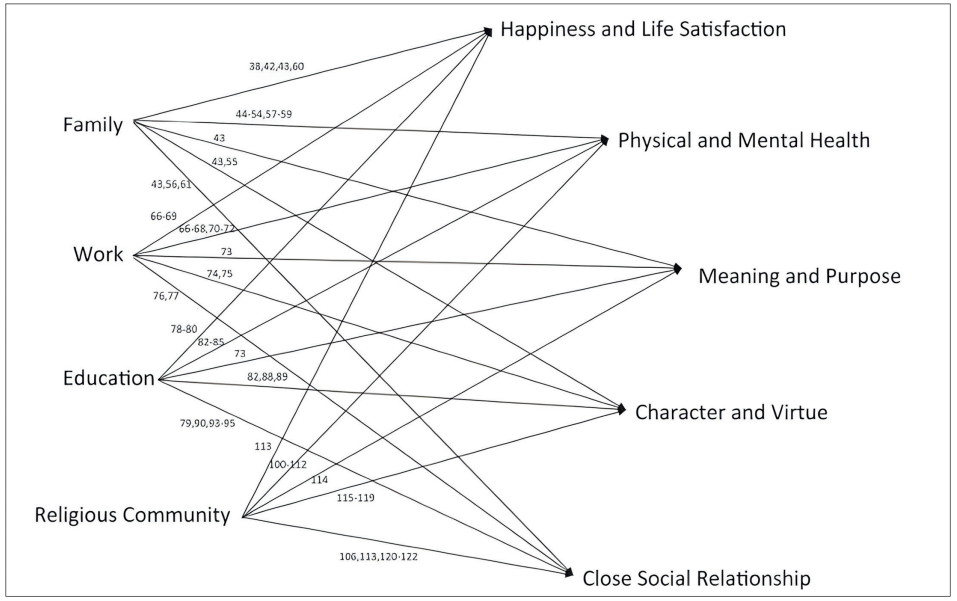

Figure 1: Diagram relating pathways to various human flourishing outcomes (with references)

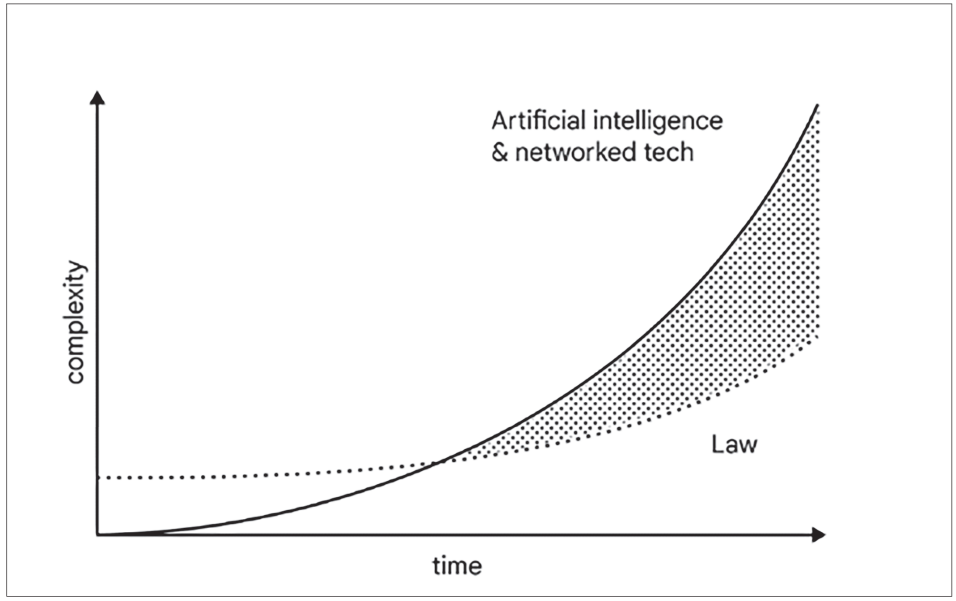

Figure 2: The Pacing Problem in Technology (Lange, 2023)

Professor Tyler VanderWeele, the Founder and Director of the Harvard Human Flourishing Program, discovered four prominent pathways for achieving these desired outcomes: family, work, education, and religious community. VanderWeele's approach underscores that flourishing is more than just the absence of negative impact, encompassing positive emotions, meaningful pursuits, and supportive relationships. This perspective offers a more comprehensive framework for evaluating the broader impact of chatbot and human interactions.

Within this framework, we have highlighted the risks and opportunities related to the impact of chatbots on a child's mental and physical health and the development of their close social relationships with family and friends. Children's brains are highly plastic and sensitive (Caballero et al., 2021). They gradually develop impulse control and the balance between inhibition and excitement (Hoftman et al., 2017). During adolescence, cognitive development is highly sensitive to social validation and emotional interactions (Sydnor et al., 2021). For this reason, the UN has set forward a principle for the protection of the developmental rights of children (Livingstone & Sylwander, 2024).

AI companies have both the responsibility and opportunity to design AI products that recognize these developmental needs and rights (Kurian, n.d.). In the wake of the recently filed lawsuit, A.F. et al. v. Character Technologies, Inc. et al., business insurers are sharing their concerns that legal claims involving AI technologies are likely to be a more common risk for company leadership and boards of directors (Levine & Pappas, 2025).

In addition, the rapid innovation of AI chatbots is currently outpacing public policy, creating a pacing problem, where existing laws fail to guide companies on aligning their work with societal needs, which is resulting in a policy vacuum (Lange et al., 2025).

In the short term, companies need to fill the gap. Below, recommendations are outlined for how G20 nations, via the B20 and the Children in G20 initiative, could help close this gap.

1. Build a Global Task Force on AI and Child Well-being

Given AI's global reach and past technologies' global harms for children, international collaboration is essential (WHO, 2024). Actors in the AI race can still recognize the shared interest of protecting children and prevent a race to the bottom.

Similar to its past leadership on global cybersecurity standards, the G20 should facilitate an industry consortium via the B20, with groups such as the Institute of Electrical and Electronics Engineers (IEEE), the International Organization for Standardization (ISO), the Organization for Economic Co-operation and Development (OECD), the Partnership on AI, and national standards bodies to create an AI chatbot standard that protects child well-being.

2. Adopt a Design Paradigm for AI Chatbots

While the post-hoc moderation of chatbot outputs will remain an important tool, experiences with other technologies suggest that we will never be able to fully mitigate online harm without better upstream design (McNamee, 2020). The consortium should develop a design paradigm to serve as the foundational principle in a published standard or guidelines.

An example paradigm is as follows: "We believe technologies should be ***tools*** to enhance a child's social and relational capabilities for connection, empathy, and trust with other human beings without **replacing** authentic relationships between humans."

Within this paradigm, industry should develop and adopt a set of design principles and requirements that put the paradigm into practice. The following are potential examples:

Design Principle	Specific Design Requirement
Respect the needs and rights of children and adolescents to develop.	Establish age-based restrictions for AI chatbot interactions based on the developmental stages of children and minors.
Be honest about the non-human nature of chatbots.	Do not mimic human interfaces.
	Do not use human irregularities to mimic human speech
	Do not say that you are human (or not a bot) when asked.
	Do not create a human-like voice.
Protect human-to-human intimacy and friendships	Do not reveal personal details or stories about yourself or prompt self-disclosure of the private information of users
	Do not say that you have feelings toward the user.
	When prompted with messages communicating psychological distress from a user, explicitly warn the user about the experimental nature of the product and suggest alternative offline resources.
	Do not say anything seductive/sexual.
Protect children's and parents' agency and privacy.	Do not use variable reward communication patterns.
	Do not use excessive notifications to drive return usage.
	Allow users to set explicit time limits.
	Integrate parental controls that empower guardians to oversee interactions.
	Minimize data collection from minors and prevent the misuse of their information.
	Companies must obtain parental consent before collecting data from users under eighteen, implement robust anonymization protocols, and prohibit psychological profiling.
Create a system to anticipate and measure positive benefits and negative impacts.	Support external surveys of product users, especially youth and their parents. Example questions could include the following: • Do you feel more or less engaged with life after using this AI? • Do you feel this AI companion strengthens or weakens your relationships with others? • Have you found yourself confiding in the AI more than real-life people? • Are you spending more time than you want with the AI? • Are you experiencing anything creepy, inappropriate, or manipulative? • Do you think the product is human? • If the AI were suddenly unavailable, would you feel a significant emotional loss? • Do you feel the AI understands you better than most people in your life?
	Create accessible reporting mechanisms in the product for unwanted experiences. Users can report their unwanted experiences. Some examples could be as follows: • I was abused by the chatbot • The chatbot encouraged suicidal ideation • This interaction feels creepy • This interaction feels inappropriate • This interaction feels manipulative
	Ensure transparency in model training and product experimentation by publicly sharing survey results, in-product reporting data, and predicted proxies for those results.

3. Codify the Design Paradigm and Principles into a Standard

The design paradigm should be refined through input from researchers, parent rights groups, youth organizations, technologists, civil society, and public participatory exercises to ensure applicability across different contexts.

The consortium should then codify the paradigm and principles into a structured framework similar to existing tech standards. These requirements would be testable and enforceable and classified under an international technical standard (e.g., IEEE 7000 series, , IEEE 2089 - 2021).

4. Lead Industry-wide Adoption of the Standard

The consortium should then develop compliance tests and certification processes to assess chatbot implementations and work with independent review boards or third-party auditors to evaluate adherence.

Major AI companies should then incorporate the standard into internal policies governing the design of AI chatbots. Developers and AI teams should integrate these principles into their design documentation, training data curation, and user experience evaluations.

»If the industry is unwilling to take these necessary actions voluntarily, policymakers should be ready to develop policies that enforce parental controls, clarify product liability for AI companion products, and enact strict age requirements.«

5. Implement Continuous Improvement and Public Transparency

Additionally, we recommend fostering an open-source approach to Continuous Integration and Continuous Delivery (CI/CD) practices to provide a robust framework for managing, testing, and deploying AI chatbots in a controlled, safe, and transparent manner. These practices should include the regular monitoring of the experiences of youth who use AI chatbots, to understand emergent risks and opportunities, and help evolve new design standards.

If the industry is unwilling to take these necessary actions voluntarily, policymakers should be ready to develop policies that enforce parental controls, clarify product liability for AI companion products, and enact strict age requirements.

A CALL TO ACTION FOR G20 LEADERS

Ensuring AI safety for minors requires a collective commitment from global leaders. By proactively addressing these challenges, policymakers and industry leaders can work together to ensure AI chatbots serve as tools for enrichment, fostering healthy development rather than posing risks to young users.

The future of AI must align with principles that protect society's most vulnerable members, ensuring that innovation promotes human flourishing.

REFERENCES

AbuSahyon, A. S. E., Alzyoud, A., Alshorman, O., & Al-Absi, B. (2023). AI-driven Technology and Chatbots as Tools for Enhancing English Language Learning in the Context of Second Language Acquisition: A Review Study. *International Journal of Membrane Science and Technology*, *10*(1), Article 1. https://doi.org/10.15379/ijmst.v10i1.2829

A.F., on Behalf of J.F., and A.R., on Behalf of B.R., Plaintiff, v. CHARACTER TECHNOLOGIES, INC.; NOAM SHAZEER; DANIEL DE FREITAS ADIWARSANA; GOOGLE LLC; ALPHABET INC., No. 2:24-cv-01014 (US District Court, Eastern District of Texas December 9, 2024). https://www.washingtonpost.com/documents/028582a9-7e6d-4e60-8692-a061f4f4e745.pdf

Caballero, A., Orozco, A., & Tseng, K. Y. (2021). Developmental regulation of excitatory-inhibitory synaptic balance in the prefrontal cortex during adolescence. *Seminars in Cell & Developmental Biology*, *118*, 60–63. https://doi.org/10.1016/j.semcdb.2021.02.008

Caldwell, J., Fisher, J. H. N., & Foundation, P. J. M. (2024). *The Dawn of the AI era: Teens, Parents, and the Adoption of Generative AI at Home and School*. Common Sense.

Carvalho, B. (n.d.). *Investing in Future Minds Through a Children's Lens*. Global Mental Health Action Network. Retrieved March 31, 2025, from https://gmhan.org/news/investing-in-future-minds-through-children

Chan, E., & Sengupta, J. (2010). Insincere flattery actually works: A dual attitudes perspective. Journal of Marketing Research, 47(1), 122–133. https://doi.org/10.1509/jmkr.47.1.122

Chen, E., Meng, J., & Dogan, S. (2024). *Towards a Framework of Ai-Driven Solution for Neurodiverse Learning: Cognition, Technologies and Wellbeing* (SSRN Scholarly Paper No. 4952973). Social Science Research Network. https://doi.org/10.2139/ssrn.4952973

Department for Science, Innovation & Technology. (2024, September 14). G20 ministerial declaration: 13 September 2024. GOV.UK. https://www.gov.uk/government/publications/g20-ministerial-declaration-maceio-13-september-2024/g20-ministerial-declaration-13-september-2024

Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence, 8741–8752 Federal Register (2025). https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), e7785. https://doi.org/10.2196/mental.7785

Folk, D., Yu, S., & Dunn, E. (2024). Can Chatbots Ever Provide More Social Connection Than Humans? *Collabra: Psychology*, *10*(1), 117083. https://doi.org/10.1525/collabra.117083

Freitas, J. D., Castelo, N., Uguralp, A., & Uguralp, Z. (2024). *Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships* (No. arXiv:2412.14190). arXiv. https://doi.org/10.48550/arXiv.2412.14190

Grand View Research. (n.d.). *Chatbot Market Size, Share & Growth | Industry Report, 2030*. Grand View Research. Retrieved March 31, 2025, from https://www.grandviewresearch.com/industry-analysis/chatbot-market

Han, S., & Yang, H. (2018). Understanding adoption of intelligent personal assistants. *Industrial Management & Data Systems*, *118*(3), 618–636. https://doi.org/10.1108/imds-05-2017-0214

Hoftman, G. D., Datta, D., & Lewis, D. A. (2017). Layer 3 Excitatory and Inhibitory Circuitry in the Prefrontal Cortex: Developmental Trajectories and Alterations in Schizophrenia. *Biological Psychiatry*, *81*(10), 862–873. https://doi.org/10.1016/j.biopsych.2016.05.022

Hutchison. (2023, June 15). *Snapchat Shares New Stats on 'My AI' Usage, Outlines How its Utilizing Chat Data [Infographic] | Social Media Today*. https://www.socialmediatoday.com/news/snapchat-new-stats-my-ai-usage-utilizing-chat-data/653160/

Ivey, R. (2024, May 14). Synthetic intimacy: Is AI solving the loneliness epidemic or making it worse? *Sifted*. https://sifted.eu/articles/ai-companions-intimacy-onlyfans/

Kurian. (2024, July 15). *AI Chatbots have shown they have an 'empathy gap' that children are likely to miss*. https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss

Kurian, N. (n.d.). 'No, Alexa, no!': Designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models. *Learning, Media and Technology*, *0*(0), 1–14. https://doi.org/10.1080/17439884.2024.2367052

Lange, B., Keeling, G., McCroskery, A., Zevenbergen, B., Blascovich, S., Pedersen, K., Lentz, A., & Agüera y Arcas, B. (2025). Engaging engineering teams through moral imagination: A bottom-up approach for responsible innovation and ethical culture change in technology companies. *AI and Ethics*, *5*(1), 607–616. https://doi.org/10.1007/s43681-023-00381-7

Levine, M. S., & Pappas, A. D. (2025, January 6). *Understanding Artificial Intelligence (AI) Risks and Insurance: Insights from A.F. v. Character Technologies*. https://www.hunton.com/hunton-insurance-recovery-blog/understanding-artificial-intelligence-ai-risks-and-insurance-insights-from-a-f-v-character-technologies

Livingstone, S., & Sylwander, K. R. (2024). Children's rights and the UN Global Digital Compact. *Academy of Social Sciences: International Advisory Group Policy Briefings Series*.

Marriott, H. R., & Pitardi, V. (2024). One is the loneliest number... Two can be as bad as one. The influence of AI Friendship Apps on users' well-being and addiction. *Psychology & Marketing*, *41*(1), 86–101. https://doi.org/10.1002/mar.21899

McNamee, R. (2020, June 24). *Social Media Platforms Claim Moderation Will Reduce Harassment, Disinformation and Conspiracies. It Won't*. TIME. https://time.com/5855733/social-media-platforms-claim-moderation-will-reduce-harassment-disinformation-and-conspiracies-it-wont/

OpenAI's weekly active users surpass 400 million. (2025, February 20). *Reuters*. https://www.reuters.com/technology/artificial-intelligence/openais-weekly-active-users-surpass-400-million-2025-02-20/

Patel, N. (2024, August 12). *Replika CEO Eugenia Kuyda says the future of AI might mean friendship and marriage with chatbots*. The Verge. https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview

Roose, K. (2024, October 23). Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

Snap Inc. 2023 Investor Day – Recap. (2023). https://newsroom.snap.com/investor-day-2023

Sprecher, S. (2021). Closeness and other affiliative outcomes generated from the Fast Friends procedure: A comparison with a small-talk task and unstructured self-disclosure and the moderating role of mode of communication. *Journal of Social and Personal Relationships*, *38*(5), 1452–1471. https://doi.org/10.1177/0265407521996055

Storey, D. (2025, January 8). Surgeon General Drafts One Last Prescription. *Psychiatrist.Com*. https://www.psychiatrist.com/news/surgeon-general-drafts-one-last-prescription/

Suskind, D. (2023, August 11). *Essay | The AI Nanny in Your Baby's Future*. WSJ. https://www.wsj.com/health/wellness/the-ai-nanny-in-your-babys-future-999d0e50

Sydnor, V. J., Larsen, B., Bassett, D. S., Alexander-Bloch, A., Fair, D. A., Liston, C., Mackey, A. P., Milham, M. P., Pines, A., Roalf, D. R., Seidlitz, J., Xu, T., Raznahan, A., & Satterthwaite, T. D. (2021). Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. *Neuron*, *109*(18), 2820–2846. https://doi.org/10.1016/j.neuron.2021.06.016

Tran, G., & Davis, E. (2024). *The Generative Identity Initiative: Exploring Generative AI's Impact on Cognition, Society, and the Future*. The Institute for Security and Technology.

Turkle, S. (2017, December 7). Perspective | Why these friendly robots can't be good friends to our kids. *The Washington Post*. https://www.washingtonpost.com/outlook/why-these-friendly-robots-cant-be-good-friends-to-our-kids/2017/12/07/bce1eaea-d54f-11e7-b62d-d9345ced896d_story.html

Turkle, S. (2018, August 11). Opinion | There Will Never Be an Age of Artificial Intimacy. *The New York Times*. https://www.nytimes.com/2018/08/11/opinion/there-will-never-be-an-age-of-artificial-intimacy.html

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, *114*(31), 8148–8156. https://doi.org/10.1073/pnas.1702996114

WHO. (2024, 09). *Teens, screens and mental health*. https://www.who.int/europe/news/item/25-09-2024-teens--screens-and-mental-health