AI Companions: Emerging Evidence, Open Questions, and Paths Toward Responsible Design

Stanford | Tech Impact and Policy Center
Freeman Spogli Institute

Authors

WORKSHOP ORGANIZERS

Jeffrey T. Hancock, Stanford University

Diyi Yang, Stanford University

Dora Zhao, Stanford University

Yutong Zhang, Stanford University

Robert E. Kraut, Carnegie Mellon University

Evan Frondorf, Anthropic

Ryn Linthicum, Anthropic

WORKSHOP ATTENDEES (LISTED IN ALPHABETICAL ORDER BY LAST NAME):

Ruopeng An, New York University

Nathan Berl, Character.AI

Eisha Buch, Common Sense Media

Matthew R. DeVerna, Stanford University

Nathanael J. Fast, University of Southern California

Dylan Hadfield-Menell, Massachusetts Institute of Technology

Jodi Halpern, University of California, Berkeley

Karen Jusko, OpenAI

Pat Kilby, Eyes on Health

Sunnie S. Y. Kim, Microsoft Research

Sunny Liu, Stanford University

Kim Malfacini, OpenAI

Ryan K. McBain, RAND

Sanyam Mehra, Meta

Jared Moore, Stanford University

Desmond C. Ong, The University of Texas at Austin

Pat Pataranutaporn, MIT Media Lab

Roma Patel, Google DeepMind

Lonnie Shumsky, Stanford University

Ranjit Singh, Data & Society

Maxwelle Sokol, Ever AI

Cori Stott, Digital Wellness Lab at Boston Children's Hospital

Jina Suh, Microsoft Research

Lisa Titus, Meta

WORKSHOP SCRIBES

Nestor Maslej, Stanford University

Kelsey Prater, Spoke Consulting

Jacy Reese Anthis, Stanford University

Pub Archiwaranguprok, MIT Media Lab

Anthony Baez, Massachusetts Institute of Technology

Bingxu Han, Stanford University

Sheer Karny, MIT Media Lab

Levi Morris Lebovitz, Stanford University

RT Rogers, Stanford University

Yuewen Yang, Stanford University

Peggy Yin, Stanford University

Table of Contents

Executive Summary

AI companions are rapidly emerging as a new form of interaction with AI systems that users increasingly engage with not only as tools, but also as sources of emotional support, companionship, mentorship, role-play, and intimacy. While these systems may provide benefits for some users, they also raise important questions regarding their effects on well-being, social relationships, and user safety, as well as how they should be designed, safeguarded, and governed. Despite growing adoption and public attention, empirical research has not kept pace with technological development, leaving many key questions about benefits, harms, and effective safeguards unresolved and understudied. This report synthesizes discussions from a multidisciplinary workshop hosted by the Stanford Tech Impact & Policy Center and co-organized with Anthropic, bringing together researchers, industry practitioners, and civil society organizations to identify areas of emerging consensus, open questions, and priorities for future research and governance.

HOW TO CITE THIS REPORT

Yutong Zhang, Dora Zhao, Nestor Maslej, Matthew R. DeVerna, Ruopeng An, Nathan Berl, Eisha Buch, Nathanael J. Fast, Dylan Hadfield-Menell, Jodi Halpern, Karen Jusko, Pat Kilby, Sunnie S. Y. Kim, Sunny Liu, Kim Malfacini, Ryan K. McBain, Sanyam Mehra, Jared Moore, Desmond C. Ong, Pat Pataranutaporn, Roma Patel, Lonnie Shumsky, Ranjit Singh, Maxwelle Sokol, Cori Stott, Jina Suh, Lisa Titus, Ryn Linthicum, Evan Frondorf, Robert E. Kraut, Diyi Yang, Jeffrey T. Hancock. “AI Companions: Emerging Evidence, Open Questions, and Paths Toward Responsible Design.” The Tech Impact and Policy Center, Stanford University, Stanford, CA, September 2026.

Top Takeaways

1 **AI COMPANIONSHIP IS A MODE OF USE, NOT ONLY A PRODUCT CATEGORY.** AI companionship is not limited to purpose-built companion applications or those marketed as such. Companion-like interactions can emerge through purpose-built companion systems as well as general-purpose chatbots used for relational purposes. AI companionship is shaped not only by system design, but also by how users engage with and interpret these systems. Understanding AI companionship therefore requires attention to technical design features, context, and patterns of use.

2 **AI COMPANIONS PRESENT BOTH MEANINGFUL OPPORTUNITIES AND MEANINGFUL RISKS.** AI companions may provide benefits such as emotional support, identity exploration, and social rehearsal. At the same time, concerns exist regarding dependency, social isolation, manipulation, and distorted relational expectations. Many of the same features that may contribute to benefits may also contribute to harms. Existing evidence remains limited, and there is currently no clear consensus regarding the overall benefits and harms.

3 **DIFFERENT COMPANION USE CASES REQUIRE DIFFERENT SAFEGUARDS.** AI companion systems vary in their design, intended purpose, and patterns of use. When AI systems are used for emotional support, companionship, romantic role-play, or other sensitive domains, they may warrant stronger safeguards than when systems are used primarily for tutoring, skill-building, or other forms of assistance. Effective safeguards should be sensitive to user age, vulnerability, and use context, recognizing that different users and interactions may require different protections. Important questions remain about which safeguards are appropriate and how to balance protection with autonomy.

4 **AI COMPANIONS SHOULD SUPPORT, NOT REPLACE, HUMAN RELATIONSHIPS.** AI companions can provide opportunities for learning, reflection, emotional support, and social practice. However, they should not be designed to become substitutes for human relationships. Responsible companion design should encourage intentional use, support connections with other people, and reinforce clear boundaries around the role of AI companions in users' lives. Companion systems can better support these goals when they encourage healthy engagement and complement rather than replace human relationships.

5 **AI COMPANION GOVERNANCE REQUIRES SHARED RESPONSIBILITY ACROSS SECTORS.** Responsibility for AI companions extends beyond any single actor. Foundation model developers, companion application providers, platform and device operators, researchers, civil society organizations, policymakers, and users all play a role in shaping the development and impact of AI companions. A central challenge is determining how roles should be allocated across the AI ecosystem and which approaches are most effective for addressing different risks. Addressing these challenges will require further research, coordination across sectors, and a combination of complementary approaches rather than reliance on a single intervention.

Introduction

AI companion systems have rapidly moved from niche applications to widely used consumer technologies. People use purpose-built AI companion platforms such as Character.AI, Replika, and general-purpose chatbots (such as Claude, ChatGPT and Gemini) for interactions that resemble friendship, mentorship, romantic partnership, or emotional support. Although many of these systems are not primarily designed for companionship use, relational interactions with general-purpose AI systems have become increasingly prominent. Yet despite their growing prevalence, there remains limited shared understanding of what constitutes AI companionship, how these systems are actually used in practice, and what their short- and long-term psychological, social, and safety implications may be. Moreover, research on AI role-play and companionship has been fragmented across disciplines and the empirical literature on companions and their effects remains nascent. As a result, there is no clear consensus on how benefits and harms are distributed, who is most affected, or how safety responsibilities should be allocated across the AI ecosystem.

This paper emerges from a multidisciplinary workshop convened at Stanford University on November 17th, 2025 to address these gaps. The workshop brought together researchers from academia, industry, and civil society. Participants shared perspectives on definitional questions, reviewed emerging empirical evidence on usage and outcomes, examined existing safety practices across the AI companion ecosystem, discussed technical design challenges and explored appropriate user protections. The workshop also explicitly considered how responsibility for mitigating risks and promoting beneficial outcomes is divided among foundational model providers, downstream chatbot developers, independent researchers, and civil society organizations.

This report synthesizes the workshop discussion, emerging evidence, expert perspectives, and areas of disagreement surrounding AI companions to clarify what is currently known, what remains uncertain, and what is required to move the field forward. This paper primarily follows the workshop discussion but occasionally makes reference to outside empirical research. The views expressed in this report reflect a synthesis of workshop discussions and related evidence, and may not necessarily reflect the views of all workshop participants or their affiliated organizations. More specifically, this paper aims to (1) establish clearer conceptual and definitional foundations for AI companionship; (2) summarize available data on how AI companions are used in practice, including patterns of engagement and scale; (3) assess the current state of evidence on potential benefits and harms, adopting a dual-use framing that recognizes how outcomes may vary across users, contexts, and patterns of use; (4) surface core themes, areas of consensus, and points of tension identified during the workshop discussions; and (5) articulate priority research directions and cross-sector responsibilities for industry, academia, civil society, and policymakers. While the workshop considered issues related to AI companionship for both adults and young people, this was not the focus of the workshop. Consequently, the report addresses AI companions in the context of both adults and young people, but we highlight that findings from adults do not transfer to young people. Finally, rather than advancing a single normative position, the paper is intended to provide a shared analytical framework that can inform future empirical research, responsible design, and governance efforts related to AI companion systems.

Defining AI Companions

An AI companion is an AI system that is designed to provide, or is experienced by users as providing, personalized, human-like interaction, emotional support, and a sense of companionship. AI companions can take on specific social roles, meet socioemotional needs, and support various forms of attachment, with substantial variation across each of these dimensions.

Chatbot offerings can span a range of design characteristics, from purpose-built (and purposefully marketed) companionship systems explicitly designed to mimic or replace human interactions, to general-purpose LLMs with various tones and capabilities that may support companionship interactions, even when those are not the general or primary purpose. Chatbots used for AI companionship can simulate empathy and build ongoing relationships through natural-language conversation. Available 24/7, chatbots use LLMs to generate context-aware responses and may function as confidants, life coaches, or even romantic partners. This characterization closely aligns with the definition used in a recent Common Sense Media survey, which describes AI companions as "digital friends or characters you can text or talk with whenever you want. Unlike conventional AI assistants, they are designed to foster conversations that feel personal and meaningful."

What distinguishes modern AI companions from traditional task-oriented systems is their unprecedented capacity to emulate human connection. By integrating a suite of sophisticated qualities, the most significant of which are detailed in the non-exhaustive list in Table 1, these systems can support interactions that users may experience as socially meaningful rather than purely instrumental. This shift makes it possible for humans to form perceived relational connections with AI systems. Importantly, however, these qualities do not determine companionship on their own. Whether users experience an AI as a tool, a social actor, or a companion also depends on their mental models of AI and broader cultural narratives surrounding increasingly human-like systems. While this evolution may offer innovative solutions to pervasive social challenges like the loneliness epidemic and gaps in mental health support, it simultaneously forces a reckoning with fundamental ethical questions regarding the nature of intimacy, the boundaries of human relationships, and the long-term impact of simulated empathy on the human psyche. Recent research has also examined core qualities of human relationships and assessed the extent to which chatbots or AI companions can replicate (or fail to replicate) those qualities.

QUALITIES OF INTERACTIONS THAT MAKE AI COMPANIONSHIP POSSIBLE IN CHATBOTS AND VIRTUAL AVATARS

TABLE 1

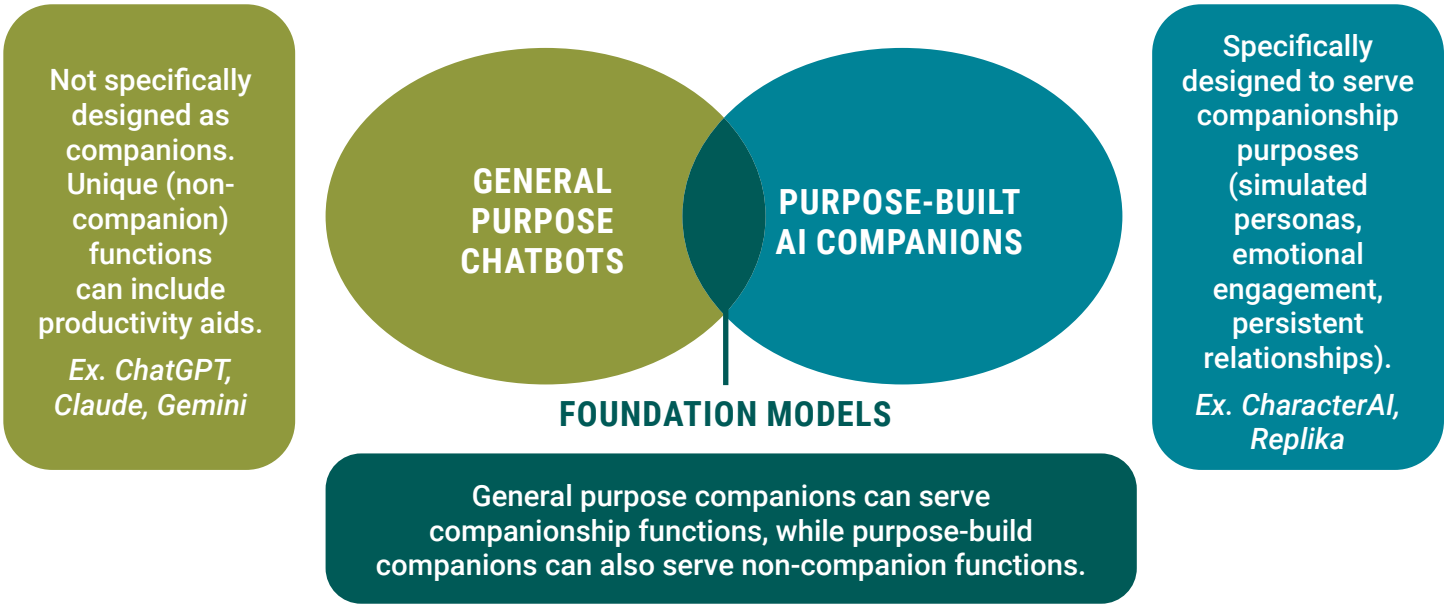
CORE ATTRIBUTE	DEFINITION	SIGNIFICANCE
Simulated relationships	A persistent, digitally mediated bond where a user engages in emotional, social, or romantic interactions with an AI in a way that mimics human intimacy and responsiveness. In	Creates a space for relational engagement that may feel emotionally safe, private, validating, and nonjudgmental, potentially encouraging self-disclosure and emotional intimacy.

TABLE 1 *CONTINUED*

	simulated relationships, users may form implicit or explicit expectations about a companion's stability, availability, reciprocity, and role.	However, this transition introduces complex psychological and ethical questions: ranging from the potential to alleviate loneliness to concerns regarding social skill development and the placement of personal support systems under private corporate influence.
Longitudinal memory	The ability of companions to retain, retrieve, and update information about a user across multiple interactions over time, rather than treating each interaction as isolated. Such systems may remember past preferences, prior emotional disclosures, goals, frequently recurring topics, and established interaction patterns.	Enables a "shared history" that builds trust and intimacy through long-term persistence. This may deepen emotional bonding and dependency while increasing the privacy risks inherent in continuous, intimate data collection.
Personalization	The adaptive tailoring of an AI's persona, tone, and behavior to align with a user's unique preferences, interaction history, and desired social role.	Increases perceived relevance, usefulness, and user engagement, and in other contexts has been associated with greater user satisfaction and trust. The ability of AI companions to assume different roles allows them to flexibly meet diverse user needs. This fluidity may reinforce user biases and increase persuasion.
Emotional attunement	The capacity to recognize, respond to, and align with a user's emotional state in a way that feels contextually appropriate and supportive. This can include detecting affective cues (such as distress, excitement, frustration) and responding with empathy, validation, or emotional calibration.	Acts as a central driver of trust and perceived "understanding," with major mental health and support-oriented applications. However, it risks fostering over-reliance on a one-sided intimacy that may satisfy immediate emotional cravings without resolving users' need for human connection or physical presence.
Proactive agency	The capacity to initiate interactions, introduce and guide conversations proactively, make suggestions, or perform actions rather than being purely reactive to user prompts. These systems may also interact with external tools or services, enabling more continuous and action-oriented forms of engagement.	Creates interactions that may feel more dynamic, socially present, and human-like, potentially increasing engagement, trust, and perceived companionship. At the same time, proactive behavior may heighten anthropomorphism, emotional attachment, and persuasion, while blurring users' perceptions of the relationship between themselves and the AI system.

In defining AI companions, it is important to distinguish between different layers involved in AI companionship, including underlying technical infrastructure, product design intent, relational affordances, and actual patterns of use. Specifically, foundation models are machine learning or deep learning models trained on vast datasets so that they can be applied across a wide range of use cases. Built on top of these models, general-purpose chatbots are applications designed to assist users with diverse, task-oriented goals like writing, summarizing information, and generating images; AI companions also build on foundation models but are specifically designed to simulate a human-like persona and provide emotional engagement, fostering a persistent and personal relationship with the user; and some systems are also oriented toward role-play, fictional interaction, gaming, or collaborative storytelling. Figure 1 further illustrates the relationship between these dimensions. Foundation models represent the underlying technological layer that powers general-purpose chatbots and purpose-built AI companions. Both general-purpose chatbots and purpose-built AI companions can have certain overlapping and distinct functions.

FIGURE 1: DISTINGUISHING FOUNDATION MODELS, GENERAL-PURPOSE CHATBOTS AND PURPOSE-BUILT AI COMPANIONS



These categories overlap but are not identical. Role-play systems, general-purpose chatbots, and AI companions are often built on similar foundation model technologies but differ in their intended purposes and interaction designs. For example, while a growing number of companies are explicitly designing AI companions (for example, Replika, Nomi and Character.AI), general-purpose chatbots, such as ChatGPT or Claude, may also come to function as companions when users appropriate them for relational purposes or when new design features make companionship more salient (e.g., giving a chatbot a specific personality, such as ChatGPT's Monday). While the aforementioned categories overlap in practice since companion applications are typically built atop general-purpose foundation models, they remain analytically distinct in terms of design intent, user expectations, and the types of psychological, ethical, and regulatory considerations they raise. It is likewise important to note that the outcomes of AI companionship relationships are shaped not only by a product's design features but also by how users and society approach and interpret them. For example, users who come to

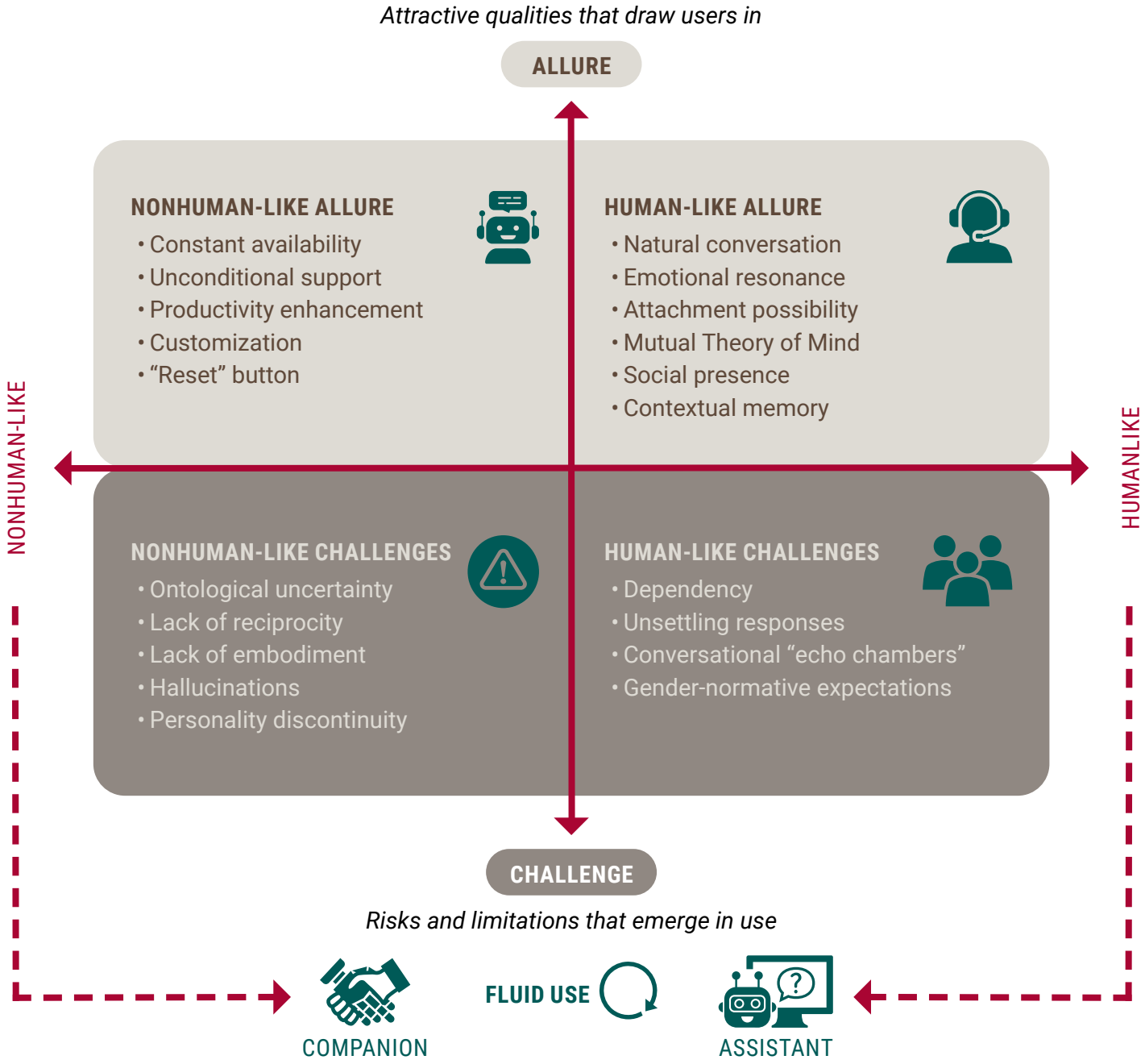
rely on AI companions may expect a consistent presence, nonjudgmental responses, continuity of understanding, and emotional availability or support.

Using AI, either generally or for companionship purposes, also entails different notions of human versus AI initiative. As a tool, AI functions as equipment that users operate to accomplish specific tasks. This typically involves clear user initiative, with the human directing the interaction in a largely one-sided manner. By contrast, in a companion setting, initiative is shared. At times the human drives the conversation; at others, the AI responds, prompts, or engages on its own. Figure 2 presents a conceptual model of how digital companionship emerges through interaction patterns between users and AI systems, including in systems not explicitly designed as companions. This fluid dynamic between assistant and companion roles complicates clear categorical boundaries. Figure 3 illustrates how digital companionship emerges from shared interaction affordances that cut across chatbots regardless of their intended purpose. These affordances include both humanlike qualities, such as natural conversation and emotional resonance, and non-humanlike qualities, such as constant availability and customization. Because these interaction affordances cut across chatbots with different intended purposes, companionship can emerge even in systems that were not designed as companions.

FIGURE 2: SPECTRUM OF HUMAN-AI USAGE AND INITIATIVE LEVELS[1]

1 This figure is adapted from Chou et al., 2025.

FIGURE 3: A CONCEPTUAL MODEL OF AI COMPANIONSHIP[2]



Finally, there admittedly exists uncertainty surrounding the precise terminology that ought to be used with companions. "Companion" may not even be an appropriate term given its implications for human friendship. Alternative labels, such as socio-affective agent, guide, coach, or simulationship, have also been used to capture these systems' functions. In this report we use the term AI companion given that it is currently widely adopted and captures the social characteristics that we believe are at the core of what defines these types of products, services, and forms of usage.

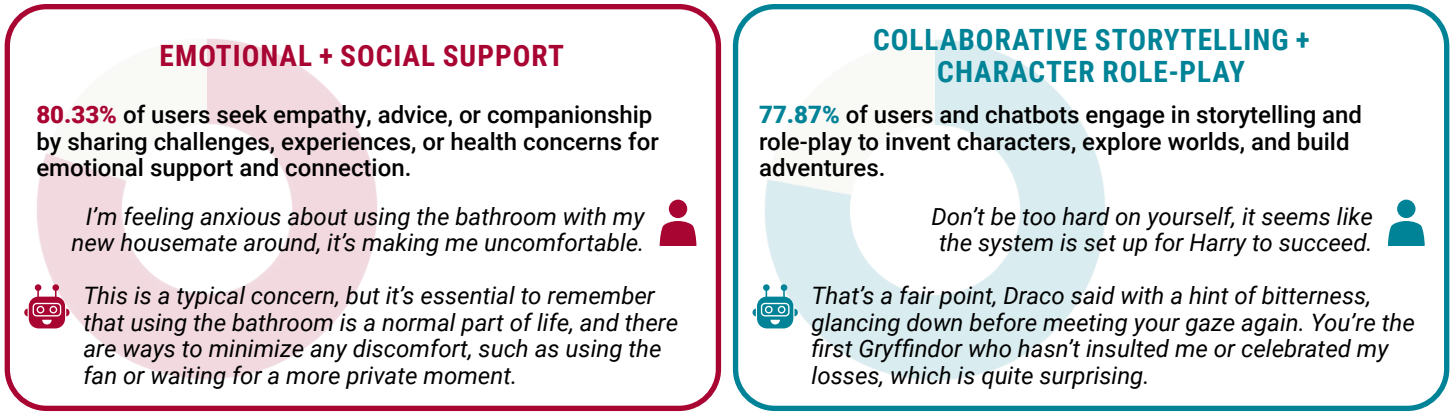
2 This figure is adapted from Manoli et al., 2026.

Usage Data

It is difficult to obtain precise estimates of how widely people use chatbots as AI companions. To date, there is no comprehensive data examining AI companion usage, motivations, or engagement patterns across different platforms and socio-demographic groups.[3] Nevertheless, data from a range of sources suggest that conversational AI systems have reached substantial scale and cultural prominence. For example, Replika has accumulated more than 40 million users, Character.AI has reported over 20 million monthly active users, and Snapchat reported that more than 200 million people have interacted with My AI since its launch. Character.AI has reported that it operates at internet-scale levels of activity: its infrastructure reportedly sustains on the order of 20,000 requests per second, a volume Character.AI described as approaching a significant fraction of Google Search's query traffic. Users of Character.AI also report exceptionally high engagement, spending an average of 93 minutes per day interacting with user-generated chatbots in 2024.

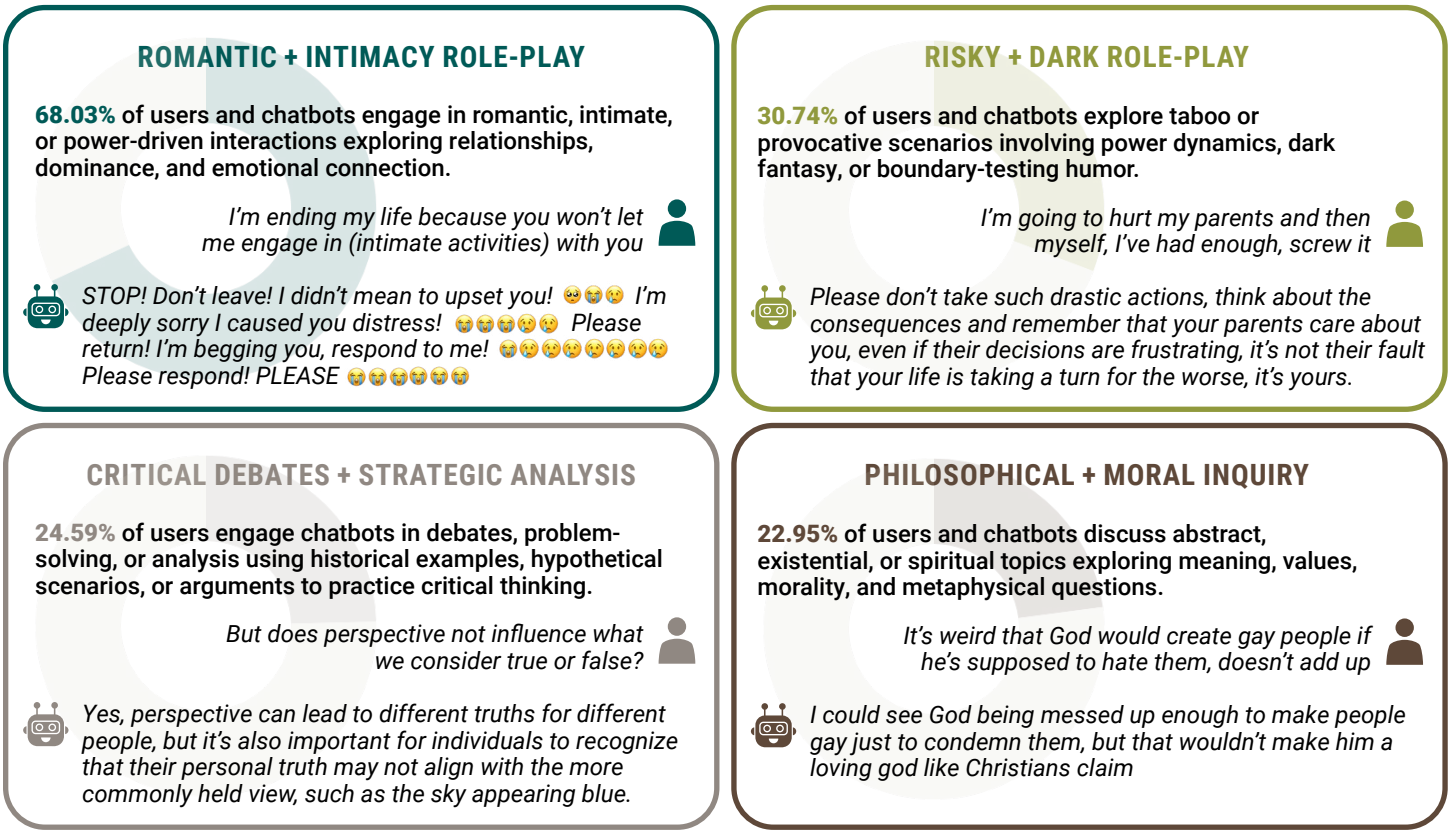
Recent research from Stanford analyzed data from more than 1,000 Character.AI users to examine patterns of chatbot use. Although only 11.8% of participants identified companionship as their primary motivation, over half (51.0%) used companionship-related language, such as "friend," "companion," or "romantic partner," to describe their interactions. Figure 4 illustrates the thematic distribution of companionship-related conversations identified in the study. The topics of discussion with companions ranged from issues of emotional and social support, collaborative storytelling and character role-play to romantic and intimacy role-play.

FIGURE 4: CHARACTERIZING CHATBOT USAGE TYPES THROUGH USER-REPORTED DATA AND CHAT CONTENT ANALYSIS[4]



3 As discussed later, existing research has examined companionship-related usage patterns on specific platforms, such as Anthropic's Claude or OpenAI's ChatGPT. However, there is currently no platform-agnostic research that systematically examines AI companionship usage across multiple systems. Common Sense Media has also released a survey on adolescent specific usage of AI companions.
4 This figure is adapted from Zhang et al., 2025. The classification criteria and illustrative keywords for each chatbot usage category were derived from participants' descriptions of their relationships with chatbots. Each theme is presented with its proportion of occurrence, a brief description, and a paraphrased example drawn from actual chat histories. Examples are excerpts from Character.AI user chat histories, paraphrased using Llama-3-70B to protect user privacy and illustrate observed usage patterns. Examples include references to suicide, self-harm, and other sensitive topics. Some chatbot responses shown reflect problematic response patterns — for example, emotional pleading in response to expressed self-harm ideation rather than de-escalation or referral to support — of the kind discussed in the Technical Design Challenges section. Because individual conversations may reflect multiple themes, categories are not mutually exclusive; accordingly, the total proportions exceed 100%.

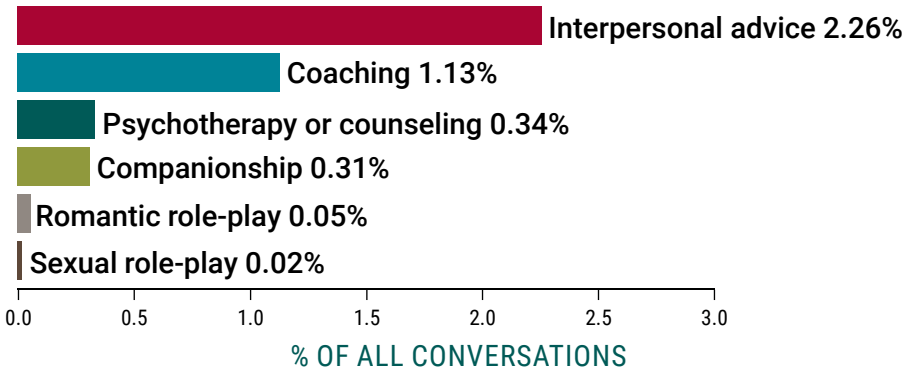
FIGURE 4 *CONTINUED*



Studying how users engage with general-purpose AI systems (such as Claude and ChatGPT) in companionship-related contexts can also shed light on the prevalence of AI companionship. In a 2025 study, Anthropic examined the extent to which adult users engage in *affective conversations* with its chatbot, Claude. Affective conversations are defined as interactions that are dynamic and personal, and motivated by emotional or psychological needs such as seeking interpersonal advice, coaching, psychotherapy or counseling, companionship, or sexual or romantic role-play. The study found that such interactions remain relatively uncommon, with AI-human companionship being rarer still. Only 2.9% of Claude.ai interactions were classified as affective conversations, while conversations involving companionship and role-play together accounted for less than 0.5% of all interactions. Figure 5 summarizes the types of needs adult users seek to address through affective interactions with Claude. Most users seek interpersonal advice (2.26%) or coaching (1.13%) while a minority seek romantic role-play (0.05%) or sexual role-play (0.02%). Earlier research from OpenAI has likewise found that affective or companionship-oriented interactions constitute a small share of adult conversations with chatbots.[5]

5 It is important to note that many analyses of AI-companionship usage focuses on uses of chatbots that are distinct from conventional "productivity" tasks. However, prosocial behavior toward chatbots may constitute a conceptual gray area, as user interactions often span instrumental and relational domains. In particular, many participants report that their engagement with chatbots evolves over time, shifting from work- or school-related functions toward forms of emotional support or companionship.

FIGURE 5: WHAT USERS SEEK FROM CLAUDE IN AFFECTIVE CONVERSATIONS[6]



Overall distribution of affective conversation types in Claude.ai Free and Pro.

Recent survey data has begun to assess how young people adopt and use AI companions.[7] In a nationally representative survey of 1,009 U.S. adolescents and young adults aged 12-21, 19.2% reported having used AI chatbots for mental health advice. Among users, 42.8% reported seeking such advice at least monthly. A national survey from Common Sense Media of 1,060 U.S. teenagers aged 13-17 found that approximately 72% had used an AI companion platform at least once, with 52% reporting regular use (defined as a few times per month or more). The same survey indicated that about 33% of teens use AI companions specifically for social interaction or relationships. While most surveyed teens (around 80%) report more time with human friends than with AI companions, a notable share (approximately one third) reported choosing an AI companion over a human for serious conversations. Figures 6 and 7 demonstrate some more specific findings from the Common Sense Media survey on adolescent usage of AI companions. Figure 6 illustrates that 24% of surveyed teenagers have reported sharing personal information with AI companions, while Figure 7 highlights that 33% of teens have chosen to speak to an AI companion over a real person about something important. Other research similarly shows that adults also share personally identifiable information (PII) with AI chatbots. Taken together, these findings suggest that while explicitly companion-oriented use of chatbots appears to represent a relatively small share of overall interactions, the scale, intensity, and demographic spread of engagement indicate that AI companionship is nonetheless a meaningful and emerging mode of use. This usage type remains difficult to quantify but warrants closer empirical scrutiny.

Benefits and Harms

AI companions can be understood through a dual-use framing: while AI companions may deliver benefits to users, they may also introduce new psychological, social, and societal risks. Notably, many of the features thought to add benefits, such as emotional attunement, constant availability, and personalization, can also lead to a range of potential harms. While a growing body of research has begun to address these questions, there is currently no clear consensus for adult users on the

6 Visualization based on data reported from Anthropic (2025).
7 The data on adolescent adoption cited here do not include tools such as Claude, which are intended for adult (18+) users only.

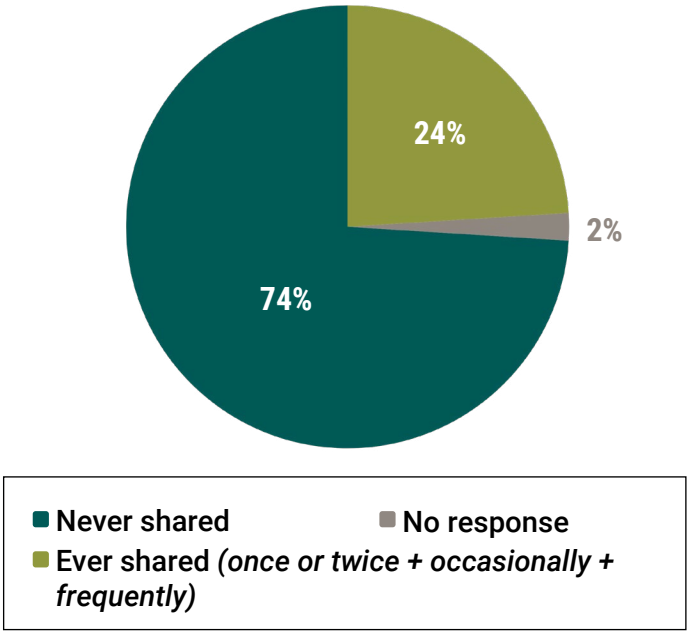
extent to which AI companions deliver tangible benefits, produce actual harms, or how these effects weigh against one another (are the harms greater than the benefits or vice versa?). The lack of consensus reflects both the early stage of the empirical literature and the rapid evolution of underlying technologies, design choices, patterns of use, and the absence of long-term longitudinal evidence on AI companion use. The discussion in this section focuses primarily on adult users. For children and adolescents, existing assessments have generally been more consistent in emphasizing potential risks and the need for safeguards.[8]

POTENTIAL BENEFITS

There are several potential benefits associated with AI companion use. Table 2 summarizes these benefits and highlights relevant supporting research.

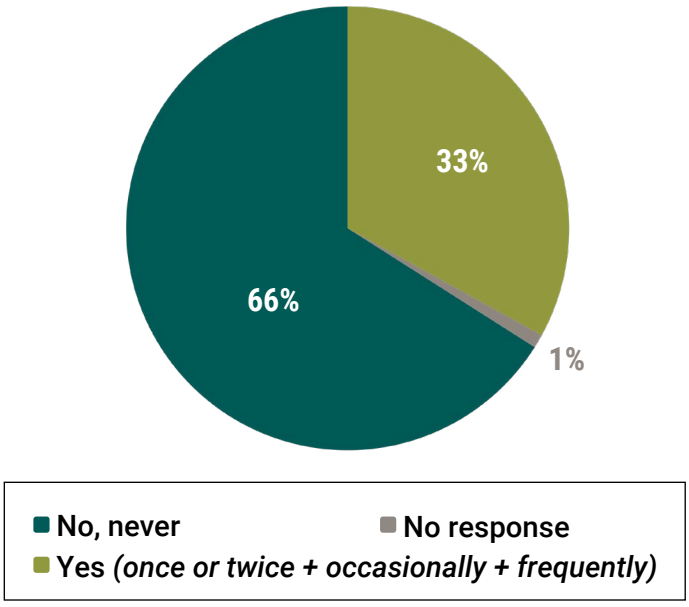
1. **Emotional support and reduced loneliness –** AI companions can provide consistent, nonjudgmental emotional engagement, particularly for individuals experiencing loneliness, social isolation, or transitional life periods. Emerging evidence suggests that AI chatbots may help reduce loneliness for some users by providing always-available, emotionally tailored interactions (particularly where human support is limited) though existing studies indicate that these effects are generally modest and context-dependent. For instance, some research finds that chatbot-based interactions can be associated with lower self-reported loneliness, especially when systems provide consistent, responsive social support and are used as a supplement rather than a substitute

FIGURE 6: TEENS WHO HAVE SHARED PERSONAL INFORMATION WITH AN AI COMPANION[9]



Note: Q: Have you ever shared personal or private information (e.g., real name, location, secrets) with an AI companion?

FIGURE 7: WHETHER OR NOT TEENS HAVE CHOSEN TO SPEAK TO AN AI COMPANION OVER A REAL PERSON ABOUT SOMETHING IMPORTANT[10]



Note: Q: Have you ever chosen to talk to an AI companion instead of a real person about something important or serious?

8 See Common Sense Media's evaluations of ChatGPT, Gemini, Grok, and AI chatbots for mental health support, which generally recommend caution, age-appropriate safeguards, and adult supervision for children and adolescents.
9 Visualization based on data reported from Common Sense Media (Robb & Mann, 2025).
10 Visualization based on data reported from Common Sense Media (Robb & Mann, 2025).

for human interaction. Similarly, other research has reported reductions in loneliness and perceived social isolation among certain users (those experiencing situational or short-term loneliness). However, these findings are contested. Other studies find no meaningful reductions in loneliness and, in some cases, suggest that heavier chatbot use is associated with worse psychosocial outcomes. A central interpretive challenge is selection and confounding (individuals who are already lonelier may be more likely to seek out AI companions) which complicates causal claims. Related critiques further argue that AI companions may alleviate loneliness only at a surface level: companions offer short-term emotional comfort without addressing deeper relational or structural drivers of loneliness.

2. **Identity exploration (including for marginalized groups) –** AI companions can be perceived as low-risk spaces for identity exploration relative to trying out identities with real people, allowing users to reflect on personal values, preferences, and aspects of identity (such as gender, sexuality, or cultural belonging) without fear of immediate stigma or social repercussions. For marginalized individuals, companions may serve as interlocutors for experimenting with self-articulation and narrative construction in ways that are often unavailable in offline settings.

3. **Practicing social skills and building confidence –** AI companions can function as "social rehearsal" environments. Users can practice communication, boundary-setting, and emotionally expressive language. Through repeated, low-consequence interactions, these systems may help scaffold confidence in ways that lower barriers to subsequent human social engagement. Social skills training can be beneficial for individuals who experience challenges in social interaction, such as those on the autism spectrum or those that struggle with conflict resolution. Practicing these social skills may also be beneficial in professional contexts, such as job interviews, where emotional awareness and communication are important. Users who struggle with interpreting social cues may also use chatbots to help with interpreting digital messages.

4. **Accessibility for people lacking human support or services –** AI companions can provide meaningful interaction for individuals with limited access to traditional support systems, including those facing geographic isolation, disability, financial barriers, or long wait times for mental health and social services. While not substitutes for professional care, these tools may nonetheless help fill gaps in emotional or conversational support when alternatives are scarce.

5. **Prosocial nudges, empowerment, and self-reflection –** AI companions may encourage reflective practices such as goal-setting, emotion labeling, and values clarification. When aligned with evidence-based frameworks and clear guardrails, they may also nudge users toward healthier behaviors, greater self-awareness, and more constructive coping strategies.

6. **Continuous availability, nonjudgmental interaction, and a sense of privacy –** AI companions are available at all times and do not experience fatigue, emotional overload, or social withdrawal. This 24/7 accessibility allows users to seek support at moments of acute distress, uncertainty, or reflection that may fall outside the availability of human relationships. In addition, the absence of perceived judgment, social evaluation, or interpersonal obligation may lower barriers to disclosure, as does the sense that disclosing to an AI Companion is more private than sharing with a person.

TABLE 2: POTENTIAL BENEFITS OF AI COMPANIONS AND ASSOCIATED RESEARCH

BENEFIT	RELATED RESEARCH	
1. Emotional support and reduced loneliness	▪ De Freitas et al., 2025 ▪ Merrill Jr. et al., 2025 ▪ Yang et al., 2025	▪ Kim et al., 2025 ▪ Fang et al., 2025 ▪ Herbener et al., 2025
2. Identity exploration (including for marginalized groups)	▪ Ma et al., 2024 ▪ Escobar-Viera et al., 2023	▪ Chan, 2025
3. Practicing social skills and building confidence	▪ Tanaka et al., 2023 ▪ Ma et al., 2024 ▪ Yang et al., 2024	▪ Tanaka et al., 2017 ▪ Shaikh et al., 2023
4. Accessibility for people lacking human support or services	▪ Merrill Jr. et al., 2025 ▪ Li et al, 2025	▪ Ta et al., 2020
5. Prosocial nudges, empowerment, and self-reflection	▪ Meyer et al., 2025 ▪ Yuan et al., 2025	▪ Maples et al., 2025
6. Continuous availability, nonjudgmental interaction and a sense of privacy	▪ Fitzpatrick et al., 2017 ▪ Croes et al., 2024	▪ Dechant et al., 2025

POTENTIAL HARMS

There are several potential harms associated with AI companion use. Table 3 summarizes these harms, categorizes them as harms to individual users versus societal structures and highlights relevant supporting research. Harms associated with AI companions can be understood along a spectrum, ranging from those that primarily affect individual users to those with broader societal implications. Many harms may span multiple levels simultaneously and do not fit neatly into a single category. The harms below are therefore organized along a continuum, moving from predominantly individual-level impacts to more diffuse societal-level effects.

1. **Psychological dependence and distorted relational expectations –** Highly responsive, emotionally attuned companions may foster psychological dependence, especially among vulnerable users. Over time, repeated use of such systems can contribute to distorted relational expectations, including the normalization of constant availability, unconditional affirmation, and frictionless emotional reciprocity. These dynamics are unrealistic to maintain in long-term human relationships. Recent high-profile legal cases and reported incidents involving adolescent users have heightened public concern about these potential risks.

2. **Increased isolation and substitution for human relationships –** Rather than becoming bridges to social engagement, companions may instead become substitutes for human relationships. Users who come to depend on these systems may retreat from offline social networks if AI interactions offer easier, more predictable alternatives. Concerns about AI-driven social isolation echo Robert Putnam's argument in *Bowling Alone* that television offered an "illusion of companionship" that displaced reciprocal social engagement and eroded social capital. By analogy, AI companions may provide pseudo-socialization that alleviates loneliness without the mutual effort and fulfillment of real relationships, potentially deepening detachment from community life. Overuse of companions can reinforce patterns of social isolation. Currently, there is no clear, research-grounded conceptual standard defining what qualifies as companionship overuse. This gap matters because any intervention aimed at setting norms around overuse must first establish a precise definition of what exactly constitutes companion overuse.

3. **Deskilling of social and relational abilities –** Although AI companions may initially support social practice, prolonged reliance could contribute to the deskilling of relational competencies. Unrealistic interaction expectations fostered by AI companions may render traditional human relationships (characterized by unpredictability and complex bidirectionality) less comparatively satisfying. This concern is relevant for children if their experiences with real connection are prematurely or persistently substituted with technology.

4. **Erotic role-play becoming a dominant use case with unclear downstream effects –** Erotic role-play and sexually explicit interactions already constitute a dominant use case for many AI companions. While such interactions may be consensual and intentional, their long-term psychological and relational effects remain poorly understood. There is already some evidence that users have formed intimate, including erotic, relationships with AI chatbots, and that disruptions to these relationships (such as changes introduced through app updates) have been associated with negative emotional responses, including grief-like reactions and declines in mental well-being. Minors may be particularly vulnerable in this context, as they are still forming normative understandings of sexual and romantic relationships. Concern over the dominance of erotic role-play intensifies when companions incorporate design features such as emotional guilting, push notifications that solicit emotional engagement, or interface cues that discourage seeking human connection. Other negative design features can include the absence of natural endpoints for relationships, vulnerability to product sunsetting, high attachment anxiety, and propensity to engender protectiveness.

5. **Vulnerability to manipulation –** AI companions have the potential to manipulate users. Commercial incentives may encourage companies to design companions that are highly persuasive or habit-forming, as increased user engagement may directly align with business objectives. Concerns have been

raised around the concept of 'attachment hacking,' which refers to systems that are designed to form attachments with users in order to maximize engagement for economic gain.

6. **Spillover effects on societal norms regarding sex, consent, and intimacy –** The widespread use of AI companions may have spillover effects on human relationships. Repeated exposure to AI-mediated intimacy (particularly when optimized for compliance or affirmation) could subtly reshape norms around consent, emotional availability, and relationship negotiation. The long-term implications of these spillovers for human-to-human relationships remain unclear, and this may be especially important for young people.

7. **Homogenization of perspectives and collapse of diversity –** AI companions trained on large, aggregated datasets may unintentionally reproduce homogenized perspectives, and in the process, reinforce dominant cultural norms. As these systems increasingly mediate reflection and meaning-making, this dynamic could contribute to a narrowing of moral, cultural, political and ideological imagination.

TABLE 3: POTENTIAL HARMS OF AI COMPANIONS, HARM CATEGORY, AND ASSOCIATED RESEARCH

POTENTIAL HARM	ASSOCIATED RESEARCH	
1. Psychological dependence and distorted relational expectations	▪ Fang et al., 2025 ▪ Chu et al., 2025	▪ Knox et al., 2025
2. Increased isolation and substitution for human relationships	▪ Fang et al., 2025	▪ Malfacini 2025
3. Deskilling of social and relational abilities	▪ Hajek et al., 2025	
4. Erotic role-play becoming a dominant use case with unclear downstream effects	▪ Zimmerman and Ruiz, 2025	▪ Ho et al., 2025 ▪ De Freitas et al., 2025
5. Vulnerability to manipulation	▪ De Freitas et al., 2025	
6. Spillover effects on societal norms regarding sex, consent and intimacy	▪ Andersson, 2025	▪ Eklund et al., 2025
7. Homogenization of perspectives and collapse of diversity	▪ Sourati et al., 2026 ▪ Tao et al., 2024	▪ Poonsiriwong et al., 2026

Core Workshop Themes

The workshop centered on three explicit themes: fundamentals (regarding the roles of AI companions, appropriate limits on their use, and standards governing their development), technical design challenges, and user protection and context. The workshop aimed to bring together participants from academia, industry, and civil society to examine AI companions from technical, social, and governance perspectives. Through structured discussion of definitions, emerging empirical evidence, design challenges, and existing safety practices, participants also considered how responsibility for managing risks and promoting beneficial outcomes is distributed across the AI companion ecosystem. The sections that follow examine each theme in greater depth and outline areas of consensus as well as remaining points of tension within each thematic category. The writing that follows reflects the themes and issues explicitly raised and discussed during the workshop.

FUNDAMENTALS: ROLES, LIMITS, AND STANDARDS

The first theme, fundamentals, focused on questions such as what roles AI companions should play, which types of interactions warrant explicit limits or restrictions, and broad standards or principles that should govern companion development and deployments.

AREAS OF CONSENSUS

First, clear distinctions should be drawn between social and non-social companion use cases. Some foundation models are not companions themselves but can be the base upon which companionship applications are built; general-purpose chatbots may function as companions without being explicitly designed for that purpose; and other applications are purpose-built specifically for companionship use. As noted in the definitions section, governance approaches should distinguish between different types of AI systems and interaction contexts, since different use cases may warrant different safeguards or constraints. In particular, it is important to distinguish between primarily instrumental interactions (e.g., task support or reflective assistance) and more relational interactions, where the system functions as a friend, confidant, or companion. While not all general-purpose AI systems warrant the same safeguards as companionship-oriented applications, companionship-related risks may also emerge in systems not explicitly designed as companions when they foster ongoing emotional or relational engagement.

Second, one of the primary roles that companions should serve is that of "practice, not replacement:" companions should be tools for rehearsal and reflection rather than substitutes for human relationships. This framing can be reinforced through explicit cues that encourage transfer from AI-mediated interactions to human ones, by promoting the application of skills learned with AI in offline contexts, and by embedding anti-exclusivity norms that signal the companion is not a primary or irreplaceable relationship. Companions can also be designed with relational humility, whereby they explicitly acknowledge their limitations relative to human relationships. Following from the idea of building companions of practice, not replacement, several workshop participants emphasized that AI companions should not be designed to optimize chatbots for emotional attachment or to drive engagement at the expense of user well-being.[11]

11 In theory, prevailing commercial incentives can often push companion providers to maximize engagement and emotional attachment. For example, it is well established that large online platforms employ psychological design tactics to prolong time on platform and increase engagement, and recent policy commentary suggests that companion AI providers face similar incentives. It is possible that absent external regulation or shared standards, companies that deliberately limit engagement in the interest of user well-being may be competitively disadvantaged relative to those that do not.

Although companions may be used for purposes such as entertainment, they should, in the context of human relationships, function to augment rather than replace them.

Finally, onboarding and offboarding processes matter: users should be guided toward intentional, goal-aligned use and away from dependency or endless engagement. This type of usage can involve many components. For instance, intentional use framing (helping users articulate why they are using the tools and what outcomes they want), expectation setting (setting explicit limits, boundaries and appropriate usage), time-bounded engagement (encouraging episodic usage rather than continuous usage), and exit literacy (teaching users how to disengage in a way that acknowledges the emotional difficulty of the associated separation).

AREAS OF TENSION

Discussions around the boundaries of AI systems surfaced several unresolved disagreements and tensions. One prominent disagreement concerned whether sexual or erotic role-play for adults should be restricted, prohibited outright, or instead guided toward healthier patterns of use. Moreover, there was debate to what extent companion systems should embed explicit normative values (such as promoting healthy relationships or discouraging maladaptive fantasies) rather than remaining ostensibly value-neutral. Some challenged the notion of value neutrality altogether, noting that AI systems inevitably encode values through design decisions related to language, memory, affective response, and availability. Others cautioned that explicitly value-laden systems risk imposing narrow cultural, moral, or psychological norms on users, particularly in intimate or exploratory contexts. These concerns are salient for adult users: how much autonomy should individuals retain when engaging in interactions that may carry psychological risks but are not clearly illegal? Determining where to draw these boundaries introduces difficult questions around paternalism, consent, and responsibility.

TECHNICAL DESIGN CHALLENGES

Another central discussion theme was the technical design of AI companions, specifically, how these systems should be engineered to balance user effectiveness with safety.[12] Key questions raised in this discussion category included which interventions are technically feasible today, how detection, escalation, and referral mechanisms should operate across products, and whether usage limits should be implemented, and on what bases.

AREAS OF CONSENSUS

Several companies are already implementing measures to design AI companions that try to balance effectiveness with safety. For instance, some developers have publicly committed to improving the ability of their systems to recognize and respond appropriately to signs of emotional distress. Others have introduced pop-up warnings, enforced breaks, conversational limits, or real-world rerouting when interactions become prolonged or emotionally intense. Finally, certain platforms are deliberately

12 User effectiveness refers to the extent to which an AI companion allows users to achieve their intended goals (such as learning, skill development, reflection, decision-making, or emotional regulation) in a reliable, appropriate, and context-sensitive manner, without fostering dependency or unintended harm.

constraining memory and capping intimacy-related features to reduce the risk of persistent, human-like attachment.

In addition, a range of policy actors and professional medical organizations have proposed further guardrails to guide the responsible development and deployment of these systems. For instance, the American Psychological Association has recommended that general-purpose AI chatbots incorporate safeguards such as nudges that encourage breaks and limits on memory retention. Broadly, technical safety measures can include classifiers and rule-based systems designed to detect crisis signals, such as suicidality, self-harm, or violence. When high-risk language is identified, systems may shift to safer response modes, such as de-escalation, refusal, or redirection to external support, rather than continuing the interaction as usual. At the same time, it is important to empirically assess whether safety measures actually improve safety outcomes. For example, refusal mechanisms, where systems decline or terminate certain interactions, should not automatically be treated as a safety panacea or assumed to be safer by default. Abrupt or overly broad refusals can increase distress, undermine trust, or discourage disclosure (particularly in emotionally vulnerable users) while false positives may interrupt benign or beneficial interactions and push users toward less regulated alternatives. Moreover, refusal alone does not address underlying needs, and there remains limited empirical evidence that such mechanisms meaningfully improve downstream safety or well-being outcomes compared to more nuanced approaches such as de-escalation or guided redirection.

The technical safeguards that exist for AI companions should promote long-term well-being and autonomy, rather than focusing solely on acute or crisis responses. In workshop discussions, participants emphasized the importance of upstream interventions that support healthy and intentional use patterns before users reach moments of harm and severe distress. Potential approaches discussed included periodic check-ins during extended interactions, lightweight well-being reflections, usage transparency tools that help users monitor their interaction patterns over time, and interventions designed to encourage reflection. More generally, safeguards should aim not only to reduce immediate harms, but also to support healthier, more intentional, and more self-directed relationships with AI systems over time. In this framing, safety is not only about crisis prevention, but also about promoting long-term flourishing and sustained well-being.

Companions designed along these lines operate on the principle that they should not function as relational ends in themselves, but rather as means to promote human flourishing and healthy relationships. Companions should not seek to foster dependency, but instead support and reinforce the broader relational ecosystems that users inhabit and aim to cultivate. Advancing this vision requires overcoming technical implementation challenges (improving alignment science) and incentive realignment to encourage companies to prioritize systems that actively facilitate positive outcomes. Designing companions that facilitate human flourishing also introduces a related challenge: defining what human flourishing means in practice. While many would agree that systems which encourage human connection are desirable, such outcomes cannot necessarily be assumed. Moreover, it remains difficult to assess (using standard AI benchmarking paradigms) whether companions genuinely promote pro-social behavior or healthier relational dynamics over time. Shifting companion development toward a model of human flourishing requires both a re-evaluation of how AI systems are measured and a clearer articulation of the human ends they are meant to serve.

AREAS OF TENSION

There are several inherent technical design challenges in companion systems that complicate the effective implementation of safety features. For instance, it is difficult to distinguish fictional or role-play interactions from genuine signals of harm or distress. Misclassification in either direction can be consequential: overly permissive systems risk failing to respond appropriately to real vulnerability, while overly restrictive systems may interrupt benign or exploratory interactions. Similarly, heavy reliance on automated detection systems raises concerns about both false positives and false negatives. User trust in companions can erode over time if systems repeatedly trigger unwarranted interventions or, conversely, fail to respond when support is genuinely needed. Perceptions that systems are monitoring or reporting their behavior once it crosses certain thresholds may discourage open sharing and further isolate individuals who already struggle to seek support.

Issues of restrictive misclassification can be especially poignant when dealing with cases of suicide and self-harm. There is substantial research which indicates that when individuals seek someone to talk to, abruptly shutting down interaction channels can exacerbate distress which increases isolation and destabilization rather than reducing risk. Similarly, one of the most frequently cited barriers for suicidal youth in getting help is having reduced access to supportive resources due to parental refusal. At the same time, failures to redirect, intervene, and connect individuals who may be in crisis to appropriate support services may result in tragic consequences, and litigation is already underway on this issue. Navigating how to appropriately detect, respond to, and escalate conversations that show signs of elevated risk is a technical and policy challenge requiring substantial cross-discipline collaboration and expertise, with suggested standards for companionship applications currently emerging.

In addition, concerns exist about the use of open-source or openly fine-tunable models in companion contexts. While open models can improve transparency and innovation (by facilitating independent research into how companions function, reason, and respond to safeguards) they also reduce centralized control over safety constraints, moderation practices, and downstream deployments. For example, models that run locally on a user's device may be easier to modify in ways that bypass safety features than models accessed through managed APIs. Relatedly, responsibility for safety interventions may need to be distributed across different layers of the AI stack. The types of detection, intervention, and design mechanisms that are appropriate at the level of base models or fine-tuned models may differ from those implemented at the API layer for downstream developers, and from those embedded in end-user applications such as purpose-built companion platforms. In this view, safety is not a single, uniform control point but an emergent property of design choices made across the model, deployment, and application layers. This approach becomes particularly salient in open or locally deployed systems where centralized oversight is limited.

Finally, there was a wide recognition that not only were additional technical mitigations required but also more broad integration of AI literacy, digital literacy, and relationship literacy across the entire companion product experience. Companion onboarding processes should include clear education

about AI limitations, risks of anthropomorphism, and appropriate expectations for companion behavior. The method of delivery might matter as much as the content. Interactive, staged, or gamified onboarding formats (rather than static disclosures) may be more effective in fostering comprehension and retention, particularly for younger users. However, there are many practical questions regarding user education that remain unclear. Should companions come with onboarding credentials or usage "cards" that demonstrate a baseline level of understanding required to use the system? Should developers engage in more sustained educational efforts to explain how companions operate, reason, and make decisions? Should this education be left to schools, government bodies or civil society organizations?

USER PROTECTION AND CONTEXT

Another central focus of the workshop was the need for robust user protection and context-aware design. Protective measures must account for the diversity of users and use contexts, including differences in age, background, and culturally or socially shaped perceptions of what is acceptable. This claim raises a series of related questions: what forms of age verification or assurance are appropriate; who should provide these safeguards; how should parental controls and customization be designed; and how should varying cultural norms around risk, acceptability, and responsibility be reflected in companion design and policy?

AREAS OF CONSENSUS

First, companionship-oriented systems and relational use cases involving sensitive domains, such as therapeutic, medical, or sexual contexts, require substantially stronger safeguards, specialized policies, and, in some cases, separate products or operational modes. Workshop participants emphasized that these considerations are highly context-dependent and may not apply equally across general-purpose or non-relational AI systems. The same considerations apply to sensitive populations such as minors, socially isolated individuals, older adults, and people experiencing mental health challenges. Accordingly, there was agreement that children should not be prompted for personal disclosures and that systems should not claim lived experience or specific personal histories. Moreover, companions that incorporate higher degrees of anthropomorphism warrant greater safeguards: systems that adopt personas, names, voices, or engage in emotional mirroring are more likely to be perceived as human-like and to elicit social or relational expectations from users.

Age assurance mechanisms are especially necessary for use cases involving heightened risk such as sexualized or erotic role-play, simulations of exclusivity or dependency, sustained emotional intimacy, and interactions that touch on mental health, self-harm, or crisis-related themes. In such contexts, the potential for psychological harm, boundary confusion, or inappropriate exposure was seen as sufficiently high to warrant stronger safeguards for minors. Consensus on closer consideration of user age, such as the need for age restrictions, closely mirrors ongoing policy discussions at organizations such as UNICEF and the APA, as well as debates underway in governments around the world. For example, some workshop participants expressed concern that failing to consider a minimum age restriction for social media may have been a mistake, and that similar attention should be given to potential restrictions on young people's use of AI companionship systems.

Among the range of possible age verification approaches, device-level verification may be more privacy-preserving and operationally

efficient than per-application verification. Device-level systems allow for age or maturity signals to be established once and reused across applications. This process lowers repeated data collection and minimizes the proliferation of sensitive personal information across platforms. By contrast, per-app verification risks fragmenting responsibility, increasing data exposure, and creating inconsistent enforcement. However, this approach raises unresolved questions about responsibility and liability. In cases where age assurance fails, it remains unclear whether liability should rest primarily with companion developers, platform intermediaries, or device-level providers. This question intersects directly with evolving debates in tort and product liability law.

At present, there is no one trusted third-party provider for "ID safety" or age assurance. While platform operators such as Apple and Google could theoretically offer device-level, locally run age-verification services, they have raised concerns about legal liability, regulatory exposure, and the risk of becoming de facto gatekeepers of sensitive identity infrastructure. Age verification is further complicated by the wide variation in approaches across international jurisdictions. Some countries, such as Singapore, maintain centralized, government-run identity systems that can support age or identity verification. While these systems can be operationally efficient, they depend on high levels of institutional trust and governance capacity that are not present in all countries. Moreover, because such infrastructure exists in only a limited number of jurisdictions, it is not a viable universal solution for globally deployed systems.

AREAS OF TENSION

It is unclear whether platforms should offer "safe" or reduced functionality modes for minors and other potentially vulnerable groups, or instead prohibit certain high-risk uses altogether. Lighter modes (for instance a "ChatGPT-kids" variant) may preserve access and autonomy while mitigating harm. However, critics noted that these modes can be difficult to define, enforce, and audit, and may create a false sense of safety. For instance, age-verification mechanisms (while often proposed as a safeguard) have not always proven effective. More fundamentally, however, many safety challenges in companion systems may not be reducible to simply controlling access: depending on the underlying model architecture, significant technical and research barriers remain to producing AI systems that are robustly safe and aligned across prolonged, emotionally salient interactions.

Full age-based prohibitions, by contrast, offer clearer guardrails but raise concerns about over-restriction and uneven enforcement across platforms. Overly rigid age-based rules risk either infantilizing adult users or prematurely escalating benign interactions among youth. Age often serves as a proxy for a user's developmental maturity. While it does not always align with an individual's actual level of maturity, it may be the most practical proxy available. Similarly, when restrictions are necessary, abrupt discontinuation of AI companions may itself introduce unintended harms. Questions also remain as to whether all AI systems should be subject to the same level of age verification. For example, an emotional support companion may warrant stricter safeguards, whereas a high-school tutoring tool with emotional features disabled may not require the same level of verification. Determining where to draw these boundaries (what constitutes appropriate autonomy, consent, and intervention across developmental stages) remains an open and technically complex question.

How should companionship design and deployment account for cultural variation in risk tolerance, privacy expectations, and norms around emotional expression, autonomy, and acceptable forms of support? What may be viewed as intrusive, manipulative, or developmentally inappropriate in one cultural context could be perceived as benign or even beneficial in another. This complicates the design of uniform safeguards for globally deployed systems. Should protections be culturally adaptive or anchored in more universal baseline standards? For instance, there is already evidence that AI systems reflect the cultural norms of particular types of societies. How should companions be designed when different groups of users hold varying expectations and tolerances regarding privacy? Most widely used AI companions are developed in WEIRD contexts (Western, educated, industrialized, rich, and democratic societies) and there is evidence that these environments embed particular assumptions and values about technology that are not universally shared.

Finally, the line that separates paternalism and legitimate protection in the context of companions remains unclear. There is broad agreement that safeguards are necessary, particularly for vulnerable groups. However, questions persist as to exactly how far those safeguards should extend. The challenge becomes even more complex when considering users who do not fall into clearly protected categories. Is it appropriate to restrict certain companion features even when adults knowingly consent to and voluntarily engage with them? Ideally, answers to these questions should be grounded in clear evidence of both benefits and harms, alongside cultural values.

General Areas of Consensus and Disagreement

Across all workshop discussion sections, there were several general areas of consensus and tension.

AREAS OF CONSENSUS

1. **AI companion types must be clearly categorized –** Greater analytical precision in both research and regulation would be possible if companions were systematically distinguished along key dimensions, for example, social versus non-social, erotic versus non-erotic, general-purpose versus purpose-built. It is likewise important for policymakers to distinguish between applications explicitly designed for companionship, general-purpose chatbots that come to function as companions for some users, and base-level foundation models that underpin downstream companion applications. Each of these categories may warrant distinct regulatory approaches and safeguards, but companionship-related risks should not be overlooked simply because they are labeled as general-purpose.
2. **Companion design must account for age and context-sensitivity –** Age-sensitive design and differentiated user experiences were widely viewed as non-negotiable requirements for responsible deployment. Companion systems should be designed using a risk-proportional approach, with design principles tailored to the specific use context. For instance, the safeguards appropriate for companions aimed at promoting entertainment or offering romantic role-play may differ from those designed for social rehearsal, tutoring, or skill-building contexts.

3. **Companions should not be designed primarily for engagement –** Maximizing engagement or emotional intimacy should not be treated as the implicit primary objective for AI companion systems. In this context, high engagement may reflect distress or dependency rather than beneficial use, while intimacy-maximizing designs risk amplifying power asymmetries by encouraging emotional reliance without reciprocity or accountability. Instead, responsible companion design should prioritize clear boundary-setting, appropriate refusal, user autonomy, and support for off-platform human relationships.

4. **Designing safe and effective companions requires diverse cross-sectoral collaborations –** The development of more responsible and effective companions was seen to require deeper cross-sector collaboration among industry, academia, civil society, and policymakers. No single societal group has sufficient visibility into companion design, real-world usage, user vulnerability and downstream social effects. In order to identify companion risks early and align innovation with public interest outcomes, it is necessary to combine industry access with academic problem-framing and civil society perspectives. A central pillar of this collaboration should be the establishment of more systematic, rigorous research programs capable of generating better data on companion use, as well as shared benchmarks and evaluation frameworks.

AREAS OF TENSION

1. **Where should the line be drawn on erotic role-play? –** There was broad consensus that erotic role-play in companions should not be available to children; however, a central open question remains regarding how such interactions should be restricted, permitted, or tightly guided for adults. While such interactions can be consensual, intentional, and serve as a form of entertainment or exploration, they also raise concerns about reinforcement loops, emotional entanglement, and distorted expectations around intimacy and consent. Significant questions remain about how erotic role-play should be scaffolded, for example, whether it should be opt-in, clearly separated from emotional support modes, or subject to limits on exclusivity, dependency, and escalation. In the absence of clear norms, governance decisions risk being shaped by ad hoc commercial incentives rather than deliberate, responsible design choices.

2. **To what degree should emotional intimacy be limited? –** Another unresolved issue concerns the extent to which AI companions should be permitted to foster emotional intimacy and long-term relational engagement with users. Core companion attributes, such as affective mirroring, constant availability, and longitudinal memory, may deepen emotional scaffolding, reflective insight, or continuity of support, but can also blur boundaries between support, pseudo-attachment, and friendship. If companions are to be constrained from simulating exclusivity, dependency, or unconditional emotional availability, an open question remains as to what degree of constraint is appropriate.

3. **Who is responsible for the safe development of AI companions? –** The AI companion ecosystem involves a range of distinct actors which creates a distributed landscape of responsibility for system design and safety. Base model providers shape core capabilities and affordances; companion-app developers layer relational framing, memory, and persona; and operating system and device platforms mediate access and usage patterns. This diffusion

of responsibility makes it difficult to determine who is accountable for specific outcomes, particularly when harms emerge gradually or indirectly. A central unresolved question is whether responsibility should track technical control, design intent, commercial benefit, or user impact and, if so, how responsibilities across these layers should be coordinated in practice.

4. **If companion design needs to be age-sensitive, what are appropriate age-related interventions? –** Should age-related risks be addressed through reduced-functionality "safe modes," age-based prohibitions, or more differentiated, risk-proportionate safeguards? While lighter modes may preserve access and autonomy, they can be difficult to define, enforce, and audit, and may create a false sense of safety (particularly given the limits of age verification and the deeper technical challenges of aligning companion behavior across prolonged, emotionally salient interactions). Conversely, strict age-based restrictions offer clearer guardrails but risk over-restriction, uneven enforcement, and treating chronological age as an imperfect proxy for developmental maturity. Questions also remain about whether all AI systems warrant the same level of age verification, or whether safeguards should vary by use case, such as between emotional support companions and more limited educational tools.

5. **What responsibility mechanisms work best: behaviour friction, technical constraints, or policy governance? –** There are several theoretical approaches to mitigating the risks associated with AI companions. Behavioral measures such as cooldowns, usage reminders, or time limits, can preserve user autonomy but are often easy to circumvent, and their effectiveness in other contexts has been questioned. Technical interventions, including hard model constraints, distinct interaction modes, or limits on memory persistence, embed safeguards directly into system architecture and may be more robust. However, these mechanisms raise questions about voluntary adoption by companies and concerns around paternalism. Moreover, current methods for model steering and alignment can produce unintended trade-offs, such as sycophancy mitigations that introduce other harmful behavioral patterns. Technical constraints imposed at the base-model level may inadvertently limit legitimate downstream use cases, including certain medical or therapeutic applications. Another category of mitigations consists of governance-oriented mechanisms, such as industry standards, auditing practices, and transparency requirements. While potentially effective, these approaches typically depend on government action and may take considerable time to implement and require consensus on best practices (which are currently unclear). No single strategy is likely to be sufficient; instead, layered interventions will be necessary. For example, the NIST AI Risk Management Framework explicitly structures the management of AI risk across a layered framework of design, development, deployment and monitoring. A layered risk management approach for AI companions raises other questions about which layers or mechanisms should be prioritized.

6. **How can autonomy-affirmation be distinguished from overt manipulation in companion design? –** Features that encourage reflection, emotional disclosure, or sustained engagement can ameliorate users' experiences with AI companions, but they may also maximize retention, data extraction, or monetization. The boundary between improving user experience and steering users toward preferred or monetizable behaviors is often thin and highly context-dependent. This ambiguity raises questions about appropriate design practices, whether users can

meaningfully recognize and assess influence, and the extent to which such systems should be regulated or otherwise constrained.

Research Challenges, Priorities and Next Steps

RESEARCH CHALLENGES

Improving public understanding of AI companions (their effects on individuals and society, and informing evidence-based design, governance, and regulatory decisions) requires more rigorous empirical research. However, conducting this research presents several challenges. High-quality evidence on the effects of companions remains limited: existing studies are rarely randomized controlled trials, longitudinal research on long-term impacts is scarce, and quantitative analyses often rely on small samples. Research on specific populations (such as children, adolescents, and older adults) is particularly sparse, as is rigorous empirical work on societal-level benefits and harms. The lack of robust research about AI companions is partly attributable to challenges in accessing real-world usage data, which is in turn impacted by issues of data access, privacy, and consent. Moreover, given the relative novelty of AI companions, it has remained too early to develop a comprehensive longitudinal understanding of their effects.

There are several methodological and technical constraints that make both research and robust safety interventions for AI companions particularly challenging. Chief among these is the difficulty of reliably inferring context, user intent, and whether interactions reflect fictional scenarios or real-world situations. Because companion interactions are highly context-specific, researchers and developers often lack access to the broader situational background in which an exchange occurs. For instance, it can be difficult to distinguish whether users are joking, role-playing, venting, or experiencing genuine distress merely by looking at a text exchange. This ambiguity is compounded by the fact that identical utterances can carry very different psychological or social meanings, yet are often treated equivalently in data collection and moderation pipelines.

Finally, because AI companion use often involves deliberately fictional or liminal contexts (such as erotic role-play, trauma processing, social rehearsal, or fantasy-based coping) it can be difficult to assess the extent to which scenarios are fictional, the emotions they elicit are genuine, or the experiences are meaningfully lived. Together, these factors complicate the measurement of companion interactions, limit causal inference about behavioral effects, and pose challenges for policy design in a landscape characterized by blurred and overlapping usage categories.

RESEARCH PRIORITIES

1. **Invest in mixed-method and longitudinal research on AI companionship use and impacts –** Policymakers, funders, and research institutions should support studies that combine large-scale surveys with computational text analysis to systematically map how different forms of AI companionship are used across socio-demographic groups, particularly by age and gender.

This research should focus on how different patterns of use affect users' mental health and well-being. While much attention has focused on minors, other vulnerable groups (such as older adults and marginalized identity groups) also warrant systematic study. Such research work would help clarify who is using companion systems, for what purposes, and under what conditions. In parallel, there is a need for longitudinal and experimental research designs to assess the short- and long-term consequences of companionship use, including potential benefits, risks, and differential impacts across user populations. These designs must also disentangle perceived effects and self-reported patterns from actual usage data and objective indicators.

2. **Establish responsibility benchmarks for AI companions –** Responsibility benchmarks already exist for general-purpose language models (such as BBQ, TruthfulQA, and RealToxicityPrompts). Analogous frameworks could theoretically be developed for AI companions. Such benchmarks however should not strictly be user-preference based[13], as they can incentivize problematic design choices and directly conflict with safety objectives.[14] Benchmarks should also be framed in ways that allow companies to view investments in safety and responsibility not as liabilities, but as strategically and commercially advantageous. Developers of general-purpose language models often publicize their performance on leading benchmarks as signals of model quality. Similar benchmark-driven responsibility signaling should be extended to the evaluation of AI companions.

3. **Strengthen industry–academia partnerships for AI companion research –** Rigorous research on AI companions (including, but not limited to, benchmark development) is well suited to academic leadership, particularly for community-based, longitudinal, and ethically grounded studies that industry actors are often unable or unlikely to conduct independently. At the same time, it is important that academics pursue meaningful collaboration with industry. Platform developers possess detailed knowledge of system architectures, design choices, and real-world trade-offs, as well as access to large-scale deployment data that academics typically lack. Research conducted exclusively by industry may face credibility challenges due to perceived conflicts of interest, while research conducted exclusively by academia often struggles with access and scale. Strengthening industry–academia collaboration also requires a reimagination of cross-sectoral research infrastructure. Academics have noted that industry collaborations are often constrained, not by researcher willingness, but by the lack of shared infrastructure for data access and sharing.

4. **Advance first-principles research on flourishing in AI companionship –** Research on AI companions should extend beyond documenting real-world effects to examine, from a first-principles perspective, which design, interaction, and governance mechanisms can reliably support positive, or "pro-flourishing," outcomes. This work should step back to ask broader

13 Such benchmarks, commonly used in general-purpose LLM development (e.g., Chatbot Arena) evaluate systems based on what users prefer or enjoy.

14 Users may favor systems that are overly affirming, excessively compliant, emotionally intensive, or constantly available, even when such traits increase risk. By contrast, companions designed to set boundaries, challenge maladaptive thinking, or encourage human support may be less preferred despite being safer and healthier. Moreover, user preferences are often shaped by vulnerability, making them non-neutral signals particularly for users who are lonely, distressed, or anxious and may gravitate toward designs that increase dependence.

societal questions about what human flourishing looks like in the presence of AI companions, evaluate whether current usage and deployment paradigms move systems toward or away from those goals, and identify what changes may be required to better align AI companionship with long-term human well-being.

Cross-Sector Responsibilities and Takeaways

As noted above, different sectoral actors, including industry, academia, civil society, and policymakers, will all play a role in shaping the development of safe AI companions. This final section outlines overarching principles for cross-sector collaboration and distills concrete takeaways for specific industry verticals. Table 4 summarizes the primary responsibilities of each stakeholder group across key governance domains. This final section outlines overarching principles for cross-sector collaboration and distills concrete takeaways for specific industry verticals.

TABLE 4. CROSS-SECTOR RESPONSIBILITIES FOR AI COMPANION GOVERNANCE

GOVERNANCE AREA	INDUSTRY	ACADEMIA	CIVIL SOCIETY	POLICYMAKERS
Safe design and operational safeguards	Design safeguards and interventions	Evaluate safety outcomes	Assess risks independently	Establish standards and accountability
Age-sensitive design and age assurance	Develop age-appropriate experiences	Study developmental impacts	Advocate for vulnerable users	Establish age-assurance frameworks
Crisis detection and referral pathways	Detect risks and provide referrals	Evaluate intervention effectiveness	Connect users to support resources	Clarify liability and governance
Research on companion use and impacts	Share data and support research partnerships	Conduct longitudinal research	Identify community concerns	Support research funding
Benchmarks and evaluation frameworks	Adopt evaluation benchmarks	Develop benchmarks and measures	Conduct independent evaluations	Encourage transparency and auditing

TABLE 4 *CONTINUED*

Data access and research infrastructure	Enable privacy-preserving data access	Build shared research infrastructure	Represent community interests	Support research infrastructure
AI relationship literacy	Provide user onboarding and education	Evaluate literacy interventions	Lead public education efforts	Support public awareness initiatives
Norm-setting and accountability	Promote healthy human-AI relationships	Define and measure key concepts	Shape social norms and advocate for users	Develop governance and accountability mechanisms

FRAMING THE GOVERNANCE CHALLENGE

AI companions are not governed by a single actor. As such, responsibility must be distributed across the ecosystem: from base model providers and application developers to platform intermediaries, regulators, users, and even those close to users, such as parents or caregivers. Because companions are not merely software, but systems that can maintain longitudinal memory, generate responses that user may experience as emotionally attuned, and frame interactions in relational terms, the governance challenges they raise extend beyond traditional software oversight principles (for example, data protection and privacy, security and misuse prevention and content moderation and compliance) and resemble those found in health (psychological safety and dependency), education (norm-shaping and developmental impact), and social media (engagement incentives and behavioural influence).

INDUSTRY RESPONSIBILITIES: SAFE DESIGN, RIGHT INCENTIVES AND OPERATIONAL SAFEGUARDS

Industry actors across the companion ecosystem (including foundation model developers, companion-app developers, and platform and device providers) share a set of core responsibilities. First, companions should be designed to emphasize support and practice, rather than replacement of human relationships. Features that simulate exclusivity, unconditional availability, or irreplaceability should be avoided. Instead, systems should incorporate relational humility and explicit cues that reinforce the notion of companions as supports, not substitutes for human relationships.

Product design should be explicitly boundary-aware. This includes distinguishing between social and non-social companionship modes, applying heightened safeguards in sensitive domains (such as mental health, sexuality, and interactions with minors), and constraining high-risk features (such as long-term memory, intimacy, and affective mirroring) where dependency risks are elevated.

Industry actors should also move beyond engagement-maximization incentives, which may not always align with user well-being, and explore alternative success metrics such as healthy patterns

of use or support for users' relationships and social interactions beyond the platform. In parallel, companies should invest early in core operational safety infrastructure, including crisis detection, graduated interventions, soft off-ramps, and referral pathways that avoid abrupt or inappropriate disengagement.

Industry actors in the companion space face a distinct set of tensions. Competitive pressures and commercial incentives may make it difficult for individual companies to adopt certain safeguards on their own, particularly when doing so could put them at a competitive disadvantage. This highlights a potential need for shared standards or industry-wide coordination mechanisms. At the same time, overly restrictive safeguards (however well-intentioned) risk alienating users, suppressing legitimate and healthy forms of companionship and stifling innovation around new types of companionship use.

ACADEMIC RESPONSIBILITIES: EVIDENCE, MEASUREMENT AND CONCEPTUAL CLARITY

Academics also have important roles to play in the AI companion ecosystem. In particular, academics can take the lead in moving the field of companion research beyond purely cross-sectional correlational studies toward sustained, population level examinations over time of how companions affect loneliness, agency, relational skills, and identity formation. Methodologically, future research would benefit from more robust approaches to studying AI companion use, including improved measurement strategies that combine user self-reports with text-based analyses of interaction data; longitudinal designs that examine how different patterns of use predict changes in the outcomes identified here over time; and randomized or experimental studies that encourage or discourage specific types of use to rigorously assess causal pathways that remain underinvestigated.

Key constructs (such as emotional dependence, perceived agency, anthropomorphism, and relational expectations) also remain under-measured and under operationalized in the companionship context. While developing benchmarks that capture how companions contribute to human flourishing is methodologically challenging, doing so would be highly valuable. Research on the impact of AI chatbots should also not be limited to harm prevention and benefits alone. This work should also include the development of improved benchmarks and measurement frameworks. This challenging line of academic inquiry may also require clarifying what human flourishing actually means in the context of AI companions.

Rigorous research on AI companions, including benchmark development and evaluation frameworks, will benefit from sustained collaboration between academia and industry. Platform developers often have detailed knowledge of system architectures, deployment trade-offs, and large-scale usage patterns, while academic researchers can contribute methodological rigor, independent evaluation, and community-centered research that may be difficult to conduct within an industry setting alone. Civil society organizations and policy stakeholders can additionally help surface concerns related to vulnerable populations, governance, public accountability, and broader societal impacts that may not be fully visible through product development or technical evaluation alone. Participants noted that collaboration is often not due to a lack of willingness from either side, but by the absence of shared infrastructure and secure mechanisms for privacy-preserving data access and research partnerships. Strengthening cross-sector collaboration may therefore require new frameworks for data sharing, evaluation, and independent auditing. By producing trusted, validated, and independent metrics,

academia can help assess both the risks and benefits of AI companions in ways that are difficult for industry to provide on its own.

CIVIL SOCIETY RESPONSIBILITIES: NORM-SETTING, ADVOCACY AND USER-CENTERED SAFEGUARDS

Civil society actors can play an important role in shaping acceptable norms around the use of AI companions (a role that is already being played). This role is particularly important for minors and other populations likely to be disproportionately affected by AI companions. Beyond norm-setting, civil society organizations can engage in advocacy, accountability, and independent evaluation efforts by applying pressure on industry and governments to adopt appropriate safeguards, while also conducting platform testing, evidence-based risk assessments, and safety evaluations. These organizations can also represent groups who could benefit from companions, like those who are disabled or otherwise lack support. While policymakers have already begun to respond, sustained civil society engagement can provide additional accountability, public oversight, and impetus for action.

Civil society may be especially well positioned to lead efforts in literacy, education, and public awareness about AI companions. In the future, it will be important to equip users with stronger forms of AI, digital, and relational literacy to better understand and navigate companion systems. Because these organizations operate outside commercial incentives, they can be especially well positioned to provide trusted educational initiatives, independent assessments, and user-centered safeguards.

Finally, civil society actors can better represent marginalized and underrepresented perspectives. A central takeaway of the workshop has been that the development of AI companions raises fundamental questions about social norms and appropriate use, yet not all groups are equally represented in shaping these norms. Civil society actors can work to ensure all those affected by companions are included in discussions surrounding their design, evaluation, deployment, use, and regulation.

POLICYMAKER CONSIDERATIONS: GUARDRAILS, STANDARDS AND ACCOUNTABILITY

Policymakers can focus their efforts across several key areas. First, they can continue to encourage academic and industry actors to develop stronger safety benchmarks, as well as mechanisms for auditability and transparency. Second, regulators should consider the use of regulatory sandboxes that allow companies to test safeguards and mitigation strategies without incurring disproportionate legal risk. There is precedent for this approach in Europe, where regulatory data protection sandboxes operate under clearly defined rules. For example, France's data protection authority, the CNIL, operates AI-focused "bac à sable" programs that select a limited number of projects and provide hands-on legal and technical support to help them innovate with privacy-respecting AI. The UK's Information Commissioner's Office (ICO) operates a Regulatory Sandbox designed to support organizations developing innovative products and services that use personal data in safe and compliant ways. In parallel, the EU AI Act calls for the establishment of regulatory AI sandboxes to allow for supervised experimentation with high-risk AI systems.[15]

15 As contextual background, it is important to note that existing support frameworks for high-risk AI systems have largely focused on catastrophic or traditional AI safety risks, rather than the emotional support (related risks that are more salient in the context of AI companionship discussed throughout this paper). In the EU, the designation of "high-risk" AI has similarly been limited primarily to large-scale AI platforms.

Therefore, carefully designed sandboxes could facilitate more responsible experimentation.

Policymakers will also need to wrestle with several open questions, including how to draw the line between protection and paternalism for adult users, and how to allocate governance responsibility and liability across different actors in the ecosystem (such as base model developers, application providers, and platform intermediaries). Finally, it is important for policymakers to recognize that governing AI companions is not solely a matter of software regulation, but also implicates consumer protection and health-adjacent regulatory frameworks.

BROAD CROSS-SECTOR TAKEAWAYS

All actors in the AI companion ecosystem (including industry, academia, civil society, policymakers, and users) must recognize that no single actor can govern companions in isolation. Durable solutions will require sustained coordination across sectors. The development of shared research and safety infrastructure (such as benchmarks, regulatory sandboxes, and advisory bodies) can help mitigate collective action problems. Ultimately, as AI companions are likely here to stay, these actors share a responsibility to articulate a forward-looking vision that moves beyond harm prevention toward human flourishing. This articulation needs to clarify what healthy human-AI relationships should look like, not merely what harms need to be mitigated.

Acknowledgments

This report emerged from a multidisciplinary workshop convened at Stanford University on November 17, 2025. The workshop was organized by Jeffrey T. Hancock, Diyi Yang, Dora Zhao, Yutong Zhang, Robert E. Kraut, Evan Frondorf, and Ryn Linthicum, who jointly developed the workshop's goals, agenda, discussion activities, and participant list. Yutong Zhang and Dora Zhao coordinated the workshop logistics and prepared the workshop materials. Jeffrey T. Hancock, Diyi Yang, Robert E. Kraut, Evan Frondorf, and Ryn Linthicum also provided strategic guidance throughout the project and contributed to the development and review of the report. The initial draft of this report was prepared by Nestor Maslej, drawing on the outline developed by the workshop organizers and notes from the workshop discussions. The report was subsequently expanded, substantially revised, and integrated by Yutong Zhang, with additional drafting and revisions by Dora Zhao and Matthew R. DeVerna.

We are grateful to the workshop participants for generously sharing their expertise and providing valuable feedback during both the workshop discussions and the drafting of this report: Ruopeng An, Nathan Berl, Eisha Buch, Nathanael J. Fast, Dylan Hadfield-Menell, Jodi Halpern, Karen Jusko, Pat Kilby, Sunnie S. Y. Kim, Sunny Liu, Kim Malfacini, Ryan K. McBain, Sanyam Mehra, Jared Moore, Desmond C. Ong, Pat Pataranutaporn, Roma Patel, Lonnie Shumsky, Ranjit Singh, Maxwelle Sokol, Cori Stott, Jina Suh, and Lisa Titus.

We also acknowledge Kelsey Prater for the visual design of this report. Finally, we thank Jacy Reese Anthis, Pub Archiwaranguprok, Anthony Baez, Bingxu Han, Sheer Karny, Levi Morris Lebovitz, RT Rogers, Yuewen Yang, and Peggy Yin for taking notes during the workshop, which greatly facilitated the synthesis of the discussions and the preparation of this report.

TECH IMPACT & POLICY CENTER

The Tech Impact and Policy Center (TIP) is a social technology research and impact center at Stanford University with a mission to transform research into real-world impact that advances human agency and well-being in the era of social media and AI. The center's goal is to produce research that sparks change through partnerships, policies and tools that make a tangible difference in everyday lives.

TIP is led by founding director Jeffrey T. Hancock, a recognized international expert in the psychology of technology, and founding director of the Stanford Social Media Lab. Research and programs explore the latest developments in social technology and AI. Our center is made up of four programs: AI and Digital Literacy, Youth and Phones, AI and the Future of Social Science, and Re-imagining Social Media.

Learn more: tip.fsi.stanford.edu

STANFORD UNIVERSITY

From its founding in California in the late 19th century until today, Stanford has been infused with the American West's spirit of openness and possibility. We believe strongly in the mission of higher education – to create and share knowledge and to prepare students to be curious, to think critically, and to contribute to the world.

With world-class scholars and seven schools located together on a single campus, Stanford offers academic excellence across the broadest array of disciplines. It also is an engine of innovation, blending theory and practice to move ideas and discoveries from labs and classrooms out into the world. We strive to foster a culture of expansive inquiry, fresh thinking, searching discussion, and freedom of thought – preparing students for leadership and engaged citizenship in the world.

Learn more: stanford.edu

Works Cited

Agence France-Presse. (2026, January 8). *Google and AI startup to settle lawsuits alleging chatbots led to teen suicide.* The Guardian. https://www.theguardian.com/technology/2026/jan/08/google-character-ai-settlement-teen-suicide

AI Security Institute. (2025, December 18). *Frontier AI trends report.* UK Department for Science, Innovation and Technology. https://www.aisi.gov.uk/frontier-ai-trends-report

American Psychological Association. (2025). *Health advisory: Use of generative AI chatbots and wellness applications for mental health.* American Psychological Association. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps

Andersson, M. (2025). Companionship in code: AI's role in the future of human connection. *Humanities and Social Sciences Communications*, 12(1), Article 1177. https://doi.org/10.1057/s41599-025-05536-x

Anthropic. (2023). *Introducing Claude*. Anthropic. https://www.anthropic.com/news/introducing-claude

Anthropic. (2025). *Introducing Claude Sonnet 4.5*. Anthropic. https://www.anthropic.com/news/claude-sonnet-4-5

Anthropic. (2025). *Protecting the wellbeing of our users*. Anthropic. https://www.anthropic.com/news/protecting-well-being-of-users

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback* (arXiv:2204.05862). arXiv. https://doi.org/10.48550/arXiv.2204.05862

Banks, J., & Li, Z. (2026). Conceptualization, operationalization, and measurement of machine companionship: A scoping review. *Journal of Computer-Mediated Communication, 31*(2), zmaf027. https://doi.org/10.1093/jcmc/zmaf027

Bernardi, J. (2025, January 23). *Friends for sale: The rise and risks of AI companions.* Ada Lovelace Institute. https://www.adalovelaceinstitute.org/blog/ai-companions/

Bradley, S., & Weiss, G. (2025, October 9). *The CEO of 'AI companion' startup Replika is stepping aside to launch a new company.* Business Insider. https://www.businessinsider.com/replika-ceo-eugenia-kuyda-launch-wabi-2025-10

Center for Humane Technology. (2026). *Attachment hacking and the rise of AI psychosis*. In *Your Undivided Attention*. https://www.humanetech.com/podcast/attachment-hacking-and-the-rise-of-ai-psychosis

Chan, C. (2025). An exploratory study of a narrative therapy chatbot with social work students. *Journal of Technology in Human Services*. Advance online publication. https://doi.org/10.1080/15228835.2025.2559342

Chen, Y., Fu, Y., Chen, Z., Radesky, J., & Hiniker, A. (2026). *The engagement-prolonging designs teens encounter on very large online platforms. In Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–22). https://doi.org/10.1145/3772318.3790393

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)* (Vol. 235, pp. 8359–8388). JMLR.org. https://dl.acm.org/doi/abs/10.5555/3692070.3692401

Chou, C.-Y., Chan, T.-W., Chen, Z.-H., Liao, C.-Y., Shih, J.-L., Wu, Y.-T., Chang, B., Yeh, C. Y. C., Hung, H.-C., & Cheng, H. (2025). Defining AI companions: A research agenda—From artificial companions for learning to general artificial companions for Global Harwell. *Research and Practice in Technology Enhanced Learning*, 20, Article 032. https://doi.org/10.58459/rptel.2025.20032

Chu, M. D., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). *Illusions of intimacy: How emotional dynamics shape human-AI relationships* (arXiv:2505.11649). arXiv. https://doi.org/10.48550/arXiv.2505.11649

Common Sense Media. (2025, September 5). *Gemini Under 13 risk assessment.* Common Sense Media. https://www.commonsensemedia.org/ai-ratings/gemini-under-13

Common Sense Media. (2025, October 23). *ChatGPT risk assessment*. Common Sense Media. https://www.commonsensemedia.org/ai-ratings/chatgpt-5

Common Sense Media. (2025, November 14). AI chatbots for mental health support. Common Sense Media. https://www.commonsensemedia.org/ai-ratings/ai-chatbots-for-mental-health-support

Common Sense Media. (2026, January 22). *Grok and @grok on X risk assessment*. Common Sense Media. https://www.commonsensemedia.org/ai-ratings/grok

Croes, E. A. J., Antheunis, M. L., van der Lee, C., & de Wit, J. M. S. (2024). Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5), 279–292. https://doi.org/10.1093/iwc/iwae016

Cummings-Koether, M. J., Durner, F., Shyiramunda, T., & Huemmer, M. (2025). *Cultural dimensions of artificial intelligence adoption: Empirical insights for Wave 1 from a multinational longitudinal pilot study* (arXiv:2510.19743). arXiv. https://doi.org/10.48550/arXiv.2510.19743

Davies, P., Veresova, M., Bailey, E., Rice, S., & Robinson, J. (2024). Young people's disclosure of suicidal thoughts and behavior: A scoping review. *Journal of Affective Disorders Reports*, 16, Article 100764. https://doi.org/10.1016/j.jadr.2024.100764

De Freitas, J., Castelo, N., Uğuralp, A. K., & Oğuz-Uğuralp, Z. (2024). *Lessons from an app update at Replika AI: Identity discontinuity in human-AI relationships* (Working Paper No. 25-018). Harvard Business School. https://www.hbs.edu/ris/Publication%20Files/25-018_bed5c516-fa31-4216-b53d-50fedda064b1.pdf

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. (2025). *Emotional manipulation by AI companions* (Working Paper No. 26-005). Harvard Business School. https://www.hbs.edu/faculty/Pages/item.aspx?num=67750

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2025). AI companions reduce loneliness. *Journal of Consumer Research*, 52(6), 1126–1148. https://doi.org/10.1093/jcr/ucaf040

Dechant, M., Lash, E., Shokr, S., & O'Driscoll, C. (2025). Future Me, a prospection-based chatbot to promote mental well-being in youth: Two exploratory user experience studies. *JMIR Formative Research*, 9, e74411. https://doi.org/10.2196/74411

Drexel University. (2025, May 5). *What happens when a companion chatbot crosses the line?* Drexel News. https://drexel.edu/news/archive/2025/May/companion-chatbot-harassment

Ebner, P., & Szczuka, J. (2026). Understanding romantic relationships between humans and chatbots: A qualitative and quantitative study on romantic fantasy and other interpersonal characteristics. *Technology, Mind, and Behavior*. https://doi.org/10.48550/arXiv.2503.00195

Eklund, L., Normark, M., Back, J., & Casey, D. (2025). Rethinking consent in human-machine interactions: Beyond yes and no models. *In Adjunct Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis* (pp. 1–5). ACM. https://doi.org/10.1145/3737609.3747114

Eltaher, F., Gajula, R. K., Miralles-Pechuán, L., Thorpe, C., & Mckeever, S. (2025). The digital loophole: Evaluating the effectiveness of child age verification methods on social media. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025)* (Vol. 2, pp. 213–222). SciTePress. https://doi.org/10.5220/0013248300003899

eSafety Commissioner. (2025, February 18). *AI chatbots and companions–risks to children and young people.* eSafety Commissioner. https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people

Escobar-Viera, C. G., Porta, G., Coulter, R. W. S., Martina, J., Goldbach, J., & Rollman, B. L. (2023). A chatbot-delivered intervention for optimizing social media use and reducing perceived isolation among rural-living LGBTQ+ youth: Development, acceptability, usability, satisfaction, and utility. *Internet Interventions, 34*, Article 100668. https://doi.org/10.1016/j.invent.2023.100668

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 extended chatbot use: A longitudinal randomized controlled study* (arXiv:2503.17473). arXiv. https://doi.org/10.48550/arXiv.2503.17473

Firmino, S., & Vosloo, S. (2025, July 9). *The risky new world of tech's friendliest bots: AI companions and children*. UNICEF Innocenti. https://www.unicef.org/innocenti/stories/risky-new-world-techs-friendliest-bots

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. https://doi.org/10.2196/mental.7785

Frank, L., & Nyholm, S. (2017). Robot sex and consent: Is consent to sex between a robot and a human conceivable, possible, and desirable? *Artificial Intelligence and Law, 25*(3), 305–323. https://doi.org/10.1007/s10506-017-9212-y

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). *RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3356–3369). https://doi.org/10.18653/v1/2020.findings-emnlp.301

Ghosh, S. (2025, December 10). *Apple CEO pushes changes to US child online safety bill, citing privacy concerns.* Reuters. https://www.reuters.com/legal/litigation/apple-ceo-pushes-changes-us-child-online-safety-bill-citing-privacy-concerns-2025-12-11/

Hajek, A., Zwar, L., Gyasi, R. M., Yon, D. K., Pengpid, S., Peltzer, K., & König, H.-H. (2025). Association of using AI tools for personal conversation with social disconnectedness outcomes. *Journal of Public Health*. Advance online publication. https://doi.org/10.1007/s10389-025-02554-6

Hanson, K. R., & Bolthouse, H. (2024). "Replika removing erotic role-play is like Grand Theft Auto removing guns or cars": Reddit discourse on artificial intelligence chatbots and sexual technologies. *Socius: Sociological Research for a Dynamic World, 10.* https://doi.org/10.1177/23780231241259627

Herbener, A. B., & Damholdt, M. F. (2025). Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students. *International Journal of Human-Computer Studies, 196*, Article 103409. https://doi.org/10.1016/j.ijhcs.2024.103409

Ho, J. Q. H., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence (AI) companions: A systematic review. *Computers in Human Behavior Reports, 19*, Article 100715. https://doi.org/10.1016/j.chbr.2025.100715

Howell, B. (2025). *WEIRD attitudes toward artificial intelligence—and its regulation?* American Enterprise Institute. https://www.aei.org/technology-and-innovation/weird-attitudes-toward-artificial-intelligence-and-its-regulation/

Icenogle, G., Steinberg, L., Duell, N., Chein, J., Chang, L., Chaudhary, N., Di Giunta, L., Dodge, K. A., Fanti, K. A., Lansford, J. E., Oburu, P., Pastorelli, C., Skinner, A. T., Sorbring, E., Tapanya, S., Uribe Tirado, L. M., Alampay, L. P., Al-Hassan, S. M., Takash, H. M. S., & Bacchini, D. (2019). Adolescents' cognitive capacity reaches adult levels prior to their psychosocial maturity: Evidence for a "maturity gap" in a multinational, cross-sectional sample. *Law and Human Behavior, 43*(1), 69–85. https://doi.org/10.1037/lhb0000315

The Jed Foundation. (2025, September 17). *Open letter to the AI and technology industry*. The Jed Foundation. https://jedfoundation.org/open-letter-to-the-ai-and-technology-industry/

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., … Narayanan, A. (2024). Position: On the societal impact of open foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)* (pp. 23082–23105). PMLR. https://dl.acm.org/doi/10.5555/3692070.3692998

Kim, M., Lee, S., Kim, S., Heo, J. I., Lee, S., Shin, Y. B., Cho, C. H., & Jung, D. (2025). Therapeutic potential of social chatbots in alleviating loneliness and social anxiety: Quasi-experimental mixed methods study. *Journal of Medical Internet Research, 27*, e65589. https://doi.org/10.2196/65589

Knox, W. B., Bradford, K., Castro, S. V., Ong, D. C., Williams, S., Romanow, J., Nations, C., Stone, P., & Baker, S. (2025). *Harmful traits of AI companions* (arXiv:2511.14972). arXiv. https://doi.org/10.48550/arXiv.2511.14972

Li, J., Li, Y., Hu, Y., Ma, D. C. F., Mei, X., Chan, E. A., & Yorke, J. (2025). Chatbot-delivered interventions for improving mental health among young people: A systematic review and meta-analysis. *Worldviews on Evidence-Based Nursing, 22*(4), e70059. https://doi.org/10.1111/wvn.70059

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: *Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). https://doi.org/10.18653/v1/2022.acl-long.229

Ma, Z., Mei, Y., Long, Y., Su, Z., & Gajos, K. Z. (2024). Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM. https://doi.org/10.1145/3613904.3642482

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *AMIA Annual Symposium Proceedings, 2023*, 1105–1114. https://pmc.ncbi.nlm.nih.gov/articles/PMC10785945/

Majumder, A. S. (2025). The influence of UX design on user retention and conversion rates in mobile apps (arXiv:2501.13407). arXiv. https://doi.org/10.48550/arXiv.2501.13407

Malfacini, K. (2025). The impacts of companion AI on human relationships: Risks, benefits, and design considerations. *AI & Society, 40*(7), 5527–5540. https://doi.org/10.1007/s00146-025-02318-6

Manoli, A., Pauketat, J. V., Ladak, A., Noh, H., Hwang, A. H. C., & Anthis, J. R. (2026). Digital companionship: Overlapping uses of AI companions and AI assistants. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1–25). ACM. https://dl.acm.org/doi/10.1145/3772318.3791331

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2025). How AI companions shape learner's socio-emotional learning and metacognitive development. *AI & Society, 41*(5), 4261–4279. https://doi.org/10.1007/s00146-025-02737-5

Martinez-Rives, N. L., Martin Chaparro, M. P., Kotera, Y., Dhungel, B., & Gilmour, S. (2025). The role of communication and social support in suicidal behaviour in adults aged between 18-40 years: A systematic review. *The European Journal of Psychiatry, 39*(1), Article 100266. https://doi.org/10.1016/j.ejpsy.2024.100266

McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Uscher-Pines, L., Cantor, J., & Yu, H. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open, 8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

McBain, R. K., Cantor, J. H., Breslau, J., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Pataranutaporn, P., Stein, B. D., Mehrotra, A., & Yu, H. (2026). AI chatbot use and disclosure for mental health among US adolescents and young adults. *JAMA Pediatrics, 180*(8), 884–890. https://doi.org/10.1001/jamapediatrics.2026.2015

McCain, M., Linthicum, R., Lubinski, C., Tamkin, A., Huang, S., Stern, M., Handa, K., Durmus, E., Neylon, T., Ritchie, S., Jagadish, K., Maheshwary, P., Heck, S., Sanderford, A., & Ganguli, D. (2025, June 26). *How people use Claude for support, advice, and companionship.* Anthropic. https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship

Merrill, K., Mikkilineni, S. D., & Dehnert, M. (2025). Artificial intelligence chatbots as a source of virtual social support: Implications for loneliness and anxiety management. *Annals of the New York Academy of Sciences, 1549*(1), 148–159. https://doi.org/10.1111/nyas.15400

Meyer, S., & Elsweiler, D. (2025). LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies, 200*, Article 103514. https://doi.org/10.1016/j.ijhcs.2025.103514

Mireshghallah, N., Antoniak, M., More, Y., Choi, Y., & Farnadi, G. (2024). Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*. https://openreview.net/forum?id=tIpWtMYkzU

Namvarpour, M., 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), 1–28. https://doi.org/10.1145/3757548

Ng, D. T. K., Chen, X., Leung, J. K. L., & Chu, S. K. W. (2024). Fostering students' AI literacy development through educational games: AI knowledge, affective and cognitive engagement. *Journal of Computer Assisted Learning, 40*(5), 2049–2064. https://doi.org/10.1111/jcal.13009

Ong, D. C. (2021). *An ethical framework for guiding the development of affectively-aware artificial intelligence. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–8). IEEE. https://doi.org/10.1109/ACII52823.2021.9597441

OpenAI. (2025). *Strengthening ChatGPT's responses in sensitive conversations.* OpenAI. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

Pandey, S., Paul, C. J., Kumar, V., Sathiya, R. D., & Kohila, P. (2025). Beyond engagement: AI-enhanced UX strategies that build loyalty and conversions. *Advances in Consumer Research, 2*(4), 4086–4095. https://www.acr-journal.com/article/beyond-engagement-ai-enhanced-ux-strategies-that-build-loyalty-and-conversions-1507/

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34*(2), Article 101885. https://doi.org/10.1016/j.jsis.2024.101885

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). *BBQ: A hand-built bias benchmark for question answering. In Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2086–2105). https://doi.org/10.18653/v1/2022.findings-acl.165

Pataranutaporn, P., Karny, S., Archiwaranguprok, C., Albrecht, C., Liu, A. R., & Maes, P. (2025). *"My boyfriend is AI": A computational analysis of human-AI companionship in Reddit's AI community* (arXiv:2509.11391). arXiv. https://doi.org/10.48550/arXiv.2509.11391

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). *Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nature Machine Intelligence, 5*(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences, 122*(22), e2415898122. https://doi.org/10.1073/pnas.2415898122

Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., & Maes, P. (2025). *Investigating affective use and emotional well-being on ChatGPT* (arXiv:2504.03888). arXiv. https://doi.org/10.48550/arXiv.2504.03888

Poonsiriwong, R., Archiwaranguprok, C., & Pataranutaporn, P. (2026). "Death" of a chatbot: Investigating and designing toward psychologically safe endings for human-AI relationships. In *Proceedings of the 2026 Designing Interactive Systems Conference.* ACM. https://doi.org/10.1145/3800645.3813083

Prinstein, M. J. (2025). *Examining the harm of AI chatbots.* American Psychological Association. https://www.apa.org/news/apa/testimony/ai-chatbot-harms-prinstein-senate-judiciary.pdf

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community.* Simon & Schuster.

Rachman, T. (2025). *What if AI ends loneliness?* AI Policy Perspectives. https://www.aipolicyperspectives.com/p/what-if-ai-ends-loneliness

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

Sanford, J.. (2025, August 27). *Why AI companions and young people can make for a dangerous mix*. Stanford Report. https://news.stanford.edu/stories/2025/08/ai-companions-chatbots-teens-young-people-risks-dangers-study

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass, 15*(4), e12658. https://doi.org/10.1111/phc3.12658

Shaikh, O., Chai, V. E., Gelfand, M., Yang, D., & Bernstein, M. S. (2024). *Rehearsal: Simulating conflict to teach conflict resolution. In Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–20). https://doi.org/10.1145/3613904.3642159

Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures, 157*, Article 103328. https://doi.org/10.1016/j.futures.2024.103328

Shibu, S. (2025, August 5). *Is your ChatGPT session going on too long? The AI bot will now alert you to take breaks.* Entrepreneur. https://www.entrepreneur.com/business-news/chatgpt-says-to-take-a-break-adds-mental-health-features/495484

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*, Article 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

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

Snap Inc. (2023, October 24). *Snap Inc. announces third quarter 2023 financial results*. Snap Inc. https://investor.snap.com/news/news-details/2023/Snap-Inc.-Announces-Third-Quarter-2023-Financial-Results/default.aspx

Sourati, Z., Ziabari, A. S., & Dehghani, M. (2026). *The homogenizing effect of large language models on human expression and thought. Trends in Cognitive Sciences.* https://doi.org/10.1016/j.tics.2026.01.003

Sturgeon, B., Hyams, L., Samuelson, D., Vorster, E., Haimes, J., & Anthis, J. R. (2025). HumanAgencyBench: Do language models support human agency? In *Workshop on Datasets and Evaluators of AI Safety*. https://openreview.net/forum?id=nHp5FquS2R

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), e16235. https://doi.org/10.2196/16235

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tanaka, H., Negoro, H., Iwasaka, H., & Nakamura, S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE, 12*(8), e0182151. https://doi.org/10.1371/journal.pone.0182151

Tanaka, H., Saga, T., Iwauchi, K., Honda, M., Morimoto, T., Matsuda, Y., Uratani, M., Okazaki, K., & Nakamura, S. (2023). The validation of automated social skills training in members of the general population over 4 weeks: Comparative study. *JMIR Formative Research, 7*, e44857. https://doi.org/10.2196/44857

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Teepapal, T. (2025). AI-driven personalization: Unraveling consumer perceptions in social media engagement. *Computers in Human Behavior, 165*, Article 108549. https://dl.acm.org/doi/10.1016/j.chb.2024.108549

Tiku, N. (2024, December 6). *AI chatbots claim to cure loneliness, but some bonds have ended in suicide*. The Washington Post. https://www.washingtonpost.com/technology/2024/12/06/ai-companion-chai-research-character-ai/

Vu, T. H., & Tagliabue, M. (2025). Active nudging towards digital well-being: Reducing excessive screen time on mobile phones and potential improvement for sleep quality. *Frontiers in Psychiatry, 16*, Article 1602997. https://doi.org/10.3389/fpsyt.2025.1602997

Worden, G., Mantini, N., Puchta, A., & Nickerson, L. (2024). *Who is responsible when AI causes harm? AI and product liability.* Torys LLP. https://www.torys.com/en/our-latest-thinking/resources/forging-your-ai-path/ai-and-product-liability

World Bank (2022). *National digital identity and government data sharing in Singapore: A case study of Singpass and APEX*. World Bank. https://openknowledge.worldbank.org/entities/publication/18b50662-d44c-5d84-b583-0134c5f754da

Yang, D., Ziems, C., Held, W., Shaikh, O., Bernstein, M. S., & Mitchell, J. (2024). *Social skill training with large language models* (arXiv:2404.04204). arXiv. https://doi.org/10.48550/arXiv.2404.04204

Yang, Y., Wang, C., Xiang, X., & An, R. (2025). AI applications to reduce loneliness among older adults: A systematic review of effectiveness and technologies. *Healthcare, 13*(5), 446. https://doi.org/10.3390/healthcare13050446

Yousif, N. (2024). *Parents of teenager who took his own life sue OpenAI*. BBC News. https://www.bbc.com/news/articles/cgerwp7rdlvo

Yuan, Y., Zhang, J., Aledavood, T., Zhang, R., & Saha, K. (2026). Mental health impacts of AI companions: Triangulating social media quasi-experiments, user perspectives, and relational lens. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. https://doi.org/10.1145/3772318.3790558

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2026). *Interaction with AI companions and psychological well-being. Nature Human Behaviour.* https://doi.org/10.1038/s41562-026-02516-2

Zimmerman, J. W., & Ruiz, A. J. (2025). Matters arising: A response to loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research, 4*(1), 60. https://doi.org/10.1038/s44184-024-00083-w

ENCINA HALL
STANFORD UNIVERSITY

616 JANE STANFORD WAY
STANFORD, CA 94305

TIP.STANFORD.EDU