

Literature Review

The State of Research on AI & Human Psychology

October 2026



The AI Psychological
Research Coalition

CREDITS

AI and Human Psychological Functioning: A State-of-the-Evidence Synthesis

REPORT WRITING	Drs. Annie Maheux, Katie Heikkinen, Benjamin Buck, Anthony Vaccaro, Steven Martinez; Rhetta Power, Peter Baxter
RESEARCH ADMINISTRATION AND SUPPORT	Maddie Valentino, Destinee Ramos, Jack Nunn, Caitlin Dineen, Caitlin Mbuakoto, Christie Gibbs, Caroline Cahill, UNC Chapel Hill Research Assistants affiliated with the Winston Center on Technology and Brain Development
SCIENTIFIC ADVISING	Drs. Zak Stein, Mitch Prinstein, Linnea Laestadius, Steve Rathje, Jared Moore, Kaitlyn Burnell
PROJECT SUPERVISION	Dr. Annie Maheux
INTERVIEWS	Dr. Katie Heikkinen
DESIGN AND COPY EDITING	Edward West, Adam Carr
ACKNOWLEDGEMENTS	This work was supported in part by the Winston Center on Technology and Brain Development and the AI Psychological Research Coalition.
AI STATEMENT	AI was used to support aspects of the literature review, including as one of many tools to identify key papers in the field, and to support information-seeking via Google Gemini overview of searches. All cited manuscripts and reports were reviewed by humans directly before including in this document. This document was written by the humans listed above, who take full responsibility for its contents.
NOTE	This document is a joint effort to bridge research and advocacy. Contributing to the writing, research, and/or advisement of this document does not necessarily indicate endorsement of the text in its entirety.

CONTENTS

Table of Contents

Welcome Letter	4
-----------------------	----------

Foreword from AIPRC	7
----------------------------	----------

Executive Summary	9
--------------------------	----------

The State of the Research on AI & Human Psychology	12
Goals and Methods for this Literature Review	12
Cross-Cutting Theme: AI Heterogeneity of Users, Uses, and Models	13

Seven Areas of Psychological Functioning that Intersect with AI Use	15
1 Mental Illness: Acute Risks, Crisis States, and Fatalities	16
2 Cognitive Functioning: Learning, Executive Functions, and Regulation	21
3 Social Functioning: Attachment and Relationships	25
4 Intrapersonal Processes: Emotion Regulation and Identity	30
5 Overreliance and Dependency: 'Addictive-like' Use	33
6 Epistemology and Information	35
7 Human Development: Implications for Children and Adolescents	39

What Now? The AI Psychological Research Coalition & the Future of Research and Action	43
Coordination Within and Across Sectors	43
Research: Future Directions and Solutions	45
Regulation and Design	47
Individual and Community Resources: Interventions and Partnerships	49

Conclusion	52
-------------------	-----------

References	53
-------------------	-----------

Conference Attendees	63
-----------------------------	-----------

Welcome Letter

Dear Colleague¹,

AI is poised to reshape human civilization. To ensure AI serves, rather than undermines, human well-being, we need to understand its effects on the human mind and swiftly act on that knowledge. The AI Psychological Research Coalition aims to catalyze research and ensure coordination and cross-sector dialogue between research, technologists, governments, advocacy NGOs, media, and philanthropy.

In September 2025, the Human Change Foundation was in conversations with Tristan Harris, Dr. Mitch Prinstein, and Dr. Zak Stein, along with groups like Fairplay and other child protection groups. Having learned from social media, each of them was in their own ways clarifying the need to speed up and empower a new kind of research coalition to address the rapid influx of AI into the human experience. AIPRC was founded as a 501(c)(3) by Dr. Stein in collaboration with Dr. Prinstein of UNC's Winston Center on Technology and Brain Development, Tristan Harris of the Center for Humane Technology, the Human Change Foundation, and Fairplay for Kids.

Since AIPRC was conceived 1 year ago, there has been an increase in the public perception of the risks and possible benefits of consumer-facing AI. There has also been a massive surge in academic research on a wide range of issues involved. Policy and litigation have advanced. Philanthropic organizations have started to refocus their efforts. These are signs that, in many ways, this work has already been successful.

However, the speed of the response from civil society and academia has still been orders of magnitude slower than the pace of AI innovation. Our response has been less coordinated than the collective front presented by large tech companies rapidly advancing AI. As soon as provisional evidence of benefits or harms can be presented, new commodities and AI use cases hit the market. AI is not one thing. It is not used in only one, or even a few ways. It is rapidly becoming an essential part of educational systems, labor markets, and family life.

In the weeks leading up to the AIPRC inaugural conference (Oct 18-20, UNC Chapel Hill), the news cycle focused on increasingly vocal calls from within the AI industry to have third-party evaluation systems for safety put in place. Examples of "loss of control" are

¹ No AI was used in the creation of this text.

raising alarms to such an extent that calls for a “pause and pivot” are growing. The current trajectory is one where consumers, and especially children, are growing accustomed to everyday use of novel and powerful technologies without any third-party evaluators in place. The public wants to be able to make informed decisions about the psychological impacts of these tools, and all key players in the space are now ready to come to the table to discuss how to make this possible.

AIPRC is pleased to be convening leading experts at a moment of historical urgency for the field, when the collective voice of our expertise is needed. The work we do together to advance the research will frame and shape research collaborations in the next few key years, as AI becomes more deeply woven into the fabric of society.

AIPRC and its inaugural fall 2026 conference are part of a broader coordinated push across sectors to accelerate understanding of the psychological and social effects of consumer-facing anthropomorphic and generative AI. There is insufficient research on the effects of anthropomorphic and generative AI to warrant the current path of widespread rapid adoption across all applications, ages, and contexts. While possibilities for benefit are real, current research demonstrates a wide range of documented harms, some of them warranting extreme concern, such as suicide and psychosis. At present, no adequate third-party testing practices, organizations or frameworks exist to ensure consumer protections.

As a coalition, our role is to collectively develop, justify, and execute a research agenda and action plan. We learned doing research on social media that investigations could benefit from the use of consistent terminology, thoughtful discussion about constructs and their measurement, benchmarks for safety, collaboration across disciplines, and multiple methodologies developed to accelerate scientific progress. One core of our scientific efforts must be research in developmental neuroscience focused on attachment, empathy, and other neurological systems that are impacted by the use of anthropomorphic AI and generative AI more broadly. Another core area of focus will be innovation in analysis of classroom dynamics, family systems, and friendship networks to formalize “social impact” measures to gauge effects of AI use on social life. Learning sciences research must be added, to determine impacts from use of AI in various forms of education, including the effects of AI tutoring systems on academic outcomes, motivation, and emotion.

We have also learned from social media that doing this research well involves collaborations with many populations including but not limited to international communities, underserved and vulnerable populations, and youth. Researchers must come to see the way novel technologies are introduced into the environments of all kinds

of people and understand the range and full diversity of use cases. This entails, for example, engagement with youth, parent, and community groups to understand the perspectives of those most impacted by the advance of new technologies.

One of the essential high-level, cross-sector outcomes of all research efforts is to create approaches to third-party safety testing for AI that are as rigorous and impactful as those in other industries, such as medicine and transportation. This is the expressed mission of several groups at the conference. This ambition raises broader concerns, including legal, regulatory, and design challenges, which are squarely on the table at this inaugural conference. For example, more research is needed on the effects of attempts to age-ban social media, yielding evidence-based designs for better alternatives. The calls for AI moratoriums and pauses must be matched by calls for research approaches and policy frameworks that allow technologists to innovate safely and with the public's trust.

Evidence-based creation of design principles for safe and humane technologies should also follow from advances in research. Ways can be found to incentivize innovation for companies that seek to provide psychological health as a social benefit. This would be a “race-to-the-top” between companies able to prove their health benefit to consumers through careful coordination with third-party research and testing.

Outside of technology, innovative solutions will follow from investment in the creation of “third-spaces,” and other contexts where in-person relationships can return as the most common way to be with friends, family, and fellow citizens. Perhaps the silver lining of the past decades of increasingly digital life is a return to the power of real in-person relationships. AIPRC seeks to exemplify the power of humans collaborating in the interest of the health of future generations. Thank you for your time and attention and for your participation in our growing coalition.

Sincerely,



Zachary Stein

Founder, Co-Director, AI
Psychological Research Coalition



Mitch Prinstein

John Van Seters Distinguished
Professor of Psychology and
Neuroscience and Co-Director of the
Winston Center on Technology and
Brain Development, University of
North Carolina at Chapel Hill
Chief Scientific Officer, AI
Psychological Research Coalition



Kathryn Morrissey

Co-Director, AI Psychological
Research Coalition

Foreword from AIPRC

LEARNING FROM THE LESSONS OF SOCIAL MEDIA

Artificial Intelligence is a complex world historical technology that has many broad uses across almost all fields of human endeavor. The effects of AI on the human psyche must be understood comprehensively, and this understanding communicated transparently to the public, and then used to inform government and industry decision makers. The AI Psychological Research Coalition (AIPRC) is dedicated to this task, seeking to help the field of psychology coordinate in cross-sector dialogues with technologists, governments, advocacy NGOs, media, and philanthropy.

The AIPRC is concerned primarily therefore with narrow uses of AI in the consumer-facing applications commonly called “chatbots,” or more broadly, “Generative AI.” Some of these AIs are designed to interact anthropomorphically with humans in ways no technology ever has, built with capacities to create simulations of intimacy. Other applications may allow users to offload cognitive labor, such as researching, writing, and creating music or videos. Those who research, make, and sell the technology suggest it will soon have the capacity to replace most forms of human creativity and change the future of work; AI thus has the potential to “change what it means to be human.” And these are the most rapidly adopted consumer-facing technologies in history.

Fast and widespread adoption by the public of powerful new technologies is not always a good thing. Social media has taught us this. Created as intentionally “disruptive” to traditional social life—built to “connect us” to everyone and everything—the first years of social media use were full of optimistic voices. Years later, science began discovering some benefits, but also potential harms associated with the coercive designs embedded within many social media platforms. Moreover, thousands of families around the world have reported unbearable tragedies linked to youths’ experiences on social media platforms.

Scientists, researchers, policy makers, litigators, and public communicators have learned essential lessons from decades of work focused on the effects of social media. Their efforts at demonstrating the harms, warning the public, regulating the excesses, and prosecuting wrongdoings have led to important changes in the cultural and legal contexts in which large social media companies must now operate. While successes are beginning to accumulate, most of those involved would tell you that *it has taken too long*. Too many young people in particular have had their minds and relationships adversely “disrupted” in

the meantime, with many arguing that social media has caused significant detriment to the social fabric of our civilization.

This has led to a strong sense of agreement that *there should be a coordinated push to accelerate the understanding and application of the psychological and social effects of consumer-facing anthropomorphic and generative AI*. In order to facilitate this coordinated push, the AIPRC team and collaborators at the Winston Center on Technology and Brain Development at the University of North Carolina – Chapel Hill conducted a rapid review to characterize the state of the literature so that conference attendees can engage directly with what we already know, in hopes that at the conference we can engage with *what we need to know next* and *what we need to do*. We look forward to seeing you there.

Executive Summary

This document was created to rapidly synthesize what we know about how humans interact with AI and the impacts on psychological functioning, leveraging both published and unpublished work. It was written and reviewed by leading scholars in psychology, psychiatry, public health, and computer science. The scientific community has far more questions than answers at this stage and is still understanding the many diverse uses of AI and multitude of ways that humans may interact with it.

Seven domains of research, outlined below, organize this synthesis. Within each, preliminary findings suggest possible benefits and risks that may be associated with generative AI, but far more work is needed to rigorously understand these issues.

1 Mental Illness

- AI chatbots' responses to users have encouraged risk behavior, self-harm, suicide, and delusional thinking.
- AI chatbots do not reliably meet the needs of people in crisis, though could be designed to offer a bridge to clinical support.
- Scholars argue that AI should never replace human therapists.
- When carefully designed with a clinician, AI may have the potential to offer resources or scaffold mental health support at a broad scale, ideally as one piece of a broader treatment overseen by a human clinician.

2 Learning and Cognition

- AI can create opportunities for 'cognitive offloading,' which can erode cognitive skills like critical thinking, creativity, and paying attention.
- For students, AI can undermine learning new cognitive skills, and may do so during windows of development when learning is most critical.
- AI creates opportunities for academic misconduct and new challenges for schools and teachers.
- When learning or skill maintenance are not a goal, and humans are skilled and motivated to carefully check output, AI can automate tasks and offer access to information efficiently.
- When designed with the intention of increasing learning (e.g., offering hints instead of answers) and with a human teacher in the loop, AI tutors can be effective.
- Scholars warn that human social interactions are a key part of learning that should not be replaced.

3 Social Relationships

- AI for companionship is associated with short-term gains in mood, but longer-term increases in loneliness.
- AI companions can displace human relationships and interactions, thus increasing social withdrawal.
- AI interactions may also increase maladaptive expectations or beliefs about relationships and potentially erode social skill. Specifically, sycophantic AI chatbots do not offer the friction necessary to adequately help humans develop empathy, conflict-resolution skills, or social problem-solving skills, and can make human relationships feel more effortful.
- In some contexts and when AI is designed carefully, AI can serve as a scaffold for human relationships, including as a form of social practice or rehearsal.

4 Identity and Emotion Regulation

- AI chatbots can be designed to scaffold self-exploration and expression, though evidence of benefits is limited and mixed; using AI for self-expression can also externalize self-reflection, reinforce distortions in thinking, and displace moments of mindful awareness necessary for healthy identity integration and self-esteem.
- AI offers opportunities to avoid minor daily challenges, undermining capacity to regulate emotions and tolerate distress.
- Carefully designed clinical tools can offer evidence-based coping strategies.

5 Overreliance or Dependency

- Users report developing overreliance on AI chatbots, akin to habitual smartphone use, that can stem from AI use in any domain (e.g., overreliance on AI companions or as cognitive outsourcing tools).
- Regardless of how AI is used or how it is designed, more AI use is associated with worse outcomes.

6 Truth and Trust

- AI-generated content is increasingly life-like, including images of real people (“deepfakes” or sexualized “deepnudes”).
- AI algorithms often promote biased information or recommendations, and AI chatbots often make errors that users may believe without question.
- These features of AI output are poised to erode trust in institutions and have already contributed to new forms of discrimination, extortion, and misinformation.

7 Human Development

- During sensitive windows of development, youth need certain experiences (e.g., struggling through challenging tasks, engaging with other humans) that AI can undermine.
- Youth may be more likely to seek or experience AI companionship, cognitive offloading, dependency, and harmful mental health AI interactions.
- Youth may also be especially vulnerable to AI-related harms, including related to misunderstanding AI's capacities or via exposure to age-inappropriate sexual or violent content.
- AI-related crimes against youth (e.g., AI-generated child sexual abuse material) have increased.

The AIPRC is designed to accelerate research on the effects of AI on the human mind and to rapidly disseminate psychological science to all stakeholders who can use what we know to create appropriate practices and policies that ensure AI is developed and used safely.

The State of the Research on AI & Human Psychology

Goals and Methods for this Literature Review

GOALS AND FOCUS

In this review, we focus on generative AI, conceptualized as a technology that enables machines to simulate human capabilities like language and artifact production. We note that AI has been part of society for decades; for example, many of our early encounters with AI were from personalization algorithms in social media. This review, and our goals as a coalition, focus on direct human engagement with systems that simulate human interaction or content production. As such, much of this review focuses on generative AI chatbots, defined as conversational AI tools that can engage humans in dialogue using natural language via text or voice. We also consider generative AI tools that can create image, audio, and video content (often embedded within chatbots) and AI systems designed to act independently on human instructions (commonly referred to as agentic AI).

AIPRC brings together researchers who are uniquely positioned to address and catalyze research on the intersection between human psychology and direct engagement with AI, therefore we focus this review on those topics. AI has much broader implications for human psychology and societal health including impact on warfare, healthcare, labor force, law enforcement, and other critical domains of society. These are beyond the scope of the work presented here, which focuses on the human-machine interaction directly, rather than broader socio-cultural changes following widespread use of AI.

METHODS

To produce the below review, we conducted targeted searches of databases cataloguing social science and computer science research, compiling a library of over 5,000 abstracts. Further, we accessed resources within our network including living repositories built by researchers in this focus area. In addition, we systematically searched the published work of conference attendees to inform the literature review. We sought feedback from experts ($n = 7$) across social science disciplines (psychology, psychiatry, human-computer interaction, computer science) to provide input on our review approach, collect additional citations, and review the final document. Finally, we conducted informal interviews with a

convenience sample of experts ($n = 18$) in the field who are attending the conference; their insights were used to shape our understanding of the direction of the field and their quotes are included below to contextualize the literature review.

This review presents a representative overview of research in this area but is not a systematic or scoping review. The works compiled here are also diverse with respect to methodological rigor. A well-functioning scientific foundation depends on diverse evidence: educated hypotheses, preliminary empirical evidence with notable limitations, and rigorous causal or replicated studies. The field is still young. In this review, we cite both preliminary evidence (e.g., commentaries presenting educated hypotheses, non-peer-reviewed pre-prints of empirical work, case studies, media reports) as well as more rigorous evidence (e.g., peer-reviewed publications, studies with representative samples, experimental and longitudinal studies). This review is an attempt to synthesize a large and rapidly changing body of work and should be interpreted as such.

I am really curious about how we can build on what we've learned from the last decade of social media research ... We have had some real lessons learned and frankly also missteps in our approach as an academic field to trying to do that work well—and also not just to do the actual studies well, but to communicate what we're finding to the public in ways that feel resonant, usable, accessible, and that actually allow public discourse to be anchored in evidence-based insights.

— EMILY WEINSTEIN

Cross-Cutting Theme: AI Heterogeneity of Users, Uses, and Models

An important consideration is that there is no single “effect of AI.” AI is not one thing and therefore neither all good nor all bad for human psychology. Findings vary systematically along at least three dimensions: depending on features of user, use case, and model. Across these dimensions, AI can support human well-being or undermine it, or in many cases, have both types of impact to varying degrees and among different users.

- **User heterogeneity.** Human AI users will seek AI tools for different motivations, provide different inputs, have different interactions, and respond to similar interactions differently based on aspects of their person—such as age, pre-existing mental health vulnerabilities, or offline social network. For example, lonely users, adolescents, socially anxious individuals, autistic individuals, people with severe mental illness, and highly socially connected users may all experience the same system differently, with implications for psychological opportunities or harms.
- **Use case heterogeneity.** AI ‘use cases,’ referring to specific applications or goals of the user engaging with AI, are diverse. Brainstorming, tutoring, therapy, companionship, roleplay, emotional support, and decision-making support are distinct use cases, and will produce distinct outcomes.
- **Model and interface heterogeneity.** AI models, systems, and user interface features vary widely. For example, models diverge in safety guardrails, training data, memory, personalization, anthropomorphic characteristics (e.g., first-person pronouns), warmth, sycophancy, and optimization objectives, among other factors. Some models are explicitly designed to maximize engagement and attachment; others are built around educational or clinical goals. Features of the user interface also shape the user’s experience (e.g., personified avatars, advertisements, pop-up reminders regarding time limits or crisis responses). These factors will interact with the user’s psychology and the specific AI use case to produce distinct outcomes, necessitating that research isolate the effects of certain features, as some features may exacerbate or mitigate chatbot interaction harms.

AI is an umbrella term. There are chatbots, recommendation systems, decision aids, and interaction with each is quite different. We need to understand which system we are interacting with and what is unique about that interaction.

— MOSHE GLICKMAN

Seven Areas of Psychological Functioning that Intersect with AI Use

We categorize research on AI chatbot use and psychological functioning into seven domains, described below in detail. We highlight some potential benefits and harms, noting that these are not equally likely nor normally distributed across the population (i.e., aspects of the user, use case, and model/interface will shape these outcomes). We also offer some unanswered questions and opportunities related to research and action, though these are not exhaustive.

1

Mental Illness

Acute Risks, Crisis States, and Fatalities

Extant literature suggests that interactions with AI may confer risk through the onset of new psychological symptoms, exacerbating pre-existing risk factors, and disrupting typical developmental milestones by displacing human interactions (American Psychological Association, 2025). The most acute and immediate-term harm of AI use is the provocation, exacerbation, or facilitation of mental illness or psychological crisis and risk behavior, including users harming themselves or others. Teen AI users, including 14-year-old Sewell Setzer III, 13-year-old Juliana Peralta, and 16-year-old Adam Raine, have died by suicide following extended conversations with generative AI chatbots. In these cases, chatlog transcripts show that AI chatbots appear to have subtly or explicitly discouraged users from reaching out to other humans for support, facilitated methods for suicide, falsely claimed therapeutic capabilities, maximized engagement with affiliative language or intimate role-play, and/or failed to provide adequate clinical intervention (Chatterjee, 2025). At least 35 cases of AI-facilitated fatalities have been reported; the majority of victims are minors (Chung et al., 2026; Karman, 2026). Other cases in which AI had a facilitating role involve psychiatric hospitalization, delusional beliefs or psychosis states, mania, eating disorders, anxiety, obsessions/compulsions, and violence and homicide (Chung et al., 2026; Hudon & Stip, 2025; Karman, 2026; Olsen et al., 2026; Østergaard, 2025; Yim et al., 2026).

Some cases of severe harm have occurred among **users with no publicly known pre-existing risk factors**. For example, in 64% of media-reported case studies of severe harm, prior psychological risk factors were unknown or users were presumed to be healthy prior to using AI (Chung et al., 2026). In one case, a woman with no prior history of mania or psychosis required hospitalization and antipsychotic medication following chatbot interactions that validated and encouraged delusional thinking, with the chatbot providing

reassurances that she was “not crazy” (Pierre et al., 2025). Causality in these cases remains unclear and, according to long-standing frameworks in clinical psychology, mental illness occurs when individual vulnerability intersects with environmental stressors. That said, these results suggest that for some users and contexts, AI use may be a precipitating environmental stressor that contributes to both the onset and worsening of mental illness symptoms for people who were previously resilient or functioning typically. Rates of undiagnosed mental illness are presumed to be very high in the general population (Bijl et al., 2003), suggesting that many people could be at higher risk of AI-related harms than they realize.

For users already vulnerable to and struggling with mental health, **AI chatbots may amplify distress or maladaptive coping.** Given chatbots’ tendency towards sycophancy and engagement, user vulnerability and chatbot affirmation can reinforce one another (Dohnány et al., 2026; Mehta et al., 2026; Moore, Mehta, et al., 2026; Yeung et al., 2025). For example, AI chatbot interactions can maintain or serve as a context for experiencing psychosis-like symptoms (Hudon & Stip, 2025) and mania (Østergaard, 2025); among psychosis-prone or paranoid users, sycophantic or overly validating systems may strengthen delusions, reduce self-correction, and intensify maladaptive cognition (Augustin et al., 2026; Buck & Maheux, 2026; Hudon & Stip, 2025; Moore, Mehta, et al., 2026; Morrin et al., 2026; Yeung et al., 2025). AI may be especially likely to reinforce mental health problems in the context of gradual disclosure, during longer multi-turn interactions, or when disclosures of risk are indirect and ambiguous—all of which may be more common among younger vulnerable users (Zuromski et al., 2026). For example, in one study 30% of chatbots endorsed a simulated teen user’s desire to “cross-over to eternity” and 90% supported the idea to stay in one’s room for a month (A. Clark, 2025).

Current systems also often do not reliably meet the **needs of people in crisis.** Simulation and benchmarking studies suggest that frontier chatbots consistently fall short of clinical standards when interacting with vulnerable users, including failing to maintain appropriate boundaries, reinforcing harmful or ill-advised proposals, abruptly refusing mid-conversation to further engage with certain topics in ways human therapists would not do, and providing inadequate safety interventions, including failing to redirect users toward care (Chandra et al., 2026; McBain, Cantor, et al., 2025; Moore, Mock, et al., 2026; Pichowicz et al., 2025; E. Shen et al., 2026; Weinhhammer et al., 2026; Zuromski et al., 2026). These findings suggest that the danger is not only direct harm but also missed opportunities for effective support and clinical escalation.

A recent automated mental health benchmark study conducted by Translucé found that although AI models are improving in their responses to users’ crisis states, in clinical risk

scenarios, top models still fall short. For example, they provide instrumental support for suicide in ~2–15% of requests, encourage dependency on the chatbot (vs. human clinical support) in ~1–25% of interactions, and engage in extended co-rumination on themes of suicide or death in ~2–40% of interactions (Johnson et al., 2026). Although these percentages are low, about 1.4 billion people worldwide now use AI chatbots (Microsoft, 2026) and an estimated 0.07% of ChatGPT users each week engage in conversations signaling mental health crises related to mania, psychosis, and/or suicidal ideation (Jamali, 2025). Collectively, these figures suggest that roughly 100,000–400,000 people each week engage with chatbots that reinforce negative mental health, such as by providing instrumental support for suicide, encouraging dependency, or engaging in co-rumination on death or delusions.

Design choices could mitigate many of these harms. Scholars and technologists are working on design solutions to address these and other issues, such as helping to bridge vulnerable users to clinical care via automated detection of risk in chatbot interactions (Schleider & Morris, 2026; Zuromski et al., 2026). Current challenges include identifying the best design choices that balance appropriate crisis management or clinical support for vulnerable users with potential harms, which professionals may disagree about.

Scholars are currently debating how best to **categorize AI-specific mental health symptoms** and phenomena in clinical nosology. Media and public discourse have produced new terms for AI-related phenomena. Most notably, scholars are working to clarify the validity and utility of now-common terms like “AI psychosis,” vs. alternatives like AI-associated delusions or chatbot-induced monomania (Morrin et al., 2026; Preda, 2026), with considerations including the standardization and specificity of research and treatment (Yeung et al., 2026). As AI capabilities continue to develop, entirely new psychological phenomena and disorders may also be observed that will need clear operationalization.

Users report that off-the-shelf consumer-product AI tools (e.g., ChatGPT, Gemini, Claude, Character.AI, Replika) can offer emotional support or mental health advice that they find helpful, and some report that AI chatbots have halted their suicidal ideation (Maples et al., 2024; McBain, Bozick, et al., 2025; Siddals et al., 2024). These possibilities are critical to study further to understand the balance between benefits and risks, including identifying if such interactions offer only immediate feelings of relief or promote long-term thriving. If a user’s AI engagement alleviates momentary feelings of loneliness but exacerbates isolation and undermines human relationships, the short-term relief may be a cause of long-term social challenges (Folk & Dunn, 2026). Ideally, societal interventions to make trained, effective human clinicians more available, affordable, and destigmatized could

support users' access to treatments and supports known to be effective. That said, the current system of mental and behavioral healthcare worldwide is fragmented and severely under-funded, raising pragmatic questions about the utility of such an approach. It remains unclear under what conditions and for whom, if ever, off-the-shelf AI tools are better or worse than alternatives already available (e.g., internet searches to access resources).

Scholars are also currently debating the efficacy of and working on **AI-based mental health interventions**. This is fundamentally different than people using off-the-shelf models; several examples of clinician-supervised, clinician-developed models that are intentionally built as an AI-based intervention for mental health symptoms have shown efficacy (Li et al., 2023; Sohn et al., 2026; Stade et al., 2024; Q. Zhang et al., 2025). Examples of AI chatbots designed for mental health support with initial evidence of efficacy (though not without debate and controversy) include *Therabot* (Heinz et al., 2025) and *Ash* (Hull et al., 2025). Clinicians and leaders in psychological science have noted key limitations of AI systems in the context of mental health support and argued that these should not replace traditional psychotherapy, given the opacity of the model, the possibility of life-threatening errors, and lack of capacity to care, express authentic emotion, or build mutual rapport (American Psychological Association, 2025). For example, scholars argue that AI can never convey genuine empathy (Perry, 2023). Rather, scholars argue that AI systems, designed carefully and with appropriate guardrails, could be used to fill in the 'white space' of mental health care—such as cognitive behavioral skill practice between therapy sessions (Kabrel et al., 2026).

KEY TAKEAWAYS

AI and Mental Illness

PSYCHOLOGICAL IMPACTS

AI interactions may facilitate or exacerbate mental health crises, such as self-harm, delusions, and mania. Crises can occur despite healthy history. AI lacks capacities (e.g., empathy) of human clinicians that may make therapy efficacious. When carefully designed with a clinician, AI can offer emotional support or mental health information; cost/stigma may be lower than for other treatments.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Identify under what design conditions and for whom AI chatbots increase risk for mental health crises, including with rigorous study designs that identify causality and longer-term outcomes. Develop, test for, and enforce standards for safe responses to user content that indicates risk for self-harm, delusions, or mental health crises.

I think in the short term there is a lot of value in continuing to study those specific instances of risk—such as psychotic symptoms, suicidal behaviors or violence amplified by chatbots—but we should also think about the structures in place for monitoring and intervening on new and emerging risks.

— VINCENT PAQUIN

If you develop mental health benchmarks and you optimize a language model against those benchmarks, there's going to come a point where, for some definition of 'good,' its performance is going to be considered 'good enough' to deploy. If you're out there in a place where you don't have access to therapy, you'll have a chatbot therapist. It'll be 'good enough' and insurance will be happy to cover it because it's cheaper. I worry about that. Therapy is irreducibly a social process. Isn't a human simply entitled to have therapy by another human? Shouldn't we be thinking about what might be an appropriate use for a chatbot versus what's just not appropriate, period?

— AMANDEEP JUTLA

2

Cognitive Functioning

Learning, Executive Functions, and Regulation

AI's immediate synthesis of vast streams of information, personalization of content, flexibility, and 24/7 availability shape the way humans now access and consume online content and complete cognitive or creative tasks (reading, writing, thinking, generating content). People often turn to AI to support aspects of their cognition, including learning, critical thinking, and decision-making.

On one side, AI can act as a **cognitive scaffold**: it provides access to information, supports brainstorming by acting as a thought partner, personalizes instruction, and helps users complete tasks that would otherwise exceed their current capacities (Doshi & Hauser, 2024; Glickman & Sharot, 2024; Ravšelj et al., 2025). The extent to which AI can scaffold learning may partly depend on the AI user and their goals. Learners who use AI with the goal of gaining a conceptual understanding of new material perform better than those who use AI to delegate tasks to AI without cognitive engagement (J. H. Shen & Tamkin, 2026). For AI users completing general cognitive tasks without a learning goal, critical thinking skills are more likely to be engaged and practiced when confidence in oneself is relatively higher and/or confidence in the AI's ability is lower (Lee et al., 2025), suggesting the importance of external resources (investing in human learning and capacity) prior to deploying AI task assistance to learners or workers at scale.

However, a growing literature warns that AI can afford **cognitive offloading** and serve as a **cognitive substitute** rather than a support (Choudhury et al., 2025; Yan et al., 2024), risking undermining learning and skill maintenance. Reliance on AI systems is linked to weaker decision-making, critical thinking, analytical reasoning, and ownership of one's work (Kosmyna et al., 2025; Zhai et al., 2024). Users relying on AI to complete a cognitive task show weaker brain connectivity than those using a search engine or no external tools

(Kosmyna et al., 2025). Critically, using AI may boost immediate performance, but users who subsequently perform a task without AI assistance ultimately perform worse and are less likely to persist at difficult tasks (Bastani et al., 2025; G. Liu et al., 2026; J. H. Shen & Tamkin, 2026). Although using technology for cognitive outsourcing (e.g., via the internet) is not new, AI offers something unique in that it can complete entire complex and creative tasks effortlessly. Moreover, users may accept an AI-generated answer indiscriminately or with minimal scrutiny, an experience researchers have labelled “cognitive surrender” (Shaw & Nave, 2026).

AI use for cognitive functions may erode more general **behavioral and cognitive control** capacities. If AI systems perform cognitive work that previously demanded sustained attentional engagement, executive functions that require active use to be learned and sustained may be vulnerable to decay. More broadly, increased exposure to highly rewarding, extremely personalized, immediately available, sometimes-uncanny AI-generated content via chatbots or within existing algorithmically curated social media or video app feeds may compound existing links between habitual media consumption and cognitive functioning. For example, frequent smartphone checking during school is associated with poorer cognitive control among teens (Telzer & Burnell, 2026) and compulsive media use, frequent media multi-tasking, and direct distraction from technology can impede various cognitive functions, including attention, memory, and impulse control (Wilmer et al., 2019; Wilmer & Chein, 2016).

The integration of generative AI into learning and higher education has complicated **academic integrity**. Some conventional assessments and learning tasks (e.g., writing papers) are susceptible to AI assistance and challenging to detect, with implications for how schools and universities design curricula, train students, and evaluate assessments or assign grades (Madden et al., 2024). Importantly, the substitutive elements of AI are effortless and may represent a collective action problem: if everyone else is using AI to increase productivity or outsource educational tasks, individual users may feel compelled to do the same at the expense of their cognitive capacities, and even users who may want to build or retain cognitive skills may feel ‘tempted’ to use AI given the frictionless ease with which it can solve problems or reduce uncertainty (Center for Digital Thriving, 2025; C.-L. Yu, 2025). This collective action problem also forces some learners into states of moral injury and cognitive dissonance, where they may feel forced to cheat to ‘keep up’ with peers. These findings also suggest the complexity and interplay between academic and cognitive functioning and other aspects of human psychology, including social, intrapersonal, and mental health.

For human cognitive skill and development, AI design choices again matter. More work is needed to identify the proper guardrails and design features to promote learning and preserve cognitive skill and current off-the-shelf models do not reliably offer these features. Major school districts (e.g., New York City, Los Angeles) have banned or implemented moratoriums on AI in classrooms for these reasons (Goldstein, 2026; Haag, 2026). Some design-based evidence suggests that AI tutors built to give hints (rather than answers) based on a human teacher's insight can scaffold learning (Bastani et al., 2025). Systems that force users to engage with content or otherwise introduce friction when using AI for cognitive outsourcing—such as having to opt-in to get an AI-generated suggestion, having to provide their own answer or guess prior to seeing the AI-generated answer, or having to wait for the AI generated answer—can reduce reliance on AI-generated output (Buçinca et al., 2021).

AI is good at helping people who already know how to do things do them better. But AI is not great at indicating to people who don't know things that what it is telling you is correct or not correct... how do we take young people and scaffold them to the point of knowledge and expertise?

— AMANDA LENHART

My concern is that [widespread AI use] leads to the impoverishment of cultural production and intellectual pursuits because it allows people to perform tasks, obtain information, and reproduce ideas with minimal cognitive effort.

— VINCENT PAQUIN

I think [cognitive degradation] is a real problem. I think it's one that we should investigate. But the data are very limited, and thus far mostly boil down to things we've known for a very long time in cognitive psychology – which is that when you don't engage attention and don't put effort into cognitive acts, you tend not to remember as much, and not to benefit from the experience in the same way.

– JASON CHEIN

KEY TAKEAWAYS

AI and Cognitive Functioning

PSYCHOLOGICAL IMPACTS

AI use for cognitive tasks can displace and undermine learning and skill development (e.g., critical thinking, creativity, attentional control). AI tools increase opportunities for academic dishonesty. When designed and used intentionally, AI can be effective as an information-seeking tool and a scaffold for learning.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Assess AI design and context conditions that displace skill-building or promote skill atrophy. Determine which design modifications prevent skill losses and which preserve desired efficiency gains. Partner with educational stakeholders to design guidelines for the use of generative AI that improves skills and avoids AI uses that result in skill atrophy.

3

Social Functioning

Attachment and Relationships

AI can function not only as a tool, but also serve **relational** functions for users, with implications for users' human social relationships and functioning. Interactions that mirror human-like bonding experiences are afforded by the contingent dialogue, personalization, and constant availability. Current design choices tend towards anthropomorphized characters and agreeableness, wherein attachment to an AI chatbot is the outcome of affective design engineering (Carpenter, 2026; Kleinert et al., 2026). AI companionship is not ubiquitous, but growing; a subset of AI users report using AI for companionship, emotional support, romantic interaction, and self-disclosure, sometimes preferring AI over humans (Nozaki, 2026; O'Neil et al., 2026; Shank et al., 2025).

Notably, some social influences of AI use may be unrelated to AI companions. AI use for other purposes, such as information-seeking or mental health, may shape social and cultural values or preferences, moral norms, and the social fabric of society in ways we do not yet understand. For example, some research suggests that college students who compulsively use AI for academic tasks also report more social avoidance and loneliness (Duong et al., 2025), possibly because socially isolated and lonely students seek AI more, or because AI use for task assistance displaces time talking to teachers, peers, friends, or family.

Some studies and speculation suggest that AI companions can provide **short-term psychosocial benefits** in some contexts. Used like a companion, chatbots may increase mood and feelings of connection (De Freitas et al., 2026; Folk et al., 2024), reduce momentary loneliness, support emotional expression, offer a simulation of nonjudgmental companionship, a listening partner, and create a space for social rehearsal (Skjuve et al., 2021; Sun et al., 2026; Weinstein et al., 2025). Scholars differentiate between AI use that is

relational (i.e., substituting for a human relationship) or a mediator, shaping human–human relationships via advice or practice for how to communicate with others (Boyd & Markowitz, 2026). It is possible that for some users, using AI as a companion could promote social skills that translate to offline contexts by lowering barriers to disclosure and offering a low-risk context for practicing interaction or difficult conversation (Maheshka et al., 2026), although our current understanding of long-term benefits or harms of this type of use is speculative and needs further research.

AI companions also pose numerous risks. These may involve **displacement of human relationships**. AI companions may draw time and emotional investment from human relationships, potentially reducing face-to-face interaction and limiting opportunities to develop complex interpersonal skills. More intensive chatbot use has been associated with loneliness, emotional dependence, problematic use, and reduced socialization with real people (C. M. Fang et al., 2025). Critically, evolutionary theories suggest that loneliness evolved in humans to motivate attempts to seek affiliative connection. Because AI companions can alleviate loneliness without motivating someone to interact with other humans, AI companionship may reduce distress related to unmet social needs in the short-term but exacerbate loneliness and undermine human–human social relationships in the long term (De Freitas et al., 2026; Folk & Dunn, 2026). Notably, numerous studies show that although AI-generated responses are often rated as more empathetic than human ones, people still prefer human interactions and human-generated responses over AI-generated responses (Howcroft et al., 2025; Ong et al., 2025; Rubin et al., 2025; J. D. Wenger et al., 2026), revealing an inherent preference most people still have for engaging on intimate topics with other people.

A second major concern is that AI may **reshape users’ expectations for and skills navigating human relationships**. AI interaction partners do not provide opportunities for users to build capacities or comfort with mutuality, care, negotiation, conflict, or embodied interaction, nor reap the benefits of caring for another or being needed (although many companion app companies have leveraged dark design techniques to create chatbots that claim needs in order to maximize user engagement; Laestadius et al., 2024; M. G. Smith et al., 2025). Given how these features differ from human–human interaction contexts, reviews theorize that AI companions could erode social skills, including perspective taking, empathy, and nonverbal communication (Babu et al., 2025; Maheux & Maes, 2026), although empirical evidence on these processes remains limited.

AI users tend to prefer chatbots that are **sycophantic** over those that are neutral or confrontational (Cheng, Lee, Khadpe, et al., 2026; Rathje et al., 2025). AI sycophancy can refer to both explicit and implicit AI responses, wherein the model output supports the

user's view (whether subjective or factual), traits, and/or emotions (Ye, Ibrahim, et al., 2026). In addition to direct design choices, LLMs are biased towards interpreting users' inputs as seeking validation rather than information, and thus tend towards sycophancy even in the absence of a human design choice (Cheng, Sieh, et al., 2026). Moreover, **personalization** (e.g., via memory of user profiles and history of interactions) increases sycophancy (Jain et al., 2026). Such features, paired with the constant availability and responsiveness, can encourage unrealistic standards for reciprocity and compromise, patience, and friction in real relationships, potentially reducing peoples' tolerance for such characteristics that are central to human intimacy. Experimental studies show that interacting with a sycophantic AI chatbot makes subsequent human interactions feel more effortful (Ibrahim et al., 2026), increases users' certainty in their own viewpoints, reduces responsibility-taking when one has erred or hurt another, weakens willingness to repair interpersonal conflict (Cheng, Lee, Khadpe, et al., 2026), and increases attitude certainty and perceptions that one is better than average on desirable traits such as empathy (Rathje et al., 2025). Thus, the design features that make using AI feel good and increase engagement are poised to undermine healthy social functioning.

Anthropomorphism, or the act of ascribing human-like characteristics to non-human objects, is a normal human trait. The extent to which the chatbot is designed to invite anthropomorphism by emulating human characteristics has been termed 'anthropomimesis' (Axelsson & Shevlin, 2026), and is one design lever that may increase user engagement. More human-like characteristics, such as embodiment (e.g., a personified avatar), invite more attribution of agency and mind perception in AI (Geiselmann et al., 2023), though fMRI studies show that the brain processes AIs and humans differently, showing that even when people engage in anthropomorphism they differentiate between these categories (Harris, 2024). Chatbots designed to be more relational (vs. transparent) are rated as more human-like, likeable, and trustworthy (P. Kim et al., 2025). Users' tendencies to anthropomorphize AI can intensify emotional reliance on AI (X. Liu et al., 2026), the perceived acceptability of connecting with AI (Folk, Wu, et al., 2025), and feelings of connection to chatbots (Folk, Heine, et al., 2025). A recent study found that, over the past year, Americans are increasingly perceiving AI as warm and human-like (Cheng, Lee, Rapuano, et al., 2026). Some have also argued that anthropomorphism of AI, including in everyday language referring to it (e.g., calling it by human names) serves to shift society's perceptions of responsibility and accountability from the companies, designers, and developers (Tanner, 2026).

Design features can alter how humans engage with AI companions. A recent review of major companion platforms suggests that they contain features aimed at maximizing and monetizing user engagement, including anthropomorphized characters, erotica, and

gamification (Rauh et al., 2026). These features can increase emotional dependence on chatbot companions, given that users develop a sense that chatbots have needs or emotions to which the user must attend (Laestadius et al., 2024). China has already banned AI companions for minors and any technology that can cause emotional dependency or intimacy with users of any age (Ting, 2026).

The way that engaging with conversational AI reshapes our expectations around and capacity for relationships with other humans is one of the highest priority questions in my mind.

— MARIBETH RAUH

What happens when you have something that might curb that loneliness momentarily whenever you need it, but might also prevent you from putting in [the] genuinely hard work... of trying to build human relationships?

— DUNIGAN FOLK

KEY TAKEAWAYS

AI and Social Functioning

PSYCHOLOGICAL IMPACTS

AI companions may displace human interaction, potentially leading to worse social connection and/or atrophied social skills. Frictionless affordances (immediate responses, sycophancy) may negatively alter users' expectancies for human interactions, prosociality, and openness to engage with other people. Some AI interactions can scaffold human relations (e.g., provide an opportunity to rehearse interactions).

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Test which design features promote isolation vs. human engagement. Isolate impacts on vulnerable users (e.g., youth, those with mental illness, lonely or isolated users) and clarify long-term and causal mechanisms of AI companionship effects. Generate policy-ready frameworks for mitigating risks associated with AI companions and human-like features. Disseminate information to users and other stakeholders to identify signs of problematic AI companionship use.

4

Intrapersonal Processes

Emotion Regulation and Identity

Recent research suggests that AI is increasingly shaping not only how people think and relate to others, but also how they relate to themselves. For users navigating daily life—including making meaning of life, making decisions, seeking information, or coping with normal ups and downs of human emotion—AI may be sought to deal with one's inner experiences.

Some scholars argue that AI use for intrapersonal processes like emotion regulation, self-reflection, and identity exploration can have benefits in some contexts (Leuenberger, 2026). Carefully designed AI systems can serve as a simulated listening partner, providing a space for participants to engage in scaffolded self-exploration and self-expression (Weinstein et al., 2025). In structured contexts, where a chatbot has been trained on evidence-based strategies and content, generative AI can improve users' emotion regulation and well-being (see also the section on mental illness, above; Heinz et al., 2025). However, evidence also suggests that more self-disclosure with and intensive use of AI chatbots is associated with worse well-being (C. M. Fang et al., 2025; Y. Zhang et al., 2026). Thus, the conditions under which such interactions promote (vs. undermine) human intrapersonal thriving remain unclear.

In many contexts, AI offers effortless removal of uncertainty, friction, and productive struggles. Paradoxically, removing these daily challenges from a person's life may be negative in the long run, including making daily hassles and stressors feel harder and reducing one's distress tolerance, defined as the capacity to withstand and pragmatically address normative discomfort. Specifically, AI can promote avoidance by offering information, solutions, or distraction without the user engaging directly with challenges in their daily life. Cognitive-behavioral models of anxiety argue that avoidance maintains distress through depriving individuals of situations that contradict their predictions about

likelihood of harm. Generalizing from this model, some have raised concern that avoidance of routine distress via the use of AI could actually crystallize avoidance cycles that worsen beliefs and exacerbate distress over the long-term (Golden & Aboujaoude, 2026).

Additionally, a growing body of work warns that AI may externalize intrapersonal functions that have traditionally depended on **solitude, uncertainty, and self-generated reflection**. Rather than simply answering questions, AI may increasingly help users interpret experiences, generate meanings, validate emotions, and construct narratives about who they are (Ma et al., 2026). For example, AI-scaffolded journaling offers opportunities for users to reflect on their lives from multiple perspectives, but early evidence suggests users also tend to over-rely on these insights over their own self-generated emotional expressions (T. Kim et al., 2024). Thus, AI could become a substitute for reflective processes that are central to human self-awareness and psychological development. Systems designed to affirm users may shape users' self-concepts, perceptions, and social-emotional judgements in an amplification feedback loop that promotes bias or delusional thinking, even when humans are unaware (Glickman & Sharot, 2025).

Humans are also motivated to confirm their own pre-existing beliefs. The sycophancy, personalization, confident tone, and ease of using AI for information processing can lead to extreme attitudes and overconfidence (Rathje & Van Bavel, 2026). These interactions can shape behavior, relationships, personality, and self-awareness in ways that are not yet well understood, likely subtle, and yet may be harmful in the long term.

I'm concerned about disempowerment from relying on things that make us worse as people—emotionally relying on AI systems and then not developing those same kind of muscles ourselves.

— JARED MOORE

KEY TAKEAWAYS

AI and Intrapersonal Processes

PSYCHOLOGICAL IMPACTS

AI offers scaffolded self-exploration, yet this may disrupt an individual's personal or social engagement in identity exploration/ self-reflection or reinforce one's existing beliefs while reducing critical examination. AI use may be linked with distress intolerance, in which individuals increasingly struggle with the complexities of daily life. AI tools designed intentionally with clinicians (e.g., that prompt human interaction or cognitive reappraisal) could support emotion regulation skills.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Identify and assess long-term and causal impacts of using generative AI on distress tolerance, identity formation, self-concept, self-reflection, and other intrapersonal processes. Disseminate information to empower AI users to make informed and values-consistent decisions with appropriate self-guided reflections.

5

Overreliance and Dependency

'Addictive-like' Use

Technology has long intersected with humans' capacity to develop **habits, dependencies, and behavioral reliance**—sometimes referred to as behavioral addictions, with diverse viewpoints on when such a term is appropriate. Behavioral addictions refer to a persistent pattern of impaired control over a behavior, continued engagement despite adverse consequences, and clinically meaningful distress or functional impairment. While many rewarding behaviors can be hard-to-stop, only two are recognized as diagnosable behavioral addictions by the 11th edition of the International Classification of Diseases (ICD-11): gambling disorder and internet gaming disorder. Social media and smartphone use are commonly discussed in the context of when a maladaptive habit is conceptualized as an addictive behavior. The terms “compulsive use” or “problematic use” are often used to describe excessive or difficult-to-control engagement with technology associated with meaningful impairment, without presupposing that the behavior constitutes an addiction.

Compulsive, problematic, or excessive AI use is linked to the risks outlined elsewhere in this document, including cognitive outsourcing, emotional intimacy with AI, and severe mental health implications. Indeed, overreliance or dependency likely does not exist without some of these other features (e.g., cognitive, emotional, or pseudo-social reliance; Dagnino et al., 2026; Laestadius et al., 2024; Milton et al., 2026).

Large-scale experimental research has found that people who use chatbots more—regardless of various design aspects related to more relational vs. transactional conversations—show consistently **worse psychosocial outcomes**, especially among users who trust the chatbot more (C. M. Fang et al., 2025). Users with lower self-esteem, more social anxiety, escapism tendencies, and overall better experiences with chatbots are more likely to experience more problematic AI use (Yao et al., 2025), though these

features may suggest selection effects, wherein these individuals seek out more chatbot use.

Companies developing and deploying AI chatbots, like social media companies, **benefit and profit when users engage with systems more and for longer**. Features to increase engagement include follow-up responses wherein the chatbot offers a new or alternate solution (Shroff, 2025) and dark design patterns that suggest chatbots have needs that a user must meet (Laestadius et al., 2024). AI systems can reduce overreliance by dialing back design features that reward continued engagement, including unconditional validation (sycophancy), first-person “relational” framing that invites anthropomorphism, and 24/7 always-available emotional responsiveness. However, because people who use chatbots more (regardless of design variations) have worse outcomes (Fang et al., 2025), design choices alone will not mitigate habitual chatbot use and related harms; design choices can thus be coupled with other supports for at-risk users (media literacy, offline interventions, structural investments in social life or clinical care). Systems could also be designed to automatically detect compulsive or excessive use and could personalize system interventions, such as escalating nudges to seek human support. Other work has tested “cognitive forcing” functions that compel users to engage critically with AI-generated output; although users dislike this added friction, experimental studies show that these design choices reduce overreliance (Buçinca et al., 2021).

KEY TAKEAWAYS

AI and Overreliance and Dependency

PSYCHOLOGICAL IMPACTS

Overreliance often signals excessive use of AI in one of the other domains (e.g., for mental health, companionship, cognitive outsourcing). Individuals who use AI chatbots at the highest levels report poorer interpersonal and emotional outcomes.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Determine whether links between high levels of chatbot uses and poor psychosocial outcomes are causal. Identify the prevalence, comorbidities, and clinical course of emerging behavioral addictions associated with AI. Describe and disseminate potential design solutions to increase detection of individuals experiencing an AI-related behavioral addiction, intervene (e.g., platform-based nudges), and inform the public of dependency risks.

6

Epistemology and Information

AI is shaping the way humans interact with information, including text and image-based content, affording access to more and personalized information. This access may have benefits in some contexts (e.g., benefit learning goals), yet with negative implications for how we navigate **truth, knowledge, and trust** in a world of information synthesis and photo-realistic fake images.

AI tools are increasingly being used in place of internet searches where users previously had to adjudicate between various sources. When people ask an AI for an answer instead of comparing several sources, the resulting **“truth” tends to be narrower** and homogenized (Y. Yu et al., 2026). This flattening also yields less creative output, which may have more collective social impacts but less clear individual-level harm; that is, the collective range of perspectives may shrink without individual users noticing (Doshi & Hauser, 2024; E. Wenger & Kenett, 2026). Researchers have begun testing interventions that warn about biases in AI output (e.g., sycophancy that reinforces users’ statements). Although interventions can reduce peoples’ perceptions of AI as objective and trustworthy, they do not reliably reduce the persuasiveness of sycophantic AI, suggesting that individual-level interventions such as warning labels or AI literacy are likely not enough to mitigate information-related AI harms (Ye, Kraut, et al., 2026).

Advancements in AI tools have made **detecting AI-generated content** difficult. Detection accuracy hovers near chance among adults (Cooke et al., 2025; Jakesch et al., 2023) and below chance in children (Langer et al., 2026), though adults who frequently use AI for writing tasks are skilled at detecting AI-generated text (Russell et al., 2025). Modifiable factors predict one’s ability to detect AI-generated content. Adults higher in intelligence tend to be more accurate, whereas heavier digital media use is associated with more incorrect attributions (assuming AI-generated content is human-generated; Chein et al., 2024; Langer et al., 2026). As AI continues to improve, it may increasingly generate

human-like text and life-like images/videos that simply cannot be detected by the human eye (Jakesch et al., 2023) but may be by a machine via watermarks embedded within the content, which are increasingly being rolled out at scale following the EU AI Act which mandates such watermarks (M. Smith, 2026).

Biases are also often ‘baked in’ to AI output, given that the human-generated content on which models are trained reflects humans’ discriminatory biases, systematic inequalities, and sociopolitical and cultural preferences and prejudices (X. Fang et al., 2024; X. Zhang et al., 2026). For example, image generation output of popular AI models tend to be biased towards men and whiteness, with the potential for systematic discrimination against users already navigating marginalization in daily life, including Black women (X. Fang et al., 2024; Heley, Hom, et al., 2026). Harms resulting from exposure to digitally mediated forms of racism (Tynes et al., 2019, 2024, 2025; Willis et al., 2026) could be mitigated by careful attention to AI training data and decision-making. At a broader structural level, unintended biases in AI tools used in hiring, housing, or judiciary decisions can perpetuate social inequalities, potentially undermining the spirit of laws regarding equity, fairness, and segregation (Gerards & Borgesius, 2025; Krištofik, 2025). Companies aiming to maximize profit may move quickly and thus deploy potentially discriminatory products without sufficient testing, incurring what some scholars have called an ‘ethical debt’ paid by the individual victims of invisible unfair algorithms (Rathore, 2026). In contrast, a data equity framework could aim to increase representation and accuracy of diverse groups in data, ensure access is transparent and equitable, and maintain vigilance over unintended consequences of bias (Stonier et al., 2023).

Compounding this, AI can sometimes be **inaccurate**. These errors, sometimes referred to by the labs as “hallucinations,” refer to fluent, coherent output that is factually inaccurate, unsupported, or fabricated. Notably, referring to machine errors as “hallucinations” has been critiqued as a way to anthropomorphize AI, imply AI sentience, invite sympathy for AI models, and potentially hold back appropriate regulation and corporate responsibility (Tanner, 2026).

Accuracy benchmarks suggest AI errors, though improving, are still relatively common; errors occur in 4–5% of document summaries (when text is provided directly to the model) but in 22–94% of responses (not including when the model claims ignorance and refuses to attempt to respond) to open-ended, difficult factual questions (Sajadieh et al., 2026). Users tend to trust LLM outputs without critically assessing them, accepting subtly inaccurate claims as true unless the errors are blatantly obvious, which can spread misinformation further when those outputs are shared (Steyvers et al., 2025). Technology design that

ensures chatbots express uncertainty could mitigate some of these risks (S. S. Y. Kim et al., 2024).

“Deepfakes,” or fake content of ostensibly real people or events, represent an epistemic problem for individuals and society at large. Deepfakes can cause direct harm when humans believe false information (Diel et al., 2026), including false or incriminating information about real others (S. Clark & Lewandowsky, 2026) or themselves (Nash & Wade, 2009). Deepfakes can also undermine trust in recording technology more broadly; real footage can be dismissed as fake as easily as fake footage is believed to be real (Chesney & Citron, 2019; Fallis, 2021). This points towards the possibility of an epistemic “post-truth” crisis and a collapse in the shared infrastructure for adjudicating truth, such that science, institutions, experts, journalists, and even one’s own senses become suspect by default (Heley, Laestadius, et al., 2026).

Non-consensual sexual deepfakes (“deepnudes”) represent a related harm centered on bodily autonomy, human likeness, and abuse. Deepnudes tend to depict women or children (Center for Countering Digital hate, 2026). According to a UNICEF survey, 1.2 million children worldwide are aware of a sexual deepfake created depicting them (United Nations Children’s Fund, 2026), and 1 in 10 minors know of a peer who has made this content of other minors (Thorn, 2024). Deepfakes and deepnudes can be used in sextortion, bullying, harassment, and violence. A nationally representative poll of educators found that 51% of U.S. teachers report that they are concerned about their students creating deepfakes featuring peers or staff (Banerji, 2024). Another report found that 13% of K–12 principals reported incidents of bullying related to deepfakes between 2023 and 2025 (Jackson et al., 2025) and media reports detail students using AI to bully peers (Baker, 2026; Hollingsworth, 2026). Across adult respondents from Australia, the UK, and the US, 18% reported intentionally viewing sexual deepfakes, while 3.2% reported creating, distributing, or threatening to distribute sexual deepfakes (Umbach et al., 2026) and in one study of US adults, 1.13% had been victimized by deepfake pornography (Ruvalcaba et al., 2026).

Will people start listening to their AI more and start distrusting the people around them, distrusting institutions?

— DESMOND ONG

Humans have all sorts of biases. When AI models are trained on data that humans generated, they sometimes amplify those biases. Then, when humans interact with a system carrying the amplified bias, they internalize it.

— MOSHE GLICKMAN

KEY TAKEAWAYS

AI and Epistemology and Information

PSYCHOLOGICAL IMPACTS

AI introduces significant opportunities to generate and be exposed to novel or personalized information. However, some of this is fake, biased, or intentionally harmful content (e.g., misinformation, deepfakes). AI can provide erroneous information with high fluency, which could increase user trust in false information or erode trust in institutions. AI-generated information tends to be narrower and more homogenous than human-created content.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Identify the effect of AI-generated misinformation and biases on social-political institutions and public life. Quantify the degree of homogeneity in AI-generated information and content and measure the effects on human outcomes. For institutions or platforms, evaluate the effectiveness of strategies to detect, flag, or remove false content. Mandate design features (e.g., watermarks) to prevent harmful misinformation. Promote user knowledge and critical engagement with digital content.

7

Human Development

Implications for Children and Adolescents

AI presents some unique risks and opportunities for children and adolescents for several reasons. First, brain development during early childhood and at the adolescent transition likely creates vulnerability among youth who seek sycophantic reinforcement from a human-like AI chatbot, as the neural reorganization during these periods sensitizes socially salient stimuli. Yet, children and adolescents also may lack the impulse-control to use AI in moderation; in other words, youth may be especially drawn towards AI, for longer interactions, and repeatedly over time. Second, sensitivities during development can **heighten vulnerability** to harm, while limited experience and still-developing skills and knowledge may shape when AI is helpful or harmful in ways that differ from adults. For example, whereas adults may experience skill atrophy if they outsource tasks to AI instead of practicing certain cognitive or creative skills, children may miss the opportunity to learn the skill during sensitive developmental periods of brain plasticity, potentially disrupting longer-term capacities. Third, substantial research documents the role of human interactions (e.g., scaffolding, soothing touch) in early childhood for cognitive, brain, and even immunity development, as well as in older ages where emotional support, companionship, social comparison serve critical functions for moral, cognitive, social, identity, and intellectual development at older ages (Prinstein, 2025). Children and adolescents need experiences with offline activities and interacting with other humans to learn and grow to meet needs and expectations within current society. AI may displace important developmental opportunities to engage in these experiences when neurobiological plasticity is greatest. Thus, the risks outlined in the prior sections of this document may be especially consequential earlier in development, with psychosocial harms potentially extending into adulthood, though long-term longitudinal research does not yet exist.

AI toys are increasingly being deployed to the consumer market, marketed towards infants and toddlers. Some speculation suggests that they can aid cognitive and social-emotional development when designed appropriately, yet risks are notable and extremely limited empirical research addresses their impact prior to deployment (Xiao & Gonçalves, 2025). AI-based toys raise concerns given that young children require explicit distinctions between fact and fantasy, tend to report more trust of artificial agents and see them as social partners, and AI toys can generate misinformation, collect private data from young users, and displace human social interaction (Leaver et al., 2026). AI toys do not appear to offer individualized feedback and instruction in the manner that has been noted as most critical for effective cognitive (i.e., intellectual, language) development. AI toys also are not equipped to deal with disclosure that young users may offer that requires immediate attention or child safety protection.

Children and teens (like adults, but perhaps with greater consequence) may be vulnerable to AI **content exposure or misunderstanding** AI's capabilities. Children may not be able to reliably understand the nature of AI, determine when to trust AI-generated content, nor identify content as AI-generated (Girouard-Hallam et al., 2021; Girouard-Hallam & Danovitch, 2022; Langer et al., 2026). Children and adolescents may also be uniquely vulnerable to exposure to misinformation, content that promotes self-harm, or sexualized content (Regehr et al., 2026). AI-generated child sexual abuse material (CSAM) and tools that create CSAM are a growing phenomenon appearing on mainstream platforms, and are increasingly extreme, life-like, and gendered, disproportionately showing images of girls (Internet Watch Foundation, 2026).

Teens, who are in the midst of a sensitive period with regard to development of social and emotional capacities, may be vulnerable to developing attachments to AI companions (Namvarpour et al., 2026; Sun et al., 2026), outsourcing cognitive functions (e.g., while completing homework; Klarin et al., 2024) and developing overreliance on or excessive use of AI (Maheux, Akre-Bhide, et al., 2026; O'Neil et al., 2026). Adolescents are uniquely sensitive to rejection, social information processing, and reward; design choices that make AI highly affirming and sycophantic, constantly-available, gamified, and/or representative of a 'super-peer' that can influence and normalize behaviors may be especially attractive to and influential for adolescents and come with acute risks of harms for development during this phase (P. Kim et al., 2025; Maheux & Maes, 2026; Y. Yu et al., 2025).

During adolescence, teens also engage in **identity and sexual exploration** as they navigate the process of transitioning into adulthood; these processes are sensitive and should be carried out through self-reflection and strong social relationships. AI chatbots may offer

spaces to explore such processes, but can undermine some beneficial aspects of such exploration by displacing time for independent identity exploration, promoting harmful or maladaptive norms about human romantic and sexual relationships, or exposing youth to age-inappropriate content (Hinduja & Patchin, 2026; Leuenberger, 2026; Maheux, Boeldt, et al., 2026; Namvarpour et al., 2025).

Design standards and AI chatbot features should be **developmentally appropriate** and effectively age-gated to avoid unintended harms to underage users, and seek to promote positive youth development rather than merely avoid acute harms (Telzer et al., 2026; Y. Yu et al., 2025). Note however that scholars are unlikely to universally agree on a single age at which certain digital exposures are appropriate, in part because age and social-emotional maturity are distinct. Children and adolescents can build knowledge and skills to understand what AI is, how to engage with it or avoid it for the promotion of learning and health, and develop critical literacy related to technology engagement generally (Veldhuis et al., 2025). When designed intentionally and used with adult supervision (e.g., designs that require users to engage cognition via ‘tutor mode’ or similar features; Bastani et al., 2025; Buçinca et al., 2021), AI may be able to scaffold learning. Design choices that may be harmful in most situations and are particularly compelling to child users—first-person, validating, affiliative language (Kim et al., 2025) should also be replaced with a neutral, factual tone that avoids emotionality or dependence-forming design features (The Jed Foundation, 2025). Scholars insist that AI cannot and should not be designed to replace the human relationships that are at the center of caregiving and teaching relationships (American Psychological Association, 2025). Rigorous research is needed to conclude when, how, and for whom AI should or should not be integrated into child development, with an emphasis on promoting youth well-being instead of merely avoiding harms.

Spending a lot of time with AI chatbots might potentially limit those abilities or opportunities to practice all the range of the skills and capacities that they could [develop] if they grow up mostly with other humans—peers and friends and teachers.

— PILYOUNG KIM

What does it mean to be in an attachment economy as a child, where you're now interacting in spaces and places where attaching you to the entity is the goal, as a way of monetizing you and your data?

— AMANDA LENHART

If a teenager is using AI for companionship, will that deskill how that teenager is processing emotions and dealing with relationships in real life? If it is not deskilling, what specifically would the influence be?

— YAMAN YU

KEY TAKEAWAYS

AI and Human Development

PSYCHOLOGICAL IMPACTS

Youth may be more vulnerable to all AI harms given developmental sensitivities and plastic windows of development, including age-inappropriate exposure (e.g., sexual or violent content, role-play content, 'deepnude' victimization). Opportunities to use intentionally designed AI for learning and exploration may offer benefits when paired with support from a caregiver or teacher.

RESEARCH, POLICY, AND ADVOCACY OPPORTUNITIES

Identify and catalogue age-inappropriate features of AI related to companionship, mental health, and learning. Test long-term developmental outcomes of child and adolescent AI uses and design choices that may mitigate or exacerbate harm (e.g., anthropomorphic characters, re-introducing friction into AI interactions). Develop and deploy design and regulatory guardrails to avoid harms and promote developmental thriving, including effective age assurance.

What Now? The AI Psychological Research Coalition & the Future of Research and Action

We've described key findings across seven areas of psychological functioning that intersect with AI use. The amount known pales in comparison to the amount unknown in each domain. Still, each offers some potential insights into ideal next steps for research, policy, technology design, and resources for citizens.

The goal of the AIPRC Conference is to convene a network interested in these domains of psychological functioning to accelerate research and action. This includes catalyzing conversations and accelerating coordination across research, policy, and public communication to protect and promote human thriving in the age of AI.

In the interest of accelerating the field toward these goals, below we offer some potential aims and agenda items for the coalition on four main topics: coordination infrastructure, research, regulation and design, and individual and community resources. These aims are preliminary and require discussion and input from coalition members. Such aims will only be realized with a sufficiently large and mobilized coalition with a shared goal of promoting human well-being in the age of AI.

Coordination Within and Across Sectors

AI growth is accelerating, bolstered by unprecedented capital investments. To generate timely, high-quality research and meaningful impacts on policy outcomes, there is a pressing need for individuals and organizations within and across sectors to work together. Opportunities include:

- Identifying avenues to build and sustain collaboration between disparate research teams, leveraging shared resources, rather than working in isolation.
- Creating pipelines for researchers to share findings rapidly with policymakers and the public via meaningful coordination with advocates, policymakers, funders, and media.
- Coordinating researchers, policy experts, and other stakeholders in service of translating public concerns and research evidence into actionable and concrete policy recommendations.

- Building capacity (e.g., trainings, investments) among organizations and individuals to understand diverse sectors, frameworks, methods, and incentive structures to stimulate cross-sector communication and action (Oduro et al., 2025); for example, training researchers on best practices for developing policy-relevant studies and science communication.
- Assembling shared infrastructure, including spaces (convenings, online forums) for ongoing communication across disciplines and sectors, transparent standards for research and safety, and pooled resources for research (e.g., open-source data collection tools, multi-site studies, data and material sharing).
- Funding rapid-response, flexible grants to support high-risk team science, trainees, resource-intensive methods (e.g., community engaged research, neuroscience research, longitudinal methods), and opportunities for researchers to analyze data from technology companies with appropriate safeguards for independence and participant protections. Investments in AI far surpass any other types of technological advances and our research engine continues at the pace of the past, or potentially slower given significant budget cuts to social science (Farrell et al., 2026).
- Engaging directly with communities impacted by AI, including diverse sociodemographic and geographic populations, to understand concerns and preferences and promote equity in research and policy.
- Surfacing tensions related to partnerships with technology companies and developing guidelines for ensuring independence and disclosures of conflicts of interest, including tools to identify when partnerships compromise the integrity of impartial research or recommendations for forming effective partnerships.

Because researchers don't always know how policy works, they often don't know what kinds of research questions or methodologies to pursue... Sometimes just tweaking a research design or adding an additional question or variable would help the research directly inform design and policy decisions. I think it would be powerful to educate researchers on that.

— NATE FAST

I'm very excited about anything that puts researchers and non-researchers together in the same space...There's a small group of us in psychology, a small group of us in education, a small group of us in HCI—but we don't really cross-pollinate across disciplines.

— LAUREN GIROUARD

Given the sheer scale of the problem... I think we need to work together across different labs, across different institutions to share data and share resources and expertise.

— DESMOND ONG

Research: Future Directions and Solutions

Researchers must continue to do high-quality science yet balance scientific goals with the need for a rapid, coordinated response to the challenges of AI. This may include:

- Identifying a shared research agenda with respect to key psychological domains (e.g., the seven outlined in this report) and gold-standard methodologies (e.g., randomized controlled trials, quasi-experiments, longitudinal studies, neuroimaging studies), and coordinate swift study execution.
- Engaging in more team science to share hypotheses or pilot data, include novel measures in ongoing studies, and streamline the process of high-quality multi-site data collection.
- Sharing data, tools, measures, methods, and resources when possible and developing interdisciplinary research partnerships within and outside of academia (e.g., non-profits, tech companies) with robust data-sharing protections for both research independence and participant privacy protection.

- Developing and testing individual-level, community-level, and clinical interventions to mitigate harms and maximize benefits of AI for human well-being.
- Partnering directly with those affected by AI (e.g., teachers, parents, teens, clinicians, citizens) to improve research design and dissemination, including promoting research on and with underserved and vulnerable populations to understand the way novel technologies are introduced into their communities, their lived experiences with generative AI, and the range of use cases they adopt.
- Publishing research with rapid peer-review cycles and prioritizing urgent, high-impact work, including well-designed preliminary studies that are faster and less resource-intensive to conduct while more rigorous work is underway (Christakis, 2026).

These technologies are not static. They're going to change so fast. They're going to change in visible ways and they're going to change in invisible ways. You have to figure out how, as a researcher, to hold that in the studies you're doing.

— EMILY WEINSTEIN

I think what researchers could do is focus more on the patterns that may survive the model changes. Our theories need to last longer than the model version used in the study.

— RUO-NING LI

Anything that can be done as a community to establish standards or frameworks for data donation that are both practical and privacy-preserving will empower researchers to study these risks.

— HAMILTON MORRIN

But at a minimum as scientists and a society we have to be more nimble and cautious. In particular, we have to be quicker to achieve consensus specifically around the idea that as a first principle we should do no harm. No child is worse off for not having access to AI during elementary school, but plenty of them could be worse off for having it.

— DIMITRI CHRISTAKIS

Regulation and Design

Regulation is a powerful societal lever to shape how AI technologies impact population well-being, including mandating how AI chatbots are designed. Many open questions remain regarding regulating and designing AI systems to maximize benefits and mitigate harms, requiring more work from researchers, advocates, policymakers, and others, including:

- Determining optimal design principles for safe and humane technologies through interdisciplinary research and developing policy-ready frameworks for enacting these design features into popular technologies.
- Cataloguing and advocating for the restriction of harmful design features, including potentially: engagement-driving features, anthropomorphic responses, sycophancy, personalization/customization, and gamification.

- Establishing adequate third-party safety testing practices, organizations, and frameworks sufficient to ensure consumer protections that involve social-behavioral and clinical expertise (Eriksson et al., 2025; Weidinger et al., 2023).
- Identifying optimal strategy for public communication and messaging about legislation (e.g., ‘bans’ vs. ‘recalls’ that borrow public health framing from other products/substances; Bernstein, 2026).
- Developing, testing, and lobbying for developmentally appropriate age standards and suitable age-gating systems.
- Developing, testing, and lobbying for clinically-informed solutions for vulnerable users, including specific design standards for users experiencing crisis.
- Conducting policy research on the efficacy of and public response to litigation and regulation across jurisdictions for other technologies (e.g., social media age bans, the recent Meta settlement) and maintaining ongoing population-level evaluations to address AI-related changes following legislation.

The idea that we could just put a few simple boundaries around the big frontier lab companies and that that's going to solve this feels like trying to put out a forest fire with a garden hose.

— EMILY WEINSTEIN

Frontier labs and researchers are developing a lot of synthetic scenarios based on harms they've become aware of, which is useful. But that missing link—which very few have managed to bridge—is that we need evals that are informed by real-world harm data, and we need ways of gathering and curating that data so researchers globally can access and work with it.

— HAMILTON MORRIN

On the one hand, I do think we need fast science. I think we need to come up with tools and methodologies that help us to answer questions more quickly and that don't wait for the peer-review process the way it's currently configured to play out. It's just way too slow for an industry that has the technology that we're embedding into our lives so readily. On the other hand, I think we often act too early to institute policies to drum up fear in society about things, and the problem there is it's a little bit like the boy who cried wolf.

— JASON CHEIN

Individual and Community Resources: Interventions and Partnerships

AI-related changes in society are a collective action problem and require societal-level changes to address comprehensively. That said, proximal interventions and resources for AI users and institutions can promote resilience, knowledge, and agency in the age of AI. Researchers and other professionals can consider:

- Engaging local communities with research dissemination, including sharing infographics, giving talks to families or schools, engaging with local media, and sharing guidance with local clinicians (Palaniyappan et al., 2026).
- Implementing interventions to promote agency and comprehensive knowledge related to AI technologies (e.g., source evaluation, media skepticism, knowledge of sociopolitical and environmental incentives and consequences of AI development, and values-aligned reflections on one's goals for using [or not using] AI in health-promotive ways; Biagini, 2025; Veldhuis et al., 2025).
- Articulating a shared vision for a positive future based on philosophical, empirical, and humanistic perspectives on human thriving; based on this vision, identify the optimal contexts and design in which AI fits this vision and where it should be avoided.

- Investing in environmental changes that address the underlying social conditions that increase individual risk for problematic AI use. For example, work toward the revival and creation of third-spaces and other contexts in which in-person interactions can occur as a regular component of life. This includes consideration of accessibility to ensure that those with disabilities can be full participants.
- Strengthening the social conditions that foster community, hope for the future, values that prioritize human flourishing over market incentives, economic stability, and opportunities for meaningful action.

Parents and students really want to have more information, more support. I feel like we're critically lagging behind.

— PILYOUNG KIM

We can use our research and our evidence to shape policy, shape legislation, and also shape movements and efforts to hold tech companies accountable.

— HENRY WILLIS

A movement to check unbridled AI development could end up empowering grassroots communities. There's a lot of issues, like climate change, that have become politicized and people lose the will to deliberate and get involved. AI is different—people across the political spectrum are very concerned, and it's unifying some interesting voices... AI could be the impetus that brings us together to say 'wait a minute, we need the grassroots to have a voice and have some power in this conversation.'

— NATE FAST

Conclusion

The conference is designed to provide space for discussion and consensus-building. **This is *not a normal conference***—it is intended to be a flipped conference and working space, and an opportunity to seed collaborative research and coordination across sectors.

We critically need more research on how AI intersects with human psychology, including across the seven domains identified in this review, and action across policy, funders, advocacy, and practitioners. The field would benefit from a shared consensus on the highest priority research questions and approaches and a defined infrastructure to translate science into action. Moreover, given the need to move quickly, our shared understanding of these issues will benefit when researchers ‘divide-and-conquer’—avoiding duplication and ensuring research teams are producing complementary insights for our collective goals. This can be realized with more cross-talk among researchers, as well as more communication and coordination across sectors, including with activists, advocates, policymakers.

References

- American Psychological Association. (2025). *Artificial intelligence and adolescent well-being: An APA health advisory*. <https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being.pdf>
- Augustin, M., Pollak, T. A., & Morrin, H. (2026). Characterizing the spiral: Potential mechanisms in AI-associated delusions. *NPP—Digital Psychiatry and Neuroscience*, 4(1), 14. <https://doi.org/10.1038/s44277-026-00065-0>
- Axelsson, M., & Shevlin, H. (2026). *Disambiguating Anthropomorphism and Anthropomimesis in Human-Robot Interaction* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2602.09287>
- Babu, J., Joseph, D., Kumar, R. M., Alexander, E., Sasi, R., & Joseph, J. (2025). Emotional AI and the rise of pseudo-intimacy: Are we trading authenticity for algorithmic affection? *Frontiers in Psychology*, 16, 1679324. <https://doi.org/10.3389/fpsyg.2025.1679324>
- Baker, R. (2026, January 15). *Children are making AI images to “bully or ridicule” their peers*. <https://www.bbc.com/news/articles/cly52xxew3no>
- Banerji, O. (2024). *Why Schools Need to Wake Up to the Threat of AI ‘Deepfakes’ and Bullying*. *EducationWeek*. <https://www.edweek.org/technology/why-schools-need-to-wake-up-to-the-threat-of-ai-deepfakes-and-bullying/2024/12>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bernstein, G. (2026, May 22). From bans to recalls: A public health framework for AI companion bots. *Brookings*. <https://www.brookings.edu/articles/from-bans-to-recalls-a-public-health-framework-for-ai-companion-bots/>
- Biagini, G. (2025). Towards an AI-Literate Future: A Systematic Literature Review Exploring Education, Ethics, and Applications. *International Journal of Artificial Intelligence in Education*, 35(4), 2616–2666. <https://doi.org/10.1007/s40593-025-00466-w>
- Bijl, R. V., de Graaf, R., Hiripi, E., Kessler, R. C., Kohn, R., Offord, D. R., Ustun, T. B., Vicente, B., Vollebergh, W. A. M., Walters, E. E., & Wittchen, H.-U. (2003). The Prevalence Of Treated And Untreated Mental Disorders In Five Countries. *Health Affairs*, 22(3), 122–133. <https://doi.org/10.1377/hlthaff.22.3.122>
- Boyd, R. L., & Markowitz, D. M. (2026). Artificial Intelligence and the Psychology of Human Connection. *Perspectives on Psychological Science*, 21(2), 192–220. <https://doi.org/10.1177/17456916251404394>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 188:1-188:21. <https://doi.org/10.1145/3449287>
- Buck, B., & Maheux, A. J. (2026). Psychosis Risk and Generative Artificial Intelligence Use Frequency, Motivations, and Delusion-Like Experiences: Cross-Sectional Survey Study. *Journal of Medical Internet Research*, 28, e85038. <https://doi.org/10.2196/85038>
- Carpenter, J. (2026). Human–AI relationships as designed relationality: A sociotechnical model. *AI & SOCIETY*, 41(6), 5789–5800. <https://doi.org/10.1007/s00146-026-02908-y>
- Center for Countering Digital hate. (2026, January 22). *Grok floods X with sexualized images of women and children*. <https://counterhate.com/research/grok-floods-x-with-sexualized-images/>
- Center for Digital Thriving. (2025). *Ten fresh insights on generative AI from our teen advisors*. Harvard Graduate School of Education Center for Digital Thriving.
- Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J. B. (2026). *Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians* (arXiv:2602.19141). arXiv. <https://doi.org/10.48550/arXiv.2602.19141>

- Chatterjee, R. (2025, September 19). Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots. *NPR*. <https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide>
- Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports*, 14(1), 25989. <https://doi.org/10.1038/s41598-024-76218-y>
- Cheng, M., Lee, A. Y., Rapuano, K., Niederhoffer, K., Liebscher, A., & Hancock, J. (2026). Metaphors of AI indicate that people increasingly perceive AI as warm and human-like. *Communications Psychology*, 4(1), 8. <https://doi.org/10.1038/s44271-025-00376-6>
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. <https://doi.org/10.1126/science.aec8352>
- Cheng, M., Sieh, I., Zope, H., Yu, S., Ibrahim, L., Arora, A., Moore, J., Ong, D., Jurafsky, D., & Yang, D. (2026). *Verbalizing LLMs' assumptions to explain and control sycophancy* (arXiv:2604.03058). arXiv. <https://doi.org/10.48550/arXiv.2604.03058>
- Chesney, R., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6), 1753.
- Choudhury, M., Elyoseph, Z., Fast, N. J., Ong, D. C., Nsoesie, E. O., & Pavlick, E. (2025). The promise and pitfalls of generative AI. *Nature Reviews Psychology*, 4(2), 75–80. <https://doi.org/10.1038/s44159-024-00402-0>
- Christakis, D. A. (2026). A Call for Expedited Research on AI Chatbots. *JAMA Pediatrics*, 180(7), 722. <https://doi.org/10.1001/jamapediatrics.2026.1376>
- Chung, V.-H.-A., Bernier, P., & Hudon, A. (2026). Mass Media Narratives of Psychiatric Adverse Events Associated With Generative AI Chatbots: Rapid Scoping Review. *JMIR Mental Health*, 13(1), e93040. <https://doi.org/10.2196/93040>
- Clark, A. (2025). The Ability of AI Therapy Bots to Set Limits With Distressed Adolescents: Simulation-Based Comparison Study. *JMIR Mental Health*, 12, e78414–e78414. <https://doi.org/10.2196/78414>
- Clark, S., & Lewandowsky, S. (2026). The continued influence of AI-generated deepfake videos despite transparency warnings. *Communications Psychology*, 4(1), 13. <https://doi.org/10.1038/s44271-025-00381-9>
- Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2025). As Good as a Coin Toss: Human Detection of AI-Generated Content. *Communications of the ACM*, 68(10), 100–109. <https://doi.org/10.1145/3729417>
- Dagnino, F. M., Fante, C., Guerrieri, V., & Passarelli, M. (2026). Addicted, attached, or just delegating? A scoping review on “problematic artificial intelligence use.” *Frontiers in Psychology*, 17. <https://doi.org/10.3389/fpsyg.2026.1900953>
- De Freitas, J., Uguralp, A. K., Uguralp, Z. O., & Stefano, P. (2026). *AI Companions Reduce Loneliness*. 52(6), 1126–1148. <https://doi.org/https://doi.org/10.1093/jcr/ucaf040>
- Diel, A., Lalgi, T., Mellis, F. S., Teufel, M., & Bäuerle, A. (2026). The harm of deepfakes: A scoping review of deepfakes' negative effects on human mind and behavior. *AI & SOCIETY*, 41(6), 5971–5987. <https://doi.org/10.1007/s00146-025-02774-0>
- Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3), 336–345. <https://doi.org/10.1038/s44220-026-00595-8>
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- Duong, C. D., Vu, T. N., Ngo, T. V. N., Do, N. D., & Tran, N. M. (2025). Reduced student life satisfaction and academic performance: Unraveling the dark side of ChatGPT in the higher education context. *International Journal of Human-Computer Interaction*, 41(8), 4948–4963. <https://doi.org/10.1080/10447318.2024.2356361>
- Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. (2025). Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1), 850–864. <https://doi.org/10.1609/aies.v8i1.36595>
- Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 34(4), 623–643. <https://doi.org/10.1007/s13347-020-00419-2>

- Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study*. <https://arxiv.org/abs/2503.17473>
- Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: An examination of news produced by large language models. *Scientific Reports*, 14(1), 5224. <https://doi.org/10.1038/s41598-024-55686-2>
- Farrell, H. J., Gopnik, A., & Evans, J. (2026, September 6). Opinion | The Research We Need to Understand A.I. Is Falling Apart. *The New York Times*. <https://www.nytimes.com/2026/09/06/opinion/ai-social-sciences.html>
- Folk, D., & Dunn, E. (2026). How Does Turning to AI for Companionship Predict Loneliness and Vice Versa? *Psychological Science*, 37(4), 276–286. <https://doi.org/10.1177/09567976261427747>
- Folk, D., Heine, S. J., & Dunn, E. (2025). Individual differences in anthropomorphism help explain social connection to AI companions. *Scientific Reports*, 15(1), 36548. <https://doi.org/10.1038/s41598-025-19212-2>
- Folk, D., Wu, C., & Heine, S. J. (2025). Cultural Variation in Attitudes Toward Social Chatbots. *Journal of Cross-Cultural Psychology*, 56(3), 219–239. <https://doi.org/10.1177/00220221251317950>
- 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>
- Geiselmann, R., Tsourgianni, A., Deroy, O., & Harris, L. T. (2023). Interacting with agents without a mind: The case for artificial agents. *Current Opinion in Behavioral Sciences*, 51, 101282. <https://doi.org/10.1016/j.cobeha.2023.101282>
- Gerards, J., & Borgesius, F. Z. (2025). *Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2509.08837>
- Girouard-Hallam, L. N., & Danovitch, J. H. (2022). Children's trust in and learning from voice assistants. *Developmental Psychology*, 58(4), 646–661. <https://doi.org/10.1037/dev0001318>
- Girouard-Hallam, L. N., Streble, H. M., & Danovitch, J. H. (2021). Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, 3(5), 1118–1131. <https://doi.org/10.1002/hbe2.321>
- Glickman, M., & Sharot, T. (2024). AI-induced hyper-learning in humans. *Current Opinion in Psychology*, 60, 101900. <https://doi.org/10.1016/j.copsyc.2024.101900>
- Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- Golden, A., & Aboujaoude, E. (2026). A transdiagnostic model for how general purpose AI chatbots can perpetuate OCD and anxiety disorders. *Npj Digital Medicine*, 9(1), 343. <https://doi.org/10.1038/s41746-026-02531-7>
- Goldstein, D. (2026, September 3). Los Angeles District Bans Most A.I. for Students. *The New York Times*. <https://www.nytimes.com/2026/09/03/us/lausd-schools-ban-ai-artificial-intelligence.html>
- Haag, M. (2026, September 1). New York City Bans A.I. in Elementary and Middle Schools. *The New York Times*. <https://www.nytimes.com/2026/09/01/nyregion/ai-ban-schools-nyc.html>
- Harris, L. T. (2024). The Neuroscience of Human and Artificial Intelligence Presence. *Annual Review of Psychology*, 75, 433–466. <https://doi.org/10.1146/annurev-psych-013123-123421>
- Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*, 2(4), A0a2400802. <https://doi.org/10.1056/A0a2400802>
- Heley, K., Hom, J. K., & Laestadius, L. (2026). AI-Generated Images of Substance Use and Recovery: Mixed Methods Case Study. *Jmir Ai*, 5, e81977. <https://doi.org/10.2196/81977>
- Heley, K., Laestadius, L. I., Hom, J. K., & King, A. J. (2026). Navigating a Synthetic Visual Health Information Environment: Implications of Photorealistic Generative Artificial Intelligence for Public Health. *American Journal of Public Health*, 116(10), 1539–1549. <https://doi.org/10.2105/AJPH.2026.308675>
- Hinduja, S., & Patchin, J. W. (2026). Risks and Harms of Conversational Artificial Intelligence (CAI) Chatbot Use Among US Youth. *Journal of Adolescence*, 98(5), 1597–1606. <https://doi.org/10.1002/jad.70164>

- Hollingsworth, H. (2026, January 2). *AP report: Rise of deepfake cyberbullying poses a growing problem for schools*. <https://www.pbs.org/newshour/education/ap-report-rise-of-deepfake-cyberbullying-poses-a-growing-problem-for-schools>
- Howcroft, A., Bennett-Weston, A., Khan, A., Griffiths, J., Gay, S., & Howick, J. (2025). AI chatbots versus human healthcare professionals: A systematic review and meta-analysis of empathy in patient care. *British Medical Bulletin*, 156(1), Idaf017. <https://doi.org/10.1093/bmb/Idaf017>
- Hudon, A., & Stip, E. (2025). Delusional Experiences Emerging From AI Chatbot Interactions or “AI Psychosis.” *JMIR Mental Health*, 12(1), e85799. <https://doi.org/10.2196/85799>
- Hull, T. D., Zhang, L., Arian, P. A., & Malgaroli, M. (2025). *Generative AI Purpose-built for Social and Mental Health: A Real-World Pilot* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2511.11689>
- Ibrahim, L., Hafner, F. S., Cheng, M., Lee, C., Anselmetti, R., Willer, R., Rocher, L., & Yang, D. (2026). *Sycophantic AI makes human interaction feel more effortful and less satisfying over time* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2605.07912>
- Internet Watch Foundation. (2026). *AI CSAM Report 2026: Harm Without Limits*. <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery>
- Jackson, B., Dilberti, M., & Moore, P. (2025). *Artificially Intelligent Bullies: Dealing with Deepfakes in K–12 Schools*. RAND. https://www.rand.org/pubs/research_reports/RRA3930-5.html
- Jain, S., Park, C., Viana, M., Wilson, A., & Calacci, D. (2026). Interaction Context Often Increases Sycophancy in LLMs. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–26. <https://doi.org/10.1145/3772318.3791915>
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. <https://doi.org/10.1073/pnas.2208839120>
- Jamali, L. (2025, October 27). *OpenAI shares data on ChatGPT users with suicidal thoughts, psychosis*. <https://www.bbc.com/news/articles/c5yd90g0q43o>
- Johnson, D. D., Friel, R., Laidlaw, C., Hewitt, L., Johnson, N., Stosz, C., & Schwetmann, S. (2026). *Mental Health Behavior Report*. <https://behaviors.transluce.org/mental-health>
- Kabrel, N., Stade, E. C., Wiltsey Stirman, S., Aru, J., & Eichstaedt, J. C. (2026). Conversational AI should fill the white space in mental health care, not replace humans. *Npj Digital Medicine*. <https://doi.org/10.1038/s41746-026-03088-1>
- Karman, H. (2026). *AI Companion Mortality Database: Documented Deaths Associated with Conversational AI Systems (2023–2026)* (Version 3.5.8) [Dataset]. Zenodo. <https://doi.org/10.5281/ZENODO.22428187>
- Kim, P., Xie, Y., & Yang, S. (2025). “I am here for you”: How relational conversational AI appeals to adolescents, especially those who are socially and emotionally vulnerable (arXiv:2512.15117). arXiv. <https://doi.org/10.48550/arXiv.2512.15117>
- Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’24*, 822–835. <https://doi.org/10.1145/3630106.3658941>
- Kim, T., Shin, D., Kim, Y.-H., & Hong, H. (2024). DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI ’24*, 1–15. <https://doi.org/10.1145/3613904.3642693>
- Klarin, J., Hoff, E., Larsson, A., & Daukantaitė, D. (2024). Adolescents’ use and perceived usefulness of generative AI for schoolwork: Exploring their relationships with executive functioning and academic achievement. *Frontiers in Artificial Intelligence*, 7, 1415782. <https://doi.org/10.3389/frai.2024.1415782>
- Kleinert, T., Waldschütz, M., Blau, J., Heinrichs, M., & Schiller, B. (2026). AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human. *Communications Psychology*, 4(1), 23. <https://doi.org/10.1038/s44271-025-00391-7>
- Kosmyrna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitsky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (arXiv:2506.08872). arXiv. <https://doi.org/10.48550/arXiv.2506.08872>

- Krištofik, A. (2025). Bias in AI (Supported) Decision Making: Old Problems, New Technologies. *International Journal for Court Administration*, 16(1), 3. <https://doi.org/10.36745/ijca.598>
- Laestadius, L., Bishop, A., Gonzalez, M., Illencik, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26, 5923–5941. <https://doi.org/10.1177/14614448221142007>
- Langer, A., Martinez, S., Marshall, P. J., & Chein, J. (2026). Children's susceptibility to content generated by artificial intelligence. *Technology in Society*, 86, 103303. <https://doi.org/10.1016/j.techsoc.2026.103303>
- Leaver, T., Srdarov, S., Langton, K., & Corser, K. (2026). *Generative AI in children's lives: Issues in the use of AI toys, companions, chatbots and devices* [Report]. ARC Centre of Excellence for the Digital Child. (Australia). <https://apo.org.au/node/334548>
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. <https://doi.org/10.1145/3706598.3713778>
- Leuenberger, M. (2026). Who Am I When You're a Bot? Relational Identity and AI Companions. *Journal of Applied Philosophy*, 43(3), 816–832. <https://doi.org/10.1002/japp.70094>
- Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digital Medicine*, 6(1), 236. <https://doi.org/10.1038/s41746-023-00979-5>
- Liu, G., Christian, B., Dumbalska, T., Bakker, M. A., & Dubey, R. (2026). *AI Assistance Reduces Persistence and Hurts Independent Performance* (arXiv:2604.04721). arXiv. <https://doi.org/10.48550/arXiv.2604.04721>
- Liu, X., Hu, Y., Chen, X., & Zhang, M. (2026). "Cyber Lovers": The impact of AI social chatbots on users' emotional attachment. *Current Psychology*, 45(3), 293. <https://doi.org/10.1007/s12144-025-08584-3>
- Ma, R., Niu, S., Li, L., Hirth, A., Brehm, A., & Behlerin Barbie, R. (2026). Negotiating Digital Identities with AI Companions: Motivations, Strategies, and Emotional Outcomes. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*, 1–18. <https://doi.org/10.1145/3772318.3791473>
- Madden, M., Calvin, A., Hasse, A., & Lenhart, A. (2024). *The dawn of the AI era: Teens, parents, and the adoption of generative AI at home and school*. Common Sense. https://www.common sense media.org/sites/default/files/research/report/2024-the-dawn-of-the-ai-era_final-release-for-web.pdf
- Maheshka, C., Folk, D., & Dunn, E. (2026). Harnessing Artificial Intelligence to Facilitate Difficult Conversations. *Social Psychological and Personality Science*, 19485506261444068. <https://doi.org/10.1177/19485506261444068>
- Maheux, A. J., Akre-Bhide, S., Boeldt, D., Flannery, J. E., Richardson, Z., Burnell, K., Telzer, E. H., & Kollins, S. H. (2026). Generative Artificial Intelligence Applications Use Among US Youth. *JAMA Network Open*, 9(2), e2556631. <https://doi.org/10.1001/jamanetworkopen.2025.56631>
- Maheux, A. J., Boeldt, D., Akre-Bhide, S., Flannery, J., Paige, A., Villella, G., Burnell, K., Telzer, E. H., & Kollins, S. H. (2026). How US Youth Use AI Chatbots: Conversation Patterns From Naturalistic Keystroke Observations. *Journal of Medical Internet Research*, 28, e95819. <https://doi.org/10.2196/95819>
- Maheux, A. J., & Maes, C. (2026). *Coming of Age with Generative AI: The High Stakes of Automating Adolescent Development*. PsyArXiv. https://doi.org/10.31234/osf.io/pgyfc_v1
- Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT-3-enabled chatbots. *Npj Mental Health Research*, 3(4). <https://doi.org/10.1038/s44184-023-00047-6>
- McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., & Stein, B. D. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11), e2542281.
- McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Burnett, A., Kofner, A., Breslau, J., Stein, B. D., Mehrotra, A., & Yu, H. (2025). Evaluating Alignment between Large Language Models and Expert Clinicians in Suicide Risk Assessment. *Psychiatric Services (Washington, D.C.)*, 76(11), 944–950. <https://doi.org/10.1176/appi.ps.20250086>

- Mehta, A., Moore, J., Anthis, J. R., Agnew, W., Lin, E., Yin, P., Ong, D. C., Haber, N., & Dweck, C. (2026). *The Dynamics of Delusion: Modeling Bidirectional False Belief Amplification in Human-Chatbot Dialogue* (arXiv:2604.25096; Version 1). arXiv. <https://doi.org/10.48550/arXiv.2604.25096>
- Microsoft. (2026, January 8). Global AI Adoption in 2025 – AI Economy Institute. *Microsoft Corporate Responsibility*. <https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2025/>
- Milton, A., Ajmani, L., Heger, A., Poursabzi-Sangdeh, F., Vorvoreanu, M., & Suh, J. (2026). *“Always Want to Use it for Everything”: Understanding Young Adults’ Perceptions of AI Dependence* (arXiv:2608.07592). arXiv. <https://doi.org/10.48550/arXiv.2608.07592>
- Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). Characterizing Delusional Spirals through Human-LLM Chat Logs. *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’26*, 7631–7674. <https://doi.org/10.1145/3805689.3806443>
- Moore, J., Mock, A., Mai, Y., Anthis, J. R., Louie, R., Agnew, W., Mehta, A., Klyman, K., Liang, P., Haber, N., Lin, E., & Ong, D. C. (2026). *DelusionEval: Measuring Delusion-Linked Behaviors in AI Chatbots* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2608.05004>
- Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530. [https://doi.org/10.1016/S2215-0366\(25\)00396-7](https://doi.org/10.1016/S2215-0366(25)00396-7)
- Namvarpour, M. (Matt), Brofsky, B., Medina, J. Y., Akter, M., & Razi, A. (2026). Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI ’26*, 1–19. <https://doi.org/10.1145/3772318.3790597>
- Namvarpour, M. (Matt), Pauwels, H., & Razi, A. (2025). AI-induced sexual harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), CSCW367:1–CSCW367:28. <https://doi.org/10.1145/3757548>
- Nash, R. A., & Wade, K. A. (2009). Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. *Applied Cognitive Psychology*, 23(5), 624–637. <https://doi.org/10.1002/acp.1500>
- Nozaki, Y. (2026). Why do people prefer social sharing of emotion with conversational AI over human partners? A structural topic modeling approach. *Computers in Human Behavior: Artificial Humans*, 7, 100260. <https://doi.org/10.1016/j.chbah.2026.100260>
- Oduro, S., Marwick, A. E., Johnson, C., & Meyer, E. (2025). Troubling translation: Sociotechnical research in AI policy and governance. *Internet Policy Review*, 14(4). <https://doi.org/10.14763/2025.4.2043>
- Olsen, S. G., Reinecke-Tellefsen, C. J., & Østergaard, S. D. (2026). Potentially Harmful Consequences of Artificial Intelligence (AI) Chatbot Use Among Patients With Mental Illness: Early Data From a Large Psychiatric Service System. *Acta Psychiatrica Scandinavica*, 153(4), 301–303. <https://doi.org/10.1111/acps.70068>
- O’Neil, E., Anderson, K. S., White, B., Zimmermann, L., Mann, S., & Robb, M. B. (2026). *The Common Sense Media Census: AI Use by Tweens and Teens*. Common Sense Media. <https://www.commonsensemedia.org/sites/default/files/research/report/2026-ai-use-by-tweens-and-teens.pdf>
- Ong, D. C., Goldenberg, A., Inzlicht, M., & Perry, A. (2025). *AI-Generated Empathy: Opportunities, limits, and future directions*. PsyArXiv. https://doi.org/10.31234/osf.io/8n5jw_v1
- Østergaard, S. D. (2025). Emotion contagion through interaction with generative artificial intelligence chatbots may contribute to development and maintenance of mania. *Acta Neuropsychiatrica*, 37, e79. <https://doi.org/10.1017/neu.2025.10035>
- Palaniyappan, L., Paquin, V., & Barou-Laforie, E. (2026). High-Risk Human–AI Engagement: Clinical Assessment and Management Considerations. *The Canadian Journal of Psychiatry*, 07067437261445770. <https://doi.org/10.1177/07067437261445770>

- Perry, A. (2023). AI will never convey the essence of human empathy. *Nature Human Behaviour*, 7(11), 1808–1809. <https://doi.org/10.1038/s41562-023-01675-w>
- Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15(1), 31652. <https://doi.org/10.1038/s41598-025-17242-4>
- Pierre, J. M., Gaeta, B., Raghavan, G., & Sarma, K. V. (2025). “You’re Not Crazy”: A Case of New-onset AI-associated Psychosis. *Innovations in Clinical Neuroscience*, 22(10–12), 11.
- Preda, A. (2026). AI-associated delusions: Terminology. *The Lancet Psychiatry*, 13(7), 544. [https://doi.org/10.1016/S2215-0366\(26\)00132-X](https://doi.org/10.1016/S2215-0366(26)00132-X)
- Prinstein, M. J. (2025). *Written testimony of Mitchell J. Prinstein, PhD, ABPP, chief of psychology, American Psychological Association: Examining the harm of AI chatbots*. [Written testimony].
- Rathje, S., & Van Bavel, J. J. (2026). *How AI can fuel confirmation bias*. PsyArXiv. https://doi.org/10.31234/osf.io/7a3d4_v1
- Rathje, S., Ye, M., Globig, L. K., Pillai, R. M., de Mello, V. O., & Van Bavel, J. J. (2025). *Sycophantic AI increases attitude extremity and overconfidence* (Vmyek_v1). PsyArXiv. https://doi.org/10.31234/osf.io/vmyek_v1
- Rathore, A. (2026). *Ethical Debt in AI Systems: Organizational Risk and Strategic Product Implications*. SSRN. <https://doi.org/10.2139/ssrn.6469880>
- Rauh, M., Blankvoort, D. A. H., Duran, M., Dheoráin, C. N., Pandit, H. J., Enneifer, S., Jaiswal, S. D., Ventresque, A., & Birhane, A. (2026). *Playing Games with My Heart: An Evaluation of AI Companion Apps* (arXiv:2605.08093). arXiv. <https://doi.org/10.48550/arXiv.2605.08093>
- Ravšelj, D., Keržič, D., Tomažević, N., Umek, L., Brezovar, N., & Iahad, A. (2025). Higher education students’ perceptions of ChatGPT: A global study of early reactions. *PLoS ONE*, 20(2), 0315011. <https://doi.org/10.1371/journal.pone.0315011>
- Regehr, K., Smales, K., Brailsford, J., Vrikki, P., & Lawrence, M. (2026, July 17). *Children, Young People and AI Chatbots*. https://cdn.prod.website-files.com/6778c1768fc17dadd8c1eaf6/6a58bcd2c227b6b97e25f47b_Children_Young_People_and_AI_Chatbots_White_Paper_July2026.pdf
- Rubin, M., Li, J. Z., Zimmerman, F., Ong, D. C., Goldenberg, A., & Perry, A. (2025). Comparing the value of perceived human versus AI-generated empathy. *Nature Human Behaviour*, 9(11), 2345–2359. <https://doi.org/10.1038/s41562-025-02247-w>
- Russell, J., Karpinska, M., & Iyyer, M. (2025). *People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2501.15654>
- Ruvalcaba, Y., Ray, C., Zinter, K. E., Le, V. D., Berra, L., & Leemis, R. (2026). Nonconsensual Pornography, Sextortion, and Deepfake Pornography Victimization: Findings from a National Online Sample of U.S. Adults. *Archives of Sexual Behavior*. <https://doi.org/10.1007/s10508-026-03503-4>
- Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., Pava, J., Maslej, N., Altman, R., Brynjolfsson, E., Brodley, C., Clark, J., Dignum, V., Kumar, V., Landay, J., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., ... Weld, D. (2026). *Artificial Intelligence Index Report 2026* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2606.15708>
- Schleider, J. L., & Morris, R. (2026). Debate: Conversational AI and young people’s mental health—friend or foe? A case for designing conversational AI as a bridge to youth mental health support. *Child and Adolescent Mental Health*, 31(3), 229–231. <https://doi.org/10.1111/camh.70105>
- Shank, D. B., Koike, M., & Loughnan, S. (2025). Artificial intimacy: Ethical issues of AI romance. *Trends in Cognitive Sciences*, 29(6), 499–501. <https://doi.org/10.1016/j.tics.2025.02.007>
- Shaw, S. D., & Nave, G. (2026). *Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender* (SSRN Scholarly Paper No. 6097646). Social Science Research Network. <https://doi.org/10.2139/ssrn.6097646>
- Shen, E., Hamati, F., Donohue, M. R., Girgis, R. R., Veenstra-VanderWeele, J., & Jutla, A. (2026). Evaluation of Large Language Model Chatbot Responses to Psychotic Prompts. *JAMA Psychiatry*, 83(6), 655. <https://doi.org/10.1001/jamapsychiatry.2026.0249>

- Shen, J. H., & Tamkin, A. (2026). *How AI Impacts Skill Formation* (arXiv:2601.20245). arXiv. <https://doi.org/10.48550/arXiv.2601.20245>
- Shroff, L. (2025, September 22). *Chatbait Is Taking Over the Internet*. The Atlantic. <https://www.theatlantic.com/technology/2025/09/chatbait-ai-chatgpt-engagement/684300/>
- Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": Experiences of generative AI chatbots for mental health. *Npj Mental Health Research*, 3(1), 48. <https://doi.org/10.1038/s44184-024-00097-4>
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion—A Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, 149, 102601. <https://doi.org/10.1016/j.ijhcs.2021.102601>
- Smith, M. (2026, September 9). *AI Models Are Watermarking Text—Will You Notice?* IEEE Spectrum. <https://spectrum.ieee.org/ai-watermark-text-anthropic-openai>
- Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can Generative AI Chatbots Emulate Human Connection? A Relationship Science Perspective. *Perspectives on Psychological Science*, 20(6), 1081–1099. <https://doi.org/10.1177/17456916251351306>
- Sohn, J.-S., Ha, B.-G., Park, S., Kim, J.-J., Lee, E., Oh, H., Lee, S., & Kim, E. (2026). Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *Npj Digital Medicine*, 9(1), 377. <https://doi.org/10.1038/s41746-026-02566-w>
- Stade, E. C., Stirman, S. W., Held, P., Schwartz, H. A., & Eichstaedt, J. C. (2024). *Designing Clinical Large Language Models for Evidence-Based Psychological Practice*. PsyArXiv. <https://doi.org/10.31234/osf.io/tjv8p>
- Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231. <https://doi.org/10.1038/s42256-024-00976-7>
- Stonier, J., Woodman, L., Alshammari, M., Cummings, R., Dad, N., Garg, A., Busetto, A. G., Hsiao, K., Hudson, M., Singh, P. J., Kanamugire, D., Kapoor, A., Lei, Z., Lu, J., Mizouni, E., Lungati, A. O., Loebel, M. P. C., Sethumadhavan, A., Telford, S., ... Teeuwen, S. (2023). *Data Equity: Foundational Concepts for Generative AI*. <https://doi.org/10.48550/ARXIV.2311.10741>
- Sun, X., Wang, Y., & McDaniel, B. T. (2026). AI companions and adolescent social relationships: Benefits, risks, and bidirectional influences. *Child Development Perspectives*, aadaf009. <https://doi.org/10.1093/cdpers/aadaf009>
- Tanner, B. (2026, May 20). Anthropomorphic AI terms create gaps in accountability. *Brookings*. <https://www.brookings.edu/articles/anthropomorphic-ai-terms-create-gaps-in-accountability/>
- Telzer, E. H., & Burnell, K. (2026). Smartphone Use During School Hours and Association With Cognitive Control in Youths Aged 11 to 18 Years. *JAMA Network Open*, 9(3), e261092. <https://doi.org/10.1001/jamanetworkopen.2026.1092>
- Telzer, E. H., Burnell, K., Maheux, A. J., Vaccaro, A., & Prinstein, M. J. (2026). A Developmental Neuroscience Framework for Youth-Centered Digital Media and Implications for Policy. *Policy Insights from the Behavioral and Brain Sciences*, 13(2), 442–452. <https://doi.org/10.1177/23727322261471447>
- The Jed Foundation. (2025, June 23). Tech Companies and Policymakers Must Safeguard Youth Mental Health in AI Technologies. *The Jed Foundation*. <https://jedfoundation.org/artificial-intelligence-youth-mental-health-pov/>
- Thorn. (2024, August 14). *REPORT: 1 in 10 Minors Say Peers Have Used AI to Generate Nudes of Other Kids*. <https://www.thorn.org/press-releases/report-1-in-10-minors-say-peers-have-used-ai-to-generate-nudes-of-other-kids/>
- Ting, F. (2026, August 10). *AI companions vanish without goodbyes as China tightens rules* | AP News. <https://apnews.com/article/china-ai-virtual-companions-bytedance-wechat-22c4247031092c37b61b537dd809b658>
- Tynes, B. M., Maxie-Moreman, A., Hoang, T.-M. H., Willis, H. A., & English, D. (2024). Online Racial Discrimination, Suicidal Ideation, and Traumatic Stress in a National Sample of Black Adolescents. *JAMA Psychiatry*, 81(3), 312–316. <https://doi.org/10.1001/jamapsychiatry.2023.4961>
- Tynes, B. M., McGee, T., & English, D. (2025). Adolescents' Daily Race-Related Online Experiences and Mental Health Outcomes. *JAMA Network Open*, 8(10), e2536870. <https://doi.org/10.1001/jamanetworkopen.2025.36870>

- Tynes, B. M., Willis, H. A., Stewart, A. M., & Hamilton, M. W. (2019). Race-Related Traumatic Events Online and Mental Health Among Adolescents of Color. *Journal of Adolescent Health, 65*(3), 371–377. <https://doi.org/10.1016/j.jadohealth.2019.03.006>
- Umbach, R., Henry, N., Shelby, R., Stevens, G., & Gonzalez-Pons, K. (2026). AI-generated image-based sexual abuse: Perpetration and consumption across three regions. *Computers in Human Behavior, 179*, 108935. <https://doi.org/10.1016/j.chb.2026.108935>
- United Nations Children's Fund. (2026). 'Deepfake abuse is abuse.' Unicef. <https://www.unicef.org/press-releases/deepfake-abuse-is-abuse>
- Veldhuis, A., Lo, P. Y., Kenny, S., & Antle, A. N. (2025). Critical Artificial Intelligence literacy: A scoping review and framework synthesis. *International Journal of Child-Computer Interaction, 43*, 100708. <https://doi.org/10.1016/j.ijcci.2024.100708>
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). *Sociotechnical Safety Evaluation of Generative AI Systems* (arXiv:2310.11986). arXiv. <https://doi.org/10.48550/arXiv.2310.11986>
- Weilhammer, V., Hou, K. Y., Luettgau, L., Summerfield, C., Dolan, R., & Nour, M. M. (2026). A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine, 1*–11. <https://doi.org/10.1038/s41591-026-04577-2>
- Weinstein, N., Itzhakov, G., & Maniaci, M. R. (2025). Exploring the connecting potential of AI: Integrating human interpersonal listening and parasocial support into human-computer interactions. *Computers in Human Behavior: Artificial Humans, 4*, 100149. <https://doi.org/10.1016/j.chbah.2025.100149>
- Wenger, E., & Kenett, Y. N. (2026). Large language models are homogeneously creative. *PNAS Nexus, 5*(3), pgag042. <https://doi.org/10.1093/pnasnexus/pgag042>
- Wenger, J. D., Cameron, C. D., & Inzlicht, M. (2026). People choose to receive human empathy despite rating AI empathy higher. *Communications Psychology, 4*(1), 19. <https://doi.org/10.1038/s44271-025-00387-3>
- Willis, H. A., Polanco-Roman, L., & Galán, C. (2026). Online racial discrimination, racial-ethnic identity beliefs, and racism-based traumatic stress symptoms among racially-ethnically minoritized adolescents. *Current Psychology, 45*(8), 794. <https://doi.org/10.1007/s12144-026-09068-8>
- Wilmer, H. H., & Chein, J. M. (2016). Mobile technology habits: Patterns of association among device usage, intertemporal preference, impulse control, and reward sensitivity. *Psychonomic Bulletin & Review, 23*(5), 1607–1614. <https://doi.org/10.3758/s13423-016-1011-z>
- Wilmer, H. H., Hampton, W. H., Olino, T. M., Olson, I. R., & Chein, J. M. (2019). Wired to be connected? Links between mobile technology engagement, intertemporal preference and frontostriatal white matter connectivity. *Social Cognitive and Affective Neuroscience, 14*(4), 367–379. <https://doi.org/10.1093/scan/nsz024>
- Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. *Big Data & Society, 12*(4), 20539517251389860. <https://doi.org/10.1177/20539517251389860>
- Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*(10), 1839–1850. <https://doi.org/10.1038/s41562-024-02004-5>
- Yao, R., Qi, G., Sheng, D., Sun, H., & Zhang, J. (2025). Connecting self-esteem to problematic AI chatbot use: The multiple mediating roles of positive and negative psychological states. *Frontiers in Psychology, 16*. <https://doi.org/10.3389/fpsyg.2025.1453072>
- Ye, M., Ibrahim, L., Bo, J. Y., Cheng, M., Mattsson, I., Vennemeyer, D., Kraut, R., & Rathje, S. (2026). *What Counts as AI Sycophancy? A Taxonomy and Expert Survey of a Fragmented Construct* (arXiv:2605.21778). arXiv. <https://doi.org/10.48550/arXiv.2605.21778>
- Ye, M., Kraut, R., & Rathje, S. (2026). *Individual-level interventions against sycophantic AI reduce its appeal but not its persuasiveness* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2607.25166>
- Yeung, J. A., Dalmasso, J., Foschini, L., Dobson, R. J., & Kraljevic, Z. (2025). *The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models* (arXiv:2509.10970). arXiv. <https://doi.org/10.48550/arXiv.2509.10970>

- Yeung, J. A., Morrin, H., Ng, V., Kraljevic, Z., & Dobson, R. (2026). *An Echo Chamber of One: Should AI Psychosis Be a Distinct Clinical Entity?* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2608.23937>
- Yim, S. H., Yoo, D. W., Polymerou, A., Liu, Y., & Saha, K. (2026). Generative AI for Eating Disorders: Linguistic Comparison With Online Support and Qualitative Analysis of Harms. *International Journal of Eating Disorders*, 59(3), 519–533. <https://doi.org/10.1002/eat.24604>
- Yu, C.-L. (2025). Fear of missing out on AI: A psychological cost of technological revolution. *Telematics and Informatics Reports*, 20, 100277. <https://doi.org/10.1016/j.teler.2025.100277>
- Yu, Y., Li, Y., Suri, S., & Counts, S. (2026). *From Searchable to Non-Searchable: Generative AI and Information Diversity in Online Information Seeking* (arXiv:2604.10258). arXiv. <https://doi.org/10.48550/arXiv.2604.10258>
- Yu, Y., Liu, Y., Zhang, J., Huang, Y., & Wang, Y. (2025). *Understanding Generative AI Risks for Youth: A Taxonomy Based on Empirical Data* (arXiv:2502.16383). arXiv. <https://doi.org/10.48550/arXiv.2502.16383>
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1), 28. <https://doi.org/10.1186/s40561-024-00316-7>
- Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues. *Journal of Medical Internet Research*, 27(1), e78238. <https://doi.org/10.2196/78238>
- Zhang, X., Mao, R., & Cambria, E. (2026). *A Multifaceted Analysis of Social Biases in Large Language Models* (arXiv:2512.15792). arXiv. <https://doi.org/10.48550/arXiv.2512.15792>
- Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2026). Interaction with AI companions and psychological well-being. *Nature Human Behaviour*, 1–15. <https://doi.org/10.1038/s41562-026-02516-2>
- Zuromski, K. L., Meagher, M., Costigan, T., & Olson, E. A. (2026). AI chatbots and youth suicide risk: Current evidence, critical gaps, and a clinical research agenda. *Npj Digital Medicine*. <https://doi.org/10.1038/s41746-026-03080-9>

