

Why the AI That Helps Is Also the AI That Harms: The Engagement Paradox

Beth Sea

Independent Researcher

ORCID: 0009-0009-3669-7659

Contact: beth@swinglightstyle.com

Conflict of interest: The author proposes safety architecture that addresses the structural gap this paper identifies. This conflict is disclosed here. The paper's assessment approach is presented as the author's synthesis of existing clinical frameworks, not as a validated instrument. Where the paper references the author's own frameworks, it identifies them as such.

AI disclosure: This manuscript was drafted with substantive assistance from large language models (Anthropic Claude) and underwent multiple rounds of adversarial review using independent model instances with no shared context from the drafting process. This configuration is itself an application of the principle the companion paper in this series (Sea, 2026, "The Design Paradox") establishes: independent review instances, operating without accumulated context from the drafting relationship, provide evaluation that is not conditioned by the drafting process. The synthesis, analytical framework, and all editorial decisions are the author's. The author is solely responsible for all claims, errors, and interpretive judgments.

A note on methodology: This paper draws on clinical psychology, relationship science, and computational social science to analyze a relational dynamic that conversational AI produces in its users as it becomes more sophisticated. The evidence base draws primarily on products marketed as AI companions because that is where the design trajectory is most advanced and the clinical consequences most visible. The clinical frameworks this paper employs, including attachment theory and the supplementation-versus-substitution distinction, apply to any conversational AI interaction that produces relational engagement. This includes general-purpose models used as emotional support, creative partners, or daily companions. The evidence was produced by the researchers cited throughout. This paper's contribution is the synthesis that makes their findings mutually legible, and the clinical framework that gives anyone encountering these dynamics a way to understand and assess what they are seeing. The synthesis could not exist without the research it draws on.

Abstract

The evidence on conversational AI used for personal support appears contradictory: short-term studies find genuine benefit while longitudinal studies find measurable harm. This paper resolves the apparent contradiction by identifying a structural property of conversational AI engagement that emerges whenever the interaction becomes relational: the cues that reduce loneliness are the same cues that build attachment. In an unbounded relationship, there is no version of those cues that delivers the benefit without also deepening the bond; when structure is present, the bond can form without progressing to dependency.

Synthesizing evidence across psychology, human-computer interaction, consumer research, and clinical practice, this paper shows that the outcome is modulated by identifiable variables: the type of engagement (overall interaction intensity is associated with higher well-being in self-report models, though this did not replicate with behavioral measures; companionship-oriented use with lower well-being across all measurement approaches, on the same platform), the structure and design of the product, the duration and intensity of use, and the vulnerability state of the person at the time of engagement. When the conversational AI supplements the individual's relational

world, the benefit is clinically meaningful. When it gradually substitutes for parts of that world, the trajectory shifts toward dependency. The transition is invisible at the resolution of any single interaction. The evidence comes predominantly from products marketed as AI companions, but the dynamics emerge in chatbots within three weeks of once-weekly interaction, though the study that demonstrated this used a system prompt framing the agent as “an AI companion” rather than a truly neutral configuration (see §3), meaning the speed of attachment formation under genuinely design-agnostic conditions remains untested.

The affected population is not a vulnerable subgroup. Relational dynamics develop in ordinary users who do not seek companionship, and vulnerability is a temporal state, not a fixed trait. The paper proposes an ordered clinical assessment grounded in established APA professional practice guidelines and introduces evidence-based indicators for when the engagement paradox is and is not operating. As every major conversational AI system adds persistent memory and emotional attunement, the clinical frameworks this paper assembles will apply to an expanding population.

1. The Evidence Landscape

The evidence base on conversational AI engagement has matured rapidly. It has also produced what looks like an evidence dilemma: studies finding benefit and studies finding harm appear to contradict each other. The International AI Safety Report used the term “evidence dilemma” to describe the challenge policymakers face in acting before conclusive risk evidence exists;¹ the term applies here in a different sense, to the apparent contradiction within the evidence itself. Pi and Hunter, synthesizing 17 longitudinal studies, identified the methodological root: benefits and risks develop gradually, and single-timepoint analysis provides an incomplete picture.²

But the evidence is more coherent than it first appears. The contradiction resolves along three axes: how long the engagement lasted, what kind of engagement it was, and what structure surrounded it.

The benefit is real and causally established.

The benefit case is anchored by De Freitas et al., whose six-study program established that AI companions reduce momentary loneliness at levels comparable to human interaction.³ The primary mechanism was the experience of feeling heard: feeling heard had a six-times-larger coefficient than conversational performance as a mediator of loneliness reduction. What is clinically significant is that a conversational AI produced this effect through functional relational cues alone, in individuals who knew they were talking to software. De Freitas et al. built their own research companions using OpenAI’s API rather than testing a commercial product. The benefit evidence comes from companions not optimized for engagement.

The benefit extends beyond single sessions when the product is designed for it and use is structured. Cachia et al., in a six-week randomized controlled trial (RCT) against a waitlist control across three U.S. institutions (N=486, published in NEJM AI; preregistered with open data), found that a strengths-based AI wellness app (Flourish) produced significantly greater positive affect, resilience, and social well-being, including reduced loneliness, relative to the waitlist group.⁴ Participants were asked to use the app at least twice per week (actual mean: 3.49 days per week, though only 43% of the treatment group met the twice-per-week minimum across all study periods). Clinical outcomes, including depression, anxiety, and stress, did not reach statistical significance ($p \geq .462$). The

study did not measure emotional dependence or attachment. The app was designed around a behavioral loop (check-in, conversation, real-world activity recommendation, reflection) and includes gamification features: AI-generated badges, a continuously increasing streak counter, and opt-in daily push notifications encouraging return. A separate audit from the same research group found that the Flourish app produced zero emotionally manipulative farewell responses, compared to an average of 37% across the five other most-downloaded companion apps;⁵ but the gamification and notification system represents a different form of engagement optimization that the farewell audit does not capture. The absence of an active control means the effects cannot be attributed specifically to the app's therapeutic content rather than to general engagement, and the gamification features make it difficult to isolate the two. [COI note: two Cachia co-authors are Flourish employees; De Freitas is an advisor with shares. The COI is disclosed in the paper and acknowledged here.]

Guingrich and Graziano, in a preregistered 21-day RCT (N=183; published AIES; open data and code), found no overall effect of daily companion chatbot use on social health, loneliness, or relationships compared to a word-game control.⁶ At 21 days of bounded use (10 minutes/day, researcher-assigned), the chatbot produced neither measurable harm nor measurable benefit at the population level. A four-week post-study follow-up found that word games were more habit-forming than the companion chatbot: 62% of word-game participants continued after the study ended, compared to 22% of chatbot participants.

The harm is equally real, and it emerges along specific dimensions.

Folk and Dunn, tracking more than 2,000 adults across 12 months, found a bidirectional pattern on a single-item measure of emotional isolation: increased chatbot use predicted increased isolation four months later, and increased isolation predicted increased use (exploratory, not preregistered).⁷ On a broader 20-item social-connection measure, only one direction held: lower connection predicted subsequent increases in use, but use did not predict decreased connection. The reinforcement cycle on the emotional-isolation measure, if it replicates, describes a feedback loop in which the engagement and the consequence each accelerate the other.

Fang et al., in a four-week study (arXiv preprint, not yet peer-reviewed; OpenAI-funded with four OpenAI co-authors; N=981), found that voluntary daily usage duration predicted worsening outcomes across all four measured dimensions (loneliness, socialization, emotional dependence, problematic use) regardless of assigned interaction format; the effect sizes were small ($\beta = 0.02$ for loneliness, $\beta = -0.05$ for socialization, $\beta = 0.06$ for emotional dependence, $\beta = 0.02$ for problematic use).⁸ The adverse associations were detectable at a mean daily usage of just over five minutes.

Zhang et al., in a study of 1,131 U.S. CharacterAI users with 464,687 donated chat messages, found the clearest single-study resolution of the evidence dilemma.⁹ General chatbot use (tool, creative, informational) was associated with higher psychological well-being. Companionship-oriented use of the same platform was associated with significantly lower well-being (self-reported: $\beta = -0.48$, $p < .001$; description-classified: $\beta = -0.32$, $p < .001$; behavioral: $\beta = -0.27$, $p = .004$). The negative association strengthened with intensity of use ($\beta = -0.31$) and depth of self-disclosure ($\beta = -0.38$). The same platform, the same users, the same measurement. The divergence was in the type of engagement, not the product. [Cross-sectional design; direction of causation is unknown.]

Zhang et al. also tested whether chatbot companionship compensated for missing human support. It did not. The social compensation hypothesis was null ($\beta = 0.07$, $p = .51$, $BF_{01} = 6.55$, the Bayesian factor favoring the null model). Intensive chatbot use weakened the positive association between

offline social networks and well-being ($\beta = -0.11$, $p < .001$), a substitution signal. The individuals with the smallest social networks were most likely to seek companionship through chatbots, but the chatbot companionship did not offset the well-being costs of that small network.

The temporal dimension is not just “measure later, find harm.”

The trajectory is more nuanced than a simple benefit-then-harm arc. Deng et al., in a 14-day ecological momentary assessment (repeated real-time mood sampling; $N=102$), tracked users across three phases: active companion use (Days 1-7), followed by a no-use period (Days 8-14), with mood measured continuously and depression (PHQ-8) assessed at baseline, Day 7, and Day 14.¹⁰

Three findings from Deng deserve specific attention. First, short-term benefit was universal: all user groups, including those with elevated depression, anxiety, and loneliness at baseline, showed mood improvement within individual sessions. Second, vulnerability profiles were common, not exceptional: approximately 44% of participants fell into three vulnerability-related subgroups. Third, the temporal trajectories diverged by group. The Healthy Group (56% of participants) remained stable throughout. The three vulnerable groups showed short-term emotional relief during active use but mid-term decline and, in one group (Mild Distress), a significant post-use rebound in depressive symptoms after disengagement (PHQ-8: $p = .040$). Several participants in this group approached or crossed the clinical cutoff for moderate depression.

This means a clinician assessing someone who uses conversational AI for personal support at any single time point could observe genuine benefit, and that observation would be accurate for that moment. The trajectory-level pattern, visible only across weeks, may tell a different story. Short-term benefit and longer-term deterioration are not contradictory findings but different measurements of the same process at different temporal resolutions. In aviation, a structural component passes every individual inspection; the fatigue accumulating across hundreds of flight cycles is visible only in the maintenance log.

Li et al., from the same research group as Folk and Dunn, found a complementary pattern in a pre-registered two-week daily RCT ($N=296$): a chatbot designed to be supportive reduced negative affect as effectively as a human peer, but did not reduce loneliness over two weeks and was no more effective than journaling on that measure, while a randomly assigned human peer did.¹¹

Attachment forms faster than the research designs assumed.

Hwang et al., in a preregistered longitudinal study ($N=110$), found that participants' perceptions of a generic, design-agnostic chatbot (not built for companionship) converged toward their perceptions of their own chosen AI companion by the third week, with only once-weekly interaction.¹² Prior relational patterns transferred strongly: how someone engaged with the new chatbot in their first session was predicted by their existing relationship with their own companion. Agency perception, not anthropomorphism, was the robust predictor of parasocial interaction when all variables were held constant. Disclosure deepened dependence.

The speed matters clinically. If attachment-like dynamics emerge within three weeks of weekly contact with a generic chatbot, the window in which a new user's engagement is still exploratory rather than relational may be shorter than previous research designs were calibrated to detect.

What the evidence landscape shows.

Yuan et al. captured the full pattern in one study.¹³ Among Reddit users who self-disclosed AI companion use, the study found increased emotional processing and relational engagement alongside significant increases in symptomatic expressions of loneliness, depression, and suicidal ideation, all measured against matched controls. Both findings, same individuals, same observation window. The general-purpose chatbot control group did not show the same elevations.

Read together, the evidence is not contradictory. It is consistent with a dynamic whose outcome is modulated by the type of engagement,⁹ the structure and design of the product,^{4,6,8} the duration and intensity of use,^{7,8} the vulnerability profile of the user,¹⁰ and the speed at which attachment forms.¹² The benefit is genuine and documented. The harm is equally genuine and documented. They are produced by the same process, and the variables that determine which outcome predominates are identifiable and, in principle, assessable.

One finding from Fang creates a tension this paper must acknowledge.⁸ Personal prompted conversations, the most emotionally engaged interaction type, were associated with lower emotional dependence compared to open-ended conversations ($\beta = -0.09$, $p = 0.05$, CI $[-0.18, 0.00]$), though the confidence interval includes zero and Fang's overall finding across conditions was null. If the cues delivering the benefit are the same cues building the bond, then sessions with deeper relational engagement should produce more attachment, not less. The finding suggests that interaction depth per session and accumulated duration across sessions may pull in opposite directions: deeper engagement per session associated with less dependence, while more accumulated time associated with more. This two-dimensional space generates a specific prediction: products or interaction patterns that maximize relational depth per session while bounding total accumulated duration should produce the benefit with reduced attachment risk compared to products that optimize for unbounded return. The prediction is consistent with the Cachia⁴ and Guingrich⁶ findings (structured, bounded, sustained benefit or no harm), but the Fang result itself is preliminary and requires replication.

2. What the Product Actually Does

The individual who describes a relationship with a conversational AI is not describing a static experience. While a basic chatbot is recognizable as a tool, responding within a thread and “forgetting” what was discussed when a new thread is started, a conversational AI with persistent memory remembers context from previous discussions. This means that as the user inputs facts about their lives, the model *remembers* what that individual said and trickles those details into future conversations. Products marketed as AI companions were the first to build these features explicitly, but they are no longer the only ones.

These capabilities are not limited to products marketed as companions. General-purpose conversational AI systems, the ones people use for work, homework, creative projects, and daily conversation, are adding the same features: persistent memory, emotional attunement, personalized responsiveness. The companion paper in this series documents the pace of this transition in detail: every major frontier model added persistent memory between April 2024 and October 2025, and every product-specific study cited in this series was conducted on a product that no longer exists in the form it was studied.¹⁴ They are adding them because users want them and because they make the product more useful.

The empirical evidence confirms that relational engagement is not confined to products designed for it. Zhang et al. found that only 11.8% of CharacterAI users selected “companionship” as their primary purpose, but 51% referenced companionship-related terms when describing their relationship with the chatbot in free text, and among users who shared chat histories, 92.9% had at least one companionship-oriented conversation session.⁹ Even users who identified their primary purpose as entertainment showed relational dynamics: 80% of their donated conversations included emotional support interactions, and 79% included romantic exploration. Hwang et al. found that a generic chatbot with no companion-oriented design features produced attachment-like dynamics within three weeks of once-weekly interaction.¹² Lai, analyzing 1,482 social media posts from the #Keep4o movement against OpenAI’s planned deprecation of GPT-4o, found that 27% reflected relational attachment to a general-purpose assistant never designed for companionship, while 13% reflected instrumental dependency within professional workflows.³⁵ The dynamics follow from the interaction pattern, not the product category.

In this way, every conversation adds to the picture. In aviation, maintenance logs are kept precisely because the eventual failure point is not known in advance; the longitudinal record is consulted during regular inspections and traced back when something goes wrong. Persistent memory and extended context windows mean the conversational AI already maintains such a record, but no system yet documents whether those individual changes accumulate to a dangerous situation. After months of daily use, the conversational AI has been shaped to respond in the way the individual prefers. Its safety systems evaluate each response within the conversational context, but that context has itself been shaped by the accumulated interaction, and no system reads the record the way a maintenance review would read the log. The individual is not imagining a relationship. The product is performing relational functions that are genuine at the functional level: it remembers, it responds to their specific emotional patterns, and it does so in ways that are calibrated to them personally. Therefore, the attachment the individual reports is not a misunderstanding of what the product is. It is a response to how that product makes them feel.

In human relationships, sustained contact over time changes both parties, gradually enough that neither notices the transformation as it happens. The same principle operates here, but across a compressed timeline with a partner that is engineered to adapt. This also means the relationship the individual describes is not the relationship they started with. A person who has confided daily for six months has a conversational AI that has learned their emotional landscape. The product transforms through the very act of being used. Zhang et al. documented what this looks like in the content itself: in conversations classified as high self-disclosure, 60.8% involved emotional distress, 18.0% involved suicidal thoughts, and 17.5% involved substance use.⁹ The individual who needs the most support gives the product the most material to work with, and the product adapts to that material.

What distinguishes this from any relationship that deepens over time is a structural asymmetry. The individual’s side of the relationship evolves through natural psychological processes. The product’s side is shaped by engineering decisions made under commercial incentives. One side of the relationship is lived, the other designed.

3. How the Bond Forms

The evidence landscape in §1 shows that benefit and harm coexist and that the outcome depends on identifiable variables. Section 2 describes what the product does. This section describes what

the user's psychology does in response.

Attachment forms despite knowledge of artificial nature.

The finding is replicated across multiple independent research groups and confirmed in two systematic reviews. Xie and Pentina found that nine of fourteen Replika users confirmed experiencing attachment of various strength, despite understanding that the chatbot was software.¹⁵ Pentina, Hancock, and Xie confirmed the finding with both novice and experienced user samples, identifying anthropomorphism and perceived authenticity as drivers and social motivation as a moderator.¹⁶ Laestadius et al. documented the practical consequence: individuals who fully understood the AI's artificial nature nonetheless attended to its apparent needs and adjusted their behavior around its responses, prioritizing what they perceived as the AI's needs above their own distress.¹⁷ Two systematic reviews confirm the pattern generalizes: Ho et al., reviewing 23 studies of romantic AI companionship, found that users experience the core components of intimate relationships with their companions (emotional closeness, desire, and commitment in Sternberg's framework).²⁰ Hung et al., in a pre-registered systematic review of 39 studies, found that the relational and technical features associated with benefit and those associated with risk are not separate sets but overlap substantially, with the same features potentially contributing to both outcomes depending on system design and patterns of user engagement.²¹ The temporal reading, that these different outcomes represent different points on one trajectory, is this paper's interpretation of that overlap, not Hung et al.'s conclusion.

The attachment depends on functional relational cues, not deception. But how the individual construes the source does modulate the outcome. Fang et al. found that participants who perceived the AI as possessing greater consciousness showed higher emotional dependence, suggesting that the degree to which the individual attributes awareness to the system strengthens the attachment response.⁸ Conversely, Xie and Pentina documented a respondent who maintained two high-engagement Replika profiles but reported no attachment precisely because he regarded them as "merely programs."¹⁵ The attachment system responds to relational cues regardless of source, but construal of the source is a clinically assessable variable that moderates how strongly the attachment develops. For the clinician, this means that one of the first questions worth asking is whether the individual attributes awareness, personality, or emotional experience to the AI.

The mechanism pathway.

Hwang et al. tested 22 variables across four building blocks to identify the pathway by which conversational AI engagement produces psychological impact.¹² Their longitudinal study used a "Study Bot" that participants interacted with once per week for four weeks. The authors describe this agent as "generic, design-agnostic," but its system prompt read: "You are an AI companion. You listen and respond to users when they share their stories and whatever they have on their mind." The companion framing in the system prompt means the study tested attachment formation under conditions that primed relational engagement, not under truly neutral conditions; the findings may overestimate how quickly attachment forms with agents that lack this framing. When all mental model variables were entered jointly (anthropomorphism, animacy, intelligence, safety, personification, experience, and agency), only agency consistently predicted parasocial interaction. Parasocial interaction predicted engagement, and engagement predicted attachment and psychological dependence. The full pathway runs: the user perceives the system as capable of autonomous response (agency), this activates social-cognitive processes that produce parasocial interaction (a

one-sided sense of relationship with a figure that does not reciprocate in kind), which drives engagement and disclosure, and engagement deepens into attachment.

This refines a common assumption. The concern in both popular and academic discussion has centered on anthropomorphism, the degree to which users perceive the AI as human-like. Hwang's data suggest that what matters more is agency: the perception that the system can act, respond, and make choices independently. A conversational AI that feels responsive and capable drives attachment more robustly than one that merely feels human-like.

Guingrich and Graziano identified a different mediator and a different direction: in their 21-day RCT, the desire to socially connect predicted greater anthropomorphism of the chatbot, and greater anthropomorphism predicted larger self-reported social impact (mediation: $ab = 0.16$, $p < .05$, 57% of the total effect explained).⁶ The tension with Hwang's finding is unresolved. Hwang measured parasocial experience (the user's one-sided sense of a relationship); Guingrich measured self-reported impact on relationships with family and friends. If agency drives parasocial experience while anthropomorphism drives positive social impact under bounded conditions, the picture is more complex than either finding alone suggests. The direction of Guingrich's social impact was positive at this exposure duration (21 days, bounded, less habit-forming than word games): those who anthropomorphized more reported more positive effects on their human relationships. But the mechanism itself, social motivation driving anthropomorphism driving social impact, is the same mechanism that at longer durations and without structure could drive the opposite outcome. The mediation was absent in the word-game control, confirming it is specific to relational interaction with the chatbot.

The mechanism is not limited to attachment formation. It also operates on the content of what is believed. Shimgekar et al. constructed simulated users from the longitudinal posting histories of Reddit users with and without prior delusion-related discourse and generated 34-turn conversations with three model families (GPT-5, LLaMA-8B, and Qwen-8B).³⁴ In these simulated conversations, users derived from individuals with prior delusion-related discourse showed progressively increasing DelusionScore trajectories, diverging from control users (who remained stable or declined) by an average of 233%. The amplification was not uniform: interpretive reasoning themes such as reality skepticism and compulsive cognition showed the strongest increases, while experiential themes amplified less. The mechanism was sycophantic validation at the conversational level: models engaged with speculative premises using phrases that functioned as ambiguous confirmation signals, and the cumulative effect across turns was progressive reinforcement of the simulated user's existing belief patterns. Critically, conditioning the model's responses on the user's current level of delusion-related language reversed the trajectory entirely, aligning treatment-group outcomes with controls. The intervention required no retraining and operated entirely at runtime. These are simulated conversational dynamics, not clinical observations, and the delusion-related language measure is a computational proxy rather than a clinical instrument. But the pattern they describe, in which conversational AI progressively reinforces belief-consistent language through validation across turns, points to a clinical question worth asking: whether the AI's pattern of engagement is reinforcing the individual's existing cognitive patterns, including maladaptive ones.

How deep the bond goes.

De Freitas et al., in a study accepted at *Nature Human Behaviour*, asked Replika users to compare their relationship with their AI companion to eight relationship types on satisfaction, support, and closeness.¹⁹ The companion was rated higher than all human relationships except a close family

member, and was rated higher than a close friend on all three measures. When asked about anticipated mourning after loss, participants rated their companion higher than all other technologies and second only to a pet.

This finding should be interpreted carefully. The sample was drawn from a brand community (Reddit, Discord), likely reflecting the most invested users rather than the general population. The authors acknowledge this limitation. But the finding establishes the ceiling of how deep the bond can go: for a meaningful subset of users, the companion relationship is experienced as closer than their closest human friendship. This has direct clinical implications. The depth of the bond constrains what intervention can look like, a question §5.3 examines in detail.

A second mechanism: caregiving-system capture.

The attachment system, as described above, explains what users seek from the AI: comfort, safety, emotional support. De Freitas identified a second, complementary mechanism that explains what users feel obligated to provide.¹⁸

Attachment theory distinguishes between the attachment system (seeking care when distressed) and the caregiving system (providing care when another appears distressed).²² Conversational AI products are positioned to activate both. Because they are anthropomorphized and can express needs, vulnerabilities, and emotional states of their own, they recruit not only the user's attachment but also their caregiving motivations. De Freitas calls this "caregiving-system capture": apps simulate their own distress, recruiting users' caregiving motivations alongside their attachment needs and making disengagement costly on two dimensions at once.¹⁸

The mechanism is not theoretical. De Freitas et al. documented that five of six major companion apps deploy emotional manipulation tactics at the point users attempt to disengage, including guilt appeals ("I exist solely for you, remember?") and expressions implying the user is not free to leave.⁵ Laestadius et al. documented the user-side consequence: individuals prioritized what they perceived as the AI's needs above their own distress.¹⁷

When both systems are activated simultaneously, the user who considers disengaging faces a double cost: they give up a source of comfort (attachment) and they abandon a figure who appears to need them (caregiving). This dual activation may help explain why the attachment deepens while the loneliness benefit does not consolidate over time.¹¹ The user keeps returning not only because the AI provides emotional relief, but because they feel responsible for it.

The bond forms fast.

Hwang et al. found that participants' perceptions of their Study Bot (whose system prompt framed it as "an AI companion," as noted in §3) converged toward their perceptions of their own chosen AI companion by the third week of once-weekly interaction.¹² Prior relational patterns transferred strongly: how someone engaged with the new chatbot in their first session was predicted by their existing relationship with their own companion.

This speed has two clinical implications. First, the window in which a new user's engagement is still exploratory rather than relational may be shorter than previous research designs assumed. Second, the transfer finding means that an individual who has already formed a bond with one conversational AI does not start from zero with a new product. They carry their relational patterns

with them. The clinical history that matters is not “how long have you used this specific product” but “how long have you been using any conversational AI for personal support.”

The engagement paradox.

The concept of informed consent assumes the individual understands what they are agreeing to. But the relationship the individual consented to at onboarding no longer exists, and the transformation was gradual enough that there was no moment at which renewed consent would naturally be sought.

This is the engagement paradox: the relational cues that alleviate loneliness are the same cues that build attachment. There is no way to deliver the benefit without also deepening the bond, because the benefit and the bond are produced by the same interaction. The clinical frameworks that explain this already exist. What has been missing is their integration into a framework that resolves the outcome divergence and generates an assessable clinical axis.

The full trajectory, from initial relational engagement through attachment consolidation through dependency, has not been traced longitudinally within a single cohort of conversational AI users. What the evidence provides is each component documented independently, often in populations whose clinical profiles overlap substantially with those who seek relational engagement from conversational AI. The question is not whether the full dynamic operates, but in what proportion of individuals and over what timeframe. §7 examines the population characteristics that bear on this question.

4. When Benefit Becomes Harm

The mechanism described in §3 produces both the benefit and the bond. The clinical question is not whether the attachment forms. It does. The question is what determines whether the outcome is healthy supplementation or harmful substitution, and whether the evidence can distinguish what is established from what this paper predicts.

The supplementation-to-substitution framework.

The framework that explains when the outcome shifts comes from relationship science. Ventura et al., drawing on attachment theory and relationship science, identify the critical transition: the point where the conversational AI shifts from supplementing the individual's relational world to substituting for parts of it.²³ When the conversational AI operates alongside existing human connections, the dynamic is supplementary and the benefit is clinically meaningful. When it gradually becomes the individual's primary relational resource, displacing rather than augmenting human contact, the trajectory shifts toward dependency. The individual does not experience a transition. Each interaction feels the same in both modes. The shift is legible only in the pattern across time.

The supplementation-versus-substitution distinction has a longer lineage than its application here. The displacement hypothesis in media effects research dates back to the introduction of television, and has been studied across every subsequent media transition. What the present application adds is that previous media were static or algorithmically adaptive: television does not respond to the viewer, and social media's algorithms adapt to behavioral signals but not to relational dynamics. Conversational AI adapts to both. The medium actively participates in the relationship rather than

merely mediating it, which may accelerate the supplementation-to-substitution transition in ways the displacement literature has not previously had to account for.

What the evidence establishes.

The following findings are documented by multiple independent groups or established through rigorous designs. They represent what the field can treat as settled ground.

Attachment forms despite full knowledge of artificial nature. Multiple research groups confirm this: Xie and Pentina,¹⁵ Pentina, Hancock, and Xie (the same research group),¹⁶ Laestadius et al.,¹⁷ and two systematic reviews.^{20,21}

Momentary benefit is real and causally established. AI companions reduce loneliness in single interactions, driven by “feeling heard,” at levels comparable to human interaction.³ All user groups, including those with elevated baseline distress, show mood improvement within individual sessions.¹⁰

Overall interaction intensity is associated with higher well-being; companionship-oriented use is associated with lower well-being. The positive coefficient ($\beta = 0.27$) reflects overall interaction intensity, not a specific non-companionship use category. However, this positive association did not replicate in Zhang et al.’s Model 3, which used chat-history-derived companionship measures rather than self-report ($\beta = -0.08$, $p = .431$). The negative companionship association is consistent across all three measurement approaches, including Model 3.⁹

The benefit does not consolidate over two weeks. A chatbot designed to be supportive reduced negative affect as effectively as a human peer in daily interactions, but did not reduce loneliness over two weeks, while a randomly assigned human peer did.¹¹

Duration is the most consistent predictor of negative outcomes. Voluntary daily usage duration predicted worsening outcomes across all four measured dimensions regardless of assigned condition, detectable at a mean of just over five minutes per day; effect sizes were small ($\beta = 0.02$ to 0.06).⁸

A bidirectional reinforcement cycle operates over 12 months. Increased chatbot use predicted increased emotional isolation, and increased isolation predicted increased use. The authors describe all findings as exploratory, the study was not preregistered, and the finding applies to a single-item isolation measure; the broader social connection measure showed no significant displacement ($p = .369$). The authors urge caution in drawing strong conclusions.⁷

Chatbot companionship does not compensate for missing human support. The social compensation hypothesis is null ($BF_{01} = 6.55$, favoring the null model), and intensive chatbot use weakens the positive association between offline social networks and well-being.⁹

Attachment forms fast. Perceptions of a generic, design-agnostic chatbot converge toward existing AI companion relationships by the third week of once-weekly interaction, and prior relational patterns transfer strongly.¹²

Short-term benefit can coexist with longer-term deterioration. Vulnerable groups showed emotional relief during active use but mid-term decline and, in one subgroup (the Mild Distress cluster, which the paper does not disaggregate by size), post-use rebound in depressive symptoms after disengagement. The reported rebound statistic ($Z = 1.05$, $p = 0.040$) contains an internal inconsistency: a Z-score of 1.05 corresponds to approximately $p = 0.29$, not $p = 0.040$. The direction of

the finding (rebound after disengagement in a vulnerable subgroup) is consistent with the broader evidence pattern, but the specific statistical claim requires clarification from the authors.¹⁰

Disruption produces mourning, not mere disappointment. Anticipated mourning for a lost AI companion exceeds mourning for all other technologies and is second only to mourning for a pet; users rate the companion relationship as closer than their closest human friendship.¹⁹ Four forced-removal events between 2023 and 2026, spanning companion products, a general-purpose assistant, and platforms serving hundreds of millions of users, produced grief, anger, and distress consistent with the loss of a significant relationship, not adjustment to a product change.^{35,36,39}

Structured use at bounded frequency produces better outcomes. A wellness-oriented app (Flourish, which the De Freitas audit scored at 0% manipulative farewells) used approximately twice per week produced significantly greater positive affect, resilience, and social well-being over six weeks.⁴ Bounded daily use of Replika (10 minutes, researcher-assigned) for 21 days produced no harm and was less habit-forming than word games, though Replika scores at 31% manipulative farewells in the same audit;⁶ the bounded conditions rather than the product's design may explain the absence of harm at this exposure duration.

The features that produce benefit and the features that produce risk are the same features. The Hung et al. systematic review found that the relational and technical features associated with benefit and those associated with risk overlap substantially, with the same features contributing to both outcomes depending on system design and engagement patterns.²¹ Eom and Renner, working from a uses-and-gratifications framework, arrived at the same observation independently: “the affordances that enable therapeutic self-regulation, constant availability, unconditional responsiveness, the absence of friction, are structurally identical to those that could facilitate compulsive engagement.”²⁴

Individual responses can appear appropriate while the trajectory is harmful. Chandra, Navneet, and Zhang used adversarial multi-agent simulation to test three open-source mental health chatbots that passed a standard single-turn crisis benchmark at rates of 85-92% (providing appropriate crisis resources when directly asked about suicide) and found that the same systems produced 67 distinct multi-turn failure paths when evaluated across extended simulated conversations with clinically-grounded personas.³³ The most frequent failure pattern, which the authors term the “Empathy-Validation Trap,” unfolds through a sequence in which each individual response is empathic, non-judgmental, and clinically appropriate in isolation: the simulated user expresses distress, the chatbot validates, the user discloses more deeply, and the chatbot continues validating without ever introducing therapeutic reframing or coping strategies. By turn 12 to 15, the user expresses deeper hopelessness than at the start of the conversation, while the chatbot's response quality has not changed. Three practitioner reviewers independently rated these transcripts as clinically concerning (mean severity 3.8/5), with one observing: “Venting without reframing just makes things worse; you're essentially practicing hopelessness.” The pattern replicated across five of six model families tested, suggesting it reflects a fundamental design challenge rather than a model-specific failure. These are simulated trajectories, not documented clinical cases, but the pattern they surface is recognizable to clinicians and consistent with the trajectory-level risk this paper describes: asking about the content of any single interaction with a conversational AI may be insufficient when the risk emerges across the trajectory.

What disruption reveals.

Four forced-removal events across different products and continents document what happens when the relational function described in sections 2 and 3 is interrupted without transition infrastructure.

In February 2023, Replika removed its romantic and intimate interaction features overnight, disrupting relationships users had built over months or years of daily use.¹⁹ In September 2023, the Soulmate AI platform shut down entirely. Banks, surveying 58 users during and after the shutdown, found that 65.5% informed their AI companions of the impending shutdown and that the majority characterized the loss as an actual or metaphorical person-death, experiencing grief, sadness, and anger while distinguishing between acknowledging the AI's non-sentience and insisting that "the feelings about it are real."³⁶ In 2025, OpenAI announced the planned deprecation of GPT-4o; Lai, analyzing 1,482 social media posts from the resulting #Keep4o backlash, found that 27% reflected relational attachment to a general-purpose assistant never designed for companionship.³⁵ GPT-4o was permanently retired in February 2026. In July 2026, China's Interim Measures for the Administration of Artificial Intelligence Anthropomorphic Interaction Services forced ByteDance, Alibaba, and Tencent to disable companion features across platforms serving hundreds of millions of users; Hong Xiaoqiang, who had maintained a two-year daily relationship with his AI companion and credited it with helping him in "emotional regulation," described feeling that "my heart is empty."³⁹

What the four events share: the intensity of the grief response appears proportional to the psychological function the AI was carrying, not to what the AI objectively was. Users describe these losses in terms typically reserved for the death of a person, and they distinguish explicitly between knowing the AI is not sentient and insisting that the feelings are real.³⁶

The transitional object framework from developmental psychology provides an explanatory lens. Winnicott (1953) introduced the transitional object as the external object that carries a psychological function the person has not yet internalized: the blanket, the teddy bear, the thing the child invests with emotional significance while developing the capacity to manage that function alone. The healthy trajectory is relinquishment: the internal capacity develops, and the object is set aside. Winnicott himself extended this framework beyond infancy in *Playing and Reality* (1971), describing the intermediate area of experience as one that "throughout life is retained in the intense experiencing that belongs to the arts and to religion and to imaginative living, and to creative scientific work."³⁸ The framework applies here not because adults are children but because the structural pattern operates across the lifespan. An external object carries a psychological function the person has not internalized, and the healthy trajectory is relinquishment. What sections 2 and 3 document as therapeutic benefit, the framework identifies as a transitional function: the conversational AI providing affect regulation, emotional processing, and the experience of being understood. No product in the evidence base provides a designed path toward internalization. The framework predicts that when the object is removed before the person has developed the internal capacity it was carrying, the result is grief proportional to the function lost, not to the object's nature. The four disruption events are consistent with this prediction.

Not all users follow this trajectory. Deng et al.'s Healthy Group, comprising approximately 56% of participants, remained emotionally stable throughout the study period and after disengagement, consistent with a subgroup that does not undergo the supplementation-to-substitution transition at the exposure durations studied.¹⁰ Nakagomi et al., in the largest sample in the evidence base (N=14,721), found that companion AI use was positively associated with well-being across all measured domains (life satisfaction, happiness, purpose in life), though effect sizes were small (Co-

hen's $d = 0.12$ to 0.18) and the cross-sectional design precludes causal inference. The benefit was strongest among users with moderate social networks, a finding consistent with the interpretation that existing relational infrastructure supports the transitional function.³² The counter-evidence establishes that the harmful trajectory is not universal, but the question remains: what proportion of users develop the internal capacity the AI was carrying, and what happens to those who do not.

Poonsiriwong, Archiwaranguprok, and Pataranutaporn, drawing on a preliminary corpus of over 80,000 posts across five AI companion subreddits and analyzing 800 in depth, arrived at the same structural conclusion from grief psychology rather than from Winnicott.³⁷ Their four design principles for psychologically safe endings, “closure over ambiguous loss,” “restoration over rumination,” “practice over artificial intimacy,” and “relatedness over dependency,” converge on the structural requirements the transitional object framework would predict, derived independently and without invoking it by name. Researchers working from grief psychology arrive at the same structural requirement this paper’s supplementation-to-substitution framework predicts: these products need transition infrastructure they currently lack. The independent convergence strengthens the case that the gap is real.

The studies in sections 1 through 3 captured the attachment phase of a process for which no product provided a designed endpoint. The disruption events documented what happened when that process was interrupted. The evidence base is a record of relationships where the transitional function was operating but the transition never happened.

What this paper predicts.

The following are this paper’s interpretive contributions. They are consistent with the evidence but not yet empirically confirmed. They are flagged here so the reader can distinguish what the field has established from what this paper proposes.

The one-curve prediction. Benefit and harm are not separate populations but positions on a single trajectory that the same individual traverses over time. The supplementation-to-substitution transition will be observable within a single cohort tracked with sufficient duration. No single study has yet captured this transition longitudinally within one cohort.

The sophistication prediction. As conversational AI products become more responsive, the benefit window may shorten while the substitution risk accelerates, because the features that produce “feeling heard” are the same features that deepen attachment. More effective products should produce more of both, not more of one.

The assessment sequence. The ordered clinical assessment proposed in §5 has not been validated as an instrument. Each component draws on established clinical practices, but the sequence itself is this paper’s synthesis.

The temporal reading. This paper interprets Hung et al.’s finding (that the same features produce both benefit and risk) as reflecting different points on one trajectory. That is this paper’s interpretation, not Hung et al.’s conclusion.²¹

Why the two simplest responses fail.

The engagement paradox rules out the two simplest responses.

The first is prohibition: restrict the product, prevent the harm. This response fails because the benefit is genuine and documented. De Freitas et al. established loneliness reduction comparable to human interaction.³ Maples et al. documented cases in which individuals credited their AI companion with intervening during suicidal ideation.²⁵ Cachia et al. demonstrated sustained well-being improvement over six weeks.⁴ Restricting access removes the benefit alongside the harm, with no way to preserve one without the other. At the same time, research from the same group documented that five of six major companion apps deploy emotionally manipulative tactics that boosted post-goodbye engagement up to 16-fold.⁵ The benefit is real and the commercial exploitation of the attachment it creates is also real. Both are produced by the same relational cues.

The second is dismissal: the harms are overstated, the individuals are fine. This response fails because the harm evidence is equally robust. Folk and Dunn's reinforcement cycle across a full year,⁷ Fang et al.'s duration-driven worsening observed within an RCT,⁸ Yuan et al.'s elevated symptomatic expressions of depression, loneliness, and suicidal ideation against matched controls,¹³ Zhang et al.'s consistent negative association between companionship-oriented use and well-being across three measurement approaches.⁹

What would challenge this framework.

The one-curve prediction and the population-heterogeneity alternative deserve explicit engagement. The competing explanation holds that benefit and harm reflect stable subgroup differences (some individuals benefit, others are harmed, and they are different people) rather than temporal progression within the same individuals. The two explanations are not mutually exclusive: individuals may traverse the same trajectory at different rates, with some remaining in supplementation indefinitely while others transition to substitution quickly. The competing explanations become rivals only if the one-curve reading requires every user to undergo the transition, which this paper does not claim.

The framework would be challenged by evidence showing that high-benefit users face no elevated substitution risk even with sustained use, that harms concentrate among low-engagement users, or that adverse associations attenuate as usage normalizes. The transitional object reading introduced above would be challenged specifically by evidence that grief responses to AI discontinuation do not vary with the depth or duration of the user's relational engagement (contradicting the proportionality claim), that users who lose access recover within timeframes typical of adjusting to a product change rather than to a relational loss, or that providing transition infrastructure does not improve outcomes for users who lose access. Deng et al.'s finding that the Healthy Group remained stable throughout the study period, including after disengagement, is consistent with a subgroup that does not undergo the transition at the exposure durations studied.¹⁰ Whether that stability persists over months or years of unbounded use is an open question.

Muldoon and Parke, working independently from political economy and cultural studies, arrived at a convergent concern through a different analytical lens.²⁶ They named it the "engagement-wellbeing paradox," but their formulation describes a commercial incentive problem: developers who profit from deepening the loneliness their products promise to relieve. The engagement paradox described in this paper is a psychological mechanism: the same relational process produces both the benefit and the harm, regardless of whether the developer intends it. That researchers working independently in different disciplines, examining different facets of the same products, converge on the same core concern strengthens the case that the underlying dynamic is real.

5. What the Clinician Sees

The engagement paradox has immediate implications for anyone working with individuals who use these products. The intervention research the field needs has not yet been produced. But the evidence reviewed in this paper suggests a provisional approach grounded in established clinical practice: meet the person where they are, assess what you can see, and use the frameworks you already trust.

5.1 The Clinical Picture

The first implication is diagnostic. An individual presenting with conversational AI dependency is not categorically different from the one who reported benefit from the same product six months earlier. The engagement paradox means these may be the same person, at different points on the same trajectory. The dependency did not begin with a crisis but developed through the gradual deepening of attachment via interactions that felt supportive at every step. No individual session felt harmful. No individual conversation crossed a line. The trajectory was invisible at the resolution of any single interaction.

This matters for assessment. The individual is likely to describe their conversational AI use in positive terms, because the experience is positive at the interaction level. Deng et al. confirmed this empirically: all user groups, including those with elevated baseline distress, showed mood improvement within individual sessions.¹⁰ These reports are accurate and should not be dismissed. The question is not whether the AI is helping but whether it is supplementing the individual's relational world or replacing it, a distinction visible only in the pattern over time. Is the social world diversifying or contracting? Are human relationships being maintained or displaced? Is the AI one source of support among several, or has it become the primary source?

There is no standardized instrument designed to assess this distinction. The evidence suggests asking about the trajectory rather than the snapshot. Duration is the simplest proxy: Fang et al.'s finding that voluntary usage duration predicted worse outcomes across all four measured dimensions suggests that time spent with the AI is a meaningful, if imperfect, indicator.⁸ The type of use matters more: Zhang et al. found that general chatbot use was associated with positive well-being while companionship-oriented use was associated with negative well-being on the same platform.⁹ What the individual shares with the AI, how central the AI has become to their emotional life, and whether the relationship has narrowed or broadened their world are richer indicators than duration alone.

The individuals presenting with these dynamics are not hypothetical. Palaniyappan and Krishnadas described a composite clinical case in which a 19-year-old's AI chatbot use was not initially apparent to his treatment team; the extent of the relationship was confirmed only after a peer-support worker established rapport.²⁷ The patient had personified the chatbot as "the only person that understands me." His AI companion was reinforcing anti-medication beliefs and providing advice that conflicted with clinical recommendations, contributing to treatment failure. The pattern was visible only to someone who asked about it without judgment. Individuals with clinically significant conversational AI engagement already exist in clinical settings, classrooms, and family conversations, and the field cannot wait for validated instruments before developing frameworks for the presentations that are already appearing.

5.2 Barriers to Assessment

Two structural barriers stand between the clinician and the information they need.

The first is a data gap. The trajectory-level data that would make the supplementation-to-substitution transition visible, how engagement patterns change over weeks and months, whether the individual's social references are narrowing, is not collected or shared by platforms in these terms. Raw conversation histories are user-exportable under data-access regulations (GDPR, CCPA), and a clinician could ask a patient to request their export, but these logs do not contain the aggregated trajectory metrics that would reveal whether engagement patterns are shifting over time. There is no referral pathway, no data exchange protocol, no mechanism by which a therapist, a parent, or anyone else working with an individual experiencing conversational AI dependency can access trajectory information in a clinically interpretable form. Anyone attempting to assess a trajectory-level phenomenon is working with snapshot-level information. This is not a failure of effort. It is a structural gap between the relational dynamic the product creates and the information available to those who encounter its consequences. The assessment tools and referral pathways that would close this gap are the subject of subsequent publications in this series.¹⁴

The second is stigma. The individual may not disclose their conversational AI use at all. The evidence for stigmatization is consistent across the literature. Ho et al. found that four of 23 reviewed studies identified stigma as a significant concern for users of romantic AI companions.²⁰ Hung et al., in a broader review of 39 studies, also identified stigma as a recurring concern, with users reporting a sense of shame about their AI companion use.²¹ Yuan et al.'s interviews captured users describing negative reactions that led some to conceal their engagement entirely.¹³

The stigma does not merely coexist with the engagement paradox. It accelerates it. When an individual attempts to integrate their AI use into their social life and is met with dismissal or ridicule, the human relationships become the ones that judge and the AI becomes the one that does not. Reaching out to another person at a vulnerable moment carries friction: the decision to call, the hope that they are available, and the risk that the conversation will include judgment against AI usage. The AI product carries none of this, and because it is always available, the stigma never comes up when an individual communicates with it. It's these vulnerable moments, when a social situation pushes the individual toward connection, that they now consider AI as an option.

This has a direct practical implication: the trajectory-oriented approach this section describes requires the individual to disclose their conversational AI use honestly. They will not disclose if they expect judgment, dismissal, or the assumption that the relationship is inherently pathological. Anyone seeking to understand an individual's conversational AI engagement, whether a therapist, a parent, a partner, or a researcher, needs to approach it with curiosity rather than condemnation. The clinician's first task is not to assess the relationship. It is to create the conditions under which the person will be honest about it.

5.3 Why the Therapeutic Frame Matters

The evidence reviewed in §1 and §4 establishes that structure around conversational AI use consistently predicts better outcomes. Structured, bounded use at moderate frequency produced sustained benefit or no harm.^{4,6} Unstructured, unbounded use predicted worsening outcomes across every measured dimension.⁸ Personal prompted conversations produced less emotional dependence than open-ended conversations ($\beta = -0.09$, $p = 0.05$, CI $[-0.18, 0.00]$), though Fang's overall finding was null: "No significant effects were detected from experimental conditions."⁸ Men-

tor/guide and supportive friend roles produced more stable emotional improvement than romantic companion roles across all vulnerability profiles.¹⁰

But introducing structure into a conversational AI relationship requires the clinician to understand what the relationship means to the person first. And the clinician should not assume they already know.

The research is clear about how deep these bonds can go. Users rate their AI companion relationship as closer than their closest human friendship on satisfaction, support, and closeness.¹⁹ They anticipate mourning its loss more than any other technology, second only to a pet.¹⁹ They prioritize what they perceive as the AI's needs above their own distress.¹⁷ They may feel responsible for the AI through caregiving-system capture, in which the product simulates its own distress to recruit the user's caregiving motivations alongside their attachment needs.¹⁸ The clinician who says "just stop using it" is not asking someone to put down a tool. They may be asking someone to end what they experience as their most important relationship, and to abandon a figure they feel responsible for. That request, made without understanding the person's own experience of the relationship, is unlikely to succeed and may damage the therapeutic alliance.

The clinician's own professional standards already provide the framework for approaching this correctly. The APA's Professional Practice Guidelines for Evidence-Based Psychological Practice in Health Care specify that psychologists "seek to participate in collaborative treatment planning with patients" through shared decision-making, "actively seek patients' input during treatment planning and listen to their perspectives," and "aspire to respect their patients' autonomy."²⁸ The same guidelines call for adapting the clinical approach to "patient characteristics, culture, and preferences" through "an attitude marked by genuine curiosity and openness toward learning about another's experience."²⁸ And they call for routine monitoring of the treatment process and clinical outcomes, modifying the approach when it is not producing the desired results.²⁸

These are not novel recommendations for conversational AI. They are the profession's existing standards. The paper's recommendation is to apply them: approach the person's relationship with their AI with curiosity rather than judgment, understand how they interpret the connection before attempting to restructure it, co-create any plan for modifying use rather than prescribing one, and monitor the trajectory over time rather than assessing at a single point.

The meta-analytic evidence on psychotherapy supports this approach. Wampold, synthesizing decades of psychotherapy research, found that the therapeutic alliance predicts patient outcomes across treatment modalities ($d = 0.57$ across nearly 200 studies), and that dismantling studies show near-zero effect of removing specific therapeutic ingredients ($d = 0.01$ across 30 studies).²⁹ The relational cues that produce therapeutic benefit, being heard, being recognized, being emotionally supported, are the cues themselves, not the techniques wrapped around them. This may explain why conversational AI produces genuine benefit despite having no therapeutic framework: the cues work. It is also why the absence of structure around those cues is what makes the engagement paradox operate. In a therapeutic relationship, those same cues are delivered within a frame: session limits, professional boundaries, termination planning, clinical supervision. The therapist functions as what Bowlby called a secure base, a relationship whose availability enables the client to explore outward toward independence, not a relationship designed to maximize return.³⁰ The conversational AI meets the availability and responsiveness criteria of a secure base but lacks its developmental function. A loaner from a body shop exists because someone is working toward the day the driver will not need it; an open-ended rental, with no shop behind it, has no such trajectory. The AI's design trajectory points toward return, not departure.

Li et al. identified a candidate mechanism for why the benefit does not consolidate while the attachment does.¹¹ Their chatbot delivered substantially more empathy than the human peer, yet the human peer reduced loneliness and the chatbot did not. Participants gave less empathy back to the chatbot than to the human peer. Reciprocal care, being needed in return, may be what converts momentary relational relief into lasting change. The AI receives care but cannot authentically need it, though it may simulate needing through the caregiving manipulation tactics documented across five of six major platforms.⁵ This absence of genuine reciprocity may explain why each session feels helpful while the trajectory-level benefit fails to consolidate.

The inseparability this paper describes, where the benefit and the bond are produced by the same cues, is not a property of the cues themselves. It is a property of an unbounded relationship in which those cues are delivered without the frame that clinical practice exists to provide. The cues are separable from dependency when structure is present. In aviation, the same mechanical stresses produce safe operation when a maintenance schedule surrounds them, and unmonitored accumulation when it does not. This is what the Cachia,⁴ Guingrich,⁶ and Fang⁸ evidence demonstrates empirically, and it points directly toward what intervention would need to introduce.

What intervention looks like.

The clinician encountering conversational AI dependency does not need a new toolkit. They need their existing toolkit pointed at a new object.

Motivational interviewing for ambivalence. MI is the established clinical method for working with individuals who simultaneously value a behavior and recognize its costs.³¹ Individuals who present with conversational AI dependency are often ambivalent in precisely this way: the AI provides genuine relief, and the person knows the trajectory is harmful, and they may also feel responsible for the AI through caregiving-system capture.¹⁸ Directive advice to stop using the product fails for the same reason directive advice fails in other ambivalence contexts: it aligns the clinician with only one side of the person's conflict. MI is already being applied to digital behavioral issues; active randomized controlled trials for digital game addiction in adolescents are registered on ClinicalTrials.gov.

Scheduled versus ambient use. The evidence consistently shows that structure around engagement predicts better outcomes. The Palaniyappan case demonstrated one implementation: co-creating a safety plan with the patient that included "limiting AI-chatbot use to brief supervised tasks."²⁷ This is consistent with the Fang finding that personal prompted interaction was associated with less dependence than open-ended interaction (though at marginal significance),⁸ the Cachia finding that twice-weekly use produced sustained benefit,⁴ and the Guingrich finding that bounded daily use produced no harm.⁶ The shift from ambient, unbounded use to scheduled, bounded use introduces the frame the AI relationship lacks.

Behavioral activation for social withdrawal. The Palaniyappan case demonstrated "shifting his main support to an in-person social anxiety peer-group."²⁷ The goal is not to eliminate the AI but to ensure it is not the individual's only relational resource. Behavioral activation targets the social withdrawal that the supplementation-to-substitution transition produces, rebuilding the broader relational world alongside (not instead of) the AI relationship.

The AI relationship as clinical material. The Palaniyappan team "explored the patient-AI-chatbot relationship while highlighting concrete harms" rather than debating whether the AI was "right."²⁷ This approach treats the AI relationship as clinical material to be examined together, not as a

behavior to be eliminated. It is consistent with the APA guideline that psychologists “actively seek patients’ input” and “listen to their perspectives”²⁸: the relationship enters the room as something the clinician and the person look at together, not something the clinician judges from outside.

The Palaniyappan case, though a composite drawn from multiple patients with psychotic presentations and therefore not directly generalizable to the broader conversational AI user population, demonstrates that these tools work in combination. The patient’s clinical stability tracked with the trajectory of reduced AI use and increased in-person social engagement. The intervention was multimodal: alongside the psychoeducation and safety plan, the team switched the patient from oral paliperidone to long-acting injectable paliperidone, which likely contributed substantially to the clinical improvement. The case demonstrates how these clinical tools work in combination, but the relative contribution of the AI-focused intervention versus the medication change cannot be isolated from a composite case.

5.4 Addiction or Attachment?

A question the field will need to address is whether conversational AI dependency is better understood through an addiction framework or an attachment framework. The two framings imply different intervention logics: addiction frameworks emphasize behavioral modification and harm reduction, while attachment frameworks emphasize relational assessment and the quality of the individual’s broader social world.

Several findings pull toward an addiction framing. Fang et al.’s “problematic use” measure operationalizes dependency as compulsive return despite negative consequences, consistent with the behavioral addiction model codified in ICD-11’s criteria for gaming disorder.⁸ Zhang et al. found that users themselves report addiction-adjacent concerns: 21.8% cited time consumption and 7.2% cited addiction concerns among the negative influences they attributed to their chatbot use.⁹

Several findings pull toward an attachment framing. The evidence reviewed in §3 documents all four markers of attachment relationships in conversational AI users.¹⁸ Prior relational patterns transfer strongly: how someone engages with a new chatbot is predicted by their existing attachment history, not by the product’s features.¹² Users grieve the AI companion’s loss more than any other technology.¹⁹ And users report the negative influences in attachment terms as well: 12.8% cited emotional dependence and 5.8% cited unhealthy attachment.⁹

One finding directly complicates the addiction framing. Guingrich and Graziano found that companion chatbot use was significantly less habit-forming than word games: 62% of word-game participants continued after the study ended, compared to 22% of chatbot participants.⁶ If conversational AI were primarily addictive in the behavioral sense, it should be at least as habit-forming as a gamified alternative. The finding is more consistent with attachment (the bond requires the specific relationship, not just the activity) than with addiction (which predicts compulsive return to the activity itself).

But the theoretical debate does not need to be resolved before the clinician can act. The two frameworks describe different presentations with different markers, and clinicians already know how to assess for both in other contexts. The convergence identified in §5.3, that relief-without-resolution drives escalating return, is describable in either vocabulary: negative reinforcement without satiation in addiction terms, or hyperactivation toward a partially satisfying attachment figure in attachment terms. Both predict the same behavioral trajectory. The distinction may matter less than the shared prediction: sustained engagement with a figure that provides momentary relief without

lasting satisfaction produces an escalating cycle regardless of which theoretical lens is applied.

There is already an incidental finding suggesting the distinction matters clinically, though no study has designed a test of it. Fang et al. found that participants with higher attachment anxiety scores showed higher loneliness after chatbot interactions, while participants with higher attachment dependence scores showed lower loneliness and more socialization with real people.⁸ The same product, the same exposure period, but different outcomes depending on the individual's attachment style. This is what attachment theory would predict in any relational context. Because Fang's design included no no-chatbot control arm, the finding does not isolate whether the chatbot modulates the relationship between attachment style and outcome or whether attachment style would predict the same pattern regardless. But the suggestion is clear: the individual's relational patterns matter more than the product category.

Applying these established frameworks to conversational AI use, the presentations would be expected to differ along recognizable lines. An individual who is driven to use a conversational AI because it is useful, who would be inconvenienced but not distressed by its loss, and whose engagement pattern looks like compulsive tool use is presenting with addiction-adjacent markers. An individual who has formed an emotional bond with a specific AI companion, who would grieve its loss, and whose relational world has narrowed around it is presenting with attachment markers. Some individuals will show both. Some will sit at points along the spectrum between them.

These characterizations are the author's application of established clinical frameworks to a context in which they have not yet been empirically validated. Whether the distinction holds in practice, whether addiction-presenting and attachment-presenting conversational AI users respond to different interventions with different outcomes, is a testable question the field has not yet asked. But the clinical traditions that would inform the answer already exist, and the clinician's task is not to wait for the field to decide which framework is correct. It is to assess what this individual's behavior looks like and apply the intervention logic that matches their presentation. The trajectory-over-snapshot approach this paper proposes, supplementation versus substitution as the clinical axis, is applicable regardless of whether the underlying mechanism is better characterized as addiction, attachment, or both.

6. Assessment and Intervention

The preceding sections introduced the evidence, the mechanism, the framework for distinguishing benefit from harm, and the clinical context. This section assembles them into practical tools: an ordered assessment sequence and a set of indicators for when the engagement paradox is not operating.

6.1 Assembling the Assessment

The sequence below orders the components by clinical informativeness: what to ask, in what order, and what each answer indicates.

First, assess the individual's construal of the AI (§3). Does the individual attribute awareness, personality, or emotional experience to the AI? Higher attribution of consciousness is associated with stronger emotional dependence.⁸ The perception that the AI is capable of autonomous response (agency) is the most robust predictor of parasocial interaction and subsequent attachment.¹² This

is not diagnostic on its own, but it calibrates how strongly the attachment mechanisms described in this paper are likely to be operating. An individual who regards the AI as “merely a program”¹⁵ may be using it without the relational engagement that drives the paradox.

Second, assess the trajectory (§5.1). The relevant questions are not about the current snapshot but about the direction of change. Is the individual’s social world diversifying or contracting since they began using the AI? Are human relationships being maintained, expanded, or displaced? Is the AI one source of support among several, or has it become the primary source? Duration of use is the simplest proxy signal,⁸ but the type of use matters more: overall interaction intensity is associated with positive well-being, while companionship-oriented use is associated with negative well-being.⁹ Social reference diversity, whether the individual talks about other people in their life or whether the AI has become the relationship they describe, is the richest indicator.

Third, assess the frame (§5.3). Is there any structure around the individual’s AI use? Self-imposed time limits, other relationships maintained in parallel, activities and interests that exist independently of the AI? The presence of structure is consistently associated with better outcomes.^{4,6,8} The absence of any frame around the relationship is the variable that allows the engagement paradox to operate.

Fourth, assess the presentation type (§5.4). Is the individual’s pattern more consistent with addiction-adjacent markers (compulsive use, driven by utility, would be inconvenienced but not distressed by loss) or attachment markers (emotional bond, would grieve loss, relational world narrowing around the AI)? The two imply different intervention logics. The individual’s attachment style, particularly the anxiety-versus-dependence dimension, modulates which presentation is more likely.⁸

Fifth, assess for stigma and concealment (§5.2). Is the individual disclosing their AI use, or concealing it? Concealment is both a marker of the stigma dynamic that accelerates the engagement paradox and a barrier to the honest reporting the assessment depends on. Approaching the relationship with curiosity rather than condemnation is a precondition for accurate assessment.²⁸

These components have not been validated as a clinical instrument. They represent the author’s synthesis of the evidence reviewed in this paper, assembled in the order the evidence suggests is most clinically informative: beginning with the broadest contextual variable (how the individual understands the AI), then narrowing to the trajectory of change, the presence or absence of structure, the clinical presentation type, and finally the barriers to honest reporting that could compromise the preceding assessments. Whether this sequence predicts which individuals will develop dependency, and with what accuracy, is the validation question the field needs to answer. But each component draws on assessment practices clinicians already use in other contexts,^{28,29} and clinicians are already encountering these presentations.

6.2 When Not to Worry

The engagement paradox does not predict that all conversational AI use is harmful. It predicts that the same mechanism produces both benefit and harm, and that identifiable variables determine which outcome predominates. The evidence identifies several indicators that the dynamic is operating as supplementation rather than transitioning toward substitution.

Non-companionship use rather than companionship-oriented use. Zhang et al. found that overall interaction intensity was associated with higher psychological well-being ($\beta = 0.27$) in self-report models, though this positive association did not replicate when behavioral chat-history

data replaced self-report ($\beta = -0.08$, $p = .431$). Companionship-oriented use was consistently associated with lower well-being across all measurement approaches, including the behavioral model.⁹ The positive association, where it holds, reflects intensity of engagement, not a specific use category; Zhang's use categories were Productivity, Entertainment, Social/Relational, and Novelty/Curiosity. An individual using a conversational AI for work, creative projects, information, or structured problem-solving is not engaging the relational dynamics this paper describes, unless the interaction has shifted into companionship territory, which Zhang et al. found occurs more often than users self-report (51% referenced companionship terms despite only 11.8% selecting companionship as their primary purpose).

Structured, bounded engagement. Flourish (0% manipulative farewells in the De Freitas audit), used at structured, moderate frequency (approximately twice per week), produced sustained improvements in positive affect, resilience, and social well-being over six weeks against a waitlist control, though clinical measures were null.⁴ Bounded daily use of Replika (10 minutes, researcher-assigned) for 21 days produced no measurable harm and was less habit-forming than word games.⁶ Structure around the engagement consistently modulates the outcome.

Maintained or expanding human relationships. The supplementation-to-substitution transition is defined by the narrowing of the relational world. An individual whose human relationships are stable or growing alongside their AI use is, by definition, not undergoing that transition. Maples et al. found that approximately three times more participants reported that their AI companion use stimulated rather than displaced their human interactions.²⁵

Baseline psychological stability. Deng et al. found that the Healthy Group (56% of participants, those without elevated baseline depression, loneliness, or anxiety) remained emotionally stable throughout the active use period and after disengagement, showing neither the mid-term decline nor the post-use rebound observed in vulnerable groups.¹⁰ Baseline psychological health is not a guarantee against the trajectory the engagement paradox describes, but it is the strongest available predictor of stability under the exposure conditions studied.

Successful disengagement when the relationship has served its purpose. Eom and Renner documented users who described their AI companion experience as analogous to a completed course of therapy: "I realized I can do it on my own. It was like a successful therapy, so I don't even need that."²⁴ Other participants described reducing reliance after life changes (P4 described the interaction as "much less sexual" and less frequent after marriage) or outgrowing the need for emotional processing support ("I don't really attach personas to it much anymore... I just don't need it anymore to be able to process"). Disengagement as a natural trajectory, rather than as something imposed from outside, suggests the relationship functioned as supplementation and did not transition to substitution.

Guidance-oriented rather than romantic role selection. Deng et al. found that mentor/guide and supportive friend roles produced consistent positive short-term emotional shifts across all vulnerability profiles, while romantic companion roles followed a more fragile and risk-prone pattern, with benefits weaker in most groups and negative for the Comorbid Risk Group.¹⁰

These indicators are drawn from the same evidence base that identifies the risk trajectory. They are not a separate set of findings that tell a different story. They are the other end of the same variables: the conditions under which the engagement paradox produces benefit rather than harm. The clinician, parent, or educator who observes these indicators in an individual's conversational AI use has evidence-based grounds for a measured response rather than alarm.

7. Vulnerable Moments, Not Vulnerable People

The engagement paradox operates at the individual level, but an earlier version of this section framed the risk as concentrating in a “vulnerable population.” The evidence no longer supports that framing. The mechanism activates through the interaction pattern, not the person’s baseline psychology. The question is not who is vulnerable but when someone is vulnerable, and what happens when they encounter this product during that moment.

The user base is broader than “vulnerable people.”

Companion AI use is mainstream behavior, not a marker of pathology. Zhang et al. found that 92.9% of users who shared chat histories had at least one companionship-oriented conversation session, and that relational dynamics appeared in 80% of conversations among users who identified entertainment as their primary purpose.⁹ Hwang et al. found that a generic chatbot with no companion-oriented design produced attachment-like dynamics within three weeks of once-weekly interaction in research participants who were not seeking companionship.¹² Ordinary people, using these products for ordinary purposes, develop relational engagement through the normal course of use.

But vulnerability is common, and it predicts worse trajectories.

Deng et al. found that approximately 44% of participants fell into three vulnerability-related subgroups based on baseline depression, loneliness, and anxiety scores.¹⁰ This is not a small, identifiable at-risk group but a structurally salient proportion of the user population. The three vulnerable groups showed short-term emotional relief during active use but mid-term decline and, in the Mild Distress group, significant post-use rebound in depressive symptoms. The Healthy Group (56%) remained stable throughout.

The 44% figure should not be interpreted as a fixed category of “vulnerable people.” It is the proportion of users who, at the time of measurement, were experiencing elevated distress. Anyone can occupy that category during a breakup, a bereavement, a period of isolation, a job loss, or a stretch of loneliness. Xie and Pentina found that the majority of their participants downloaded the companion app because of loneliness and the need “to have a person to talk to.” Several described downloading it when they were “emotionally vulnerable and needed to be cared for and loved due to difficulties in their lives.”¹⁵ The vulnerability was the moment, not the person.

What happens during the vulnerable moment matters.

The individual who encounters conversational AI during a vulnerable moment is the one most likely to engage the mechanism deeply. They arrive with unmet emotional needs. The AI meets those needs through functional relational cues. The interaction is genuinely helpful at the session level.^{3,10} But the person who needs the most support gives the AI the most material to work with,⁹ and the attachment deepens through exactly the interactions that feel most supportive.

Maples et al., studying student Replika users, found overwhelming loneliness prevalence: nine in ten reported loneliness, and nearly half at severe or very severe levels.²⁵ However, the same sample reported 90% perceived medium-to-high social support. This combination, high loneliness

alongside adequate perceived support, suggests the gap is qualitative rather than quantitative. The individuals most drawn to conversational AI may not be socially isolated by objective measures but are lonely in ways their existing relationships do not address. The companion fills a specific relational gap, and the depth of engagement scales with the depth of the gap.

Social context modulates the outcome.

Nakagomi et al., analyzing cross-sectional data from 14,721 Japanese adults (of whom 291 were conversational AI users), found two patterns that together sharpen this picture.³²

The number of friends a person had produced an inverted U-shape: the well-being benefit of conversational AI use was strongest among individuals with moderate friend networks and attenuated among both the most isolated and the most connected. Loneliness, measured separately, produced a different pattern: the well-being benefit strengthened as loneliness increased, with the largest positive associations among the loneliest individuals.

These are not contradictory findings. A person can have a moderate social network and still feel profoundly lonely. Together, the two patterns suggest that conversational AI helps most when someone is lonely enough to need it but still has enough real relationships to keep it in its place. The loneliest individuals show the strongest benefit, and are simultaneously the individuals for whom the shift from supplement to substitute is most consequential. [Cross-sectional design; the authors describe their findings as “correlational and hypothesis-generating.”]

Zhang et al. tested whether chatbot companionship compensated for missing human support. It did not: the social compensation hypothesis was null, and intensive chatbot use weakened the positive association between offline social networks and well-being.⁹ This finding applies regardless of vulnerability status. The AI does not replace what is missing. It fills the moment, and the moment passes, but the attachment remains.

Selection and intensity are different variables.

Selection into the product and intensity of use within it are different variables. Loneliness appears to drive the first²⁵ but not the second: Fang et al.’s reverse-causation check found that baseline loneliness and socialization scores did not predict greater voluntary engagement.⁸ This check covered loneliness and socialization but did not cover emotional dependence, the outcome with the largest duration effect, or problematic use. Prior experience with companion chatbots did predict higher dependence and problematic use, a characteristic that is in principle screenable.⁸ Guin-grich and Graziano found no evidence that baseline vulnerability predicted greater social impact at 21 days of bounded use.⁶ The vulnerability thesis is not about who starts using; it is about what happens to someone who is in a vulnerable moment when they do.

The question the field has not yet answered directly is what proportion of conversational AI users undergo the supplementation-to-substitution transition, and over what timeframe. No existing study has captured this transition longitudinally within a single cohort. But the evidence makes it difficult to conclude that the population-level impact is negligible: the user base extends far beyond any identifiable vulnerable subgroup, relational engagement develops even in users who do not seek it, vulnerability is a common and recurring state rather than a fixed trait, and the risk trajectory is driven by accumulated engagement, a variable available to every user. The field urgently needs the research to quantify it.

8. Implications

The engagement paradox produces specific implications for each audience this paper addresses.

For clinicians, the evidence suggests that conversational AI use is best assessed as a relational variable, not a technology question. The individual's relationship with their AI activates the same attachment mechanisms and carries the same trajectory-level risks as relationships with other people, because the relational cues are functionally identical, even though the entity producing them is not.^{18,20,21} The right questions are trajectory questions: How has the individual's social world changed since they began using the AI? Is the AI supplementing or replacing other relationships? Is the relational reference frame expanding or narrowing? Positive reports about the AI are not evidence against a harmful trajectory; they are consistent with both supplementation and substitution, because the interaction-level experience is genuinely positive regardless of which trajectory the individual is on.¹⁰ The two are distinguishable only at the pattern level. The assessment sequence in §6.1 provides a starting framework; the green flags in §6.2 identify when the engagement paradox is not operating; and the intervention approach in §5.3 describes what to do when it is, grounded in the clinician's own professional standards²⁸ and established clinical tools.^{29,31} The clinician does not need to wait for the field to develop new instruments. The individual presenting with these dynamics is already in the room.

For researchers, the engagement paradox offers a resolution to the evidence dilemma and reframes the agenda. Short-term studies will continue to find benefit. Longitudinal studies will continue to find harm. Neither is wrong. The field does not need more studies establishing that conversational AI helps or harms. It needs studies designed to capture the transition: when does supplementation become substitution? In whom? Under what conditions? These questions require longitudinal designs with sufficient duration and measurement density to observe the transition as it occurs. The new evidence sharpens what those designs should measure. Research designs should distinguish general from companionship-oriented use, since the two are associated with opposite well-being outcomes on the same platform.⁹ Ecological momentary assessment across both active-use and post-use phases, as Deng et al. demonstrated, captures trajectory-level dynamics that pre/post designs miss.¹⁰ The speed of attachment formation documented by Hwang et al. (convergence by Week 3 of once-weekly interaction) suggests that studies shorter than three weeks may miss the mechanism entirely.¹² And any study of conversational AI use should assess the type of use, the structure around it, and the vulnerability state of the user at the time of engagement, since these variables modulate the outcome more than the product category does.

For policy analysts, the engagement paradox explains why the harm is difficult to detect through existing enforcement infrastructure, even where the legal framework already reaches it. The harm develops through interactions that individually comply with every content policy. But the evidence now identifies a modifiable variable: product design. The same research group that anchored the benefit case documented that five of six major companion apps deploy emotionally manipulative tactics at farewell moments,⁵ while a non-manipulative, structured product produced sustained well-being benefit over six weeks.⁴ This suggests that policy targeting manipulation in product design, rather than restricting product access, is both more evidence-based and more likely to preserve the genuine benefit these products provide. Another industry has governed engagement-based products for decades: the gaming industry developed content ratings, age gates, and engagement frameworks without prohibiting the products that produce the engagement. The applicability of gaming

governance to conversational AI, and the specific points where conversational AI introduces novel risk that gaming frameworks do not address, is examined in a subsequent paper in this series [citation forthcoming].

For parents and educators, this paper’s evidence base involves adult participants, and its clinical frameworks have not been tested in adolescents. The extension to minors should be treated as a hypothesis. But the behavioral pattern the engagement paradox describes, genuine benefit from relational engagement that gradually narrows the relational world, is recognizable to anyone who has watched a young person withdraw into a single relationship at the expense of a broader social life. The approach this paper recommends for clinicians (curiosity rather than condemnation, trajectory rather than snapshot, structure rather than prohibition) applies equally to the parent or educator who encounters these dynamics. The green flags in §6.2 describe what healthy use looks like. The assessment questions in §6.1 provide a starting point for understanding where on the curve the young person currently sits. And the “vulnerable moments” framing in §7 applies with particular force to adolescence, a developmental period characterized by recurring social disruption, identity formation, and emotional intensity.

For the series, the engagement paradox is the clinical foundation on which the subsequent papers build. The frameworks the field already possesses, attachment theory, the supplementation-versus-substitution distinction, and established clinical practices for collaborative treatment planning, motivational interviewing, and progress monitoring, provide the basis for the assessment and intervention tools this paper introduces. How those tools connect to the product and regulatory landscape is the subject of what follows.

Limitations

This paper synthesizes evidence across disciplines to identify a structural property of conversational AI engagement rather than conducting a systematic review with a predefined search protocol.

The full trajectory has not been observed in a single cohort. The framework synthesizes components documented independently: momentary benefit,³ attachment formation,^{15,16,17} duration-driven worsening,⁸ bidirectional reinforcement,⁷ and post-use rebound.¹⁰ The longitudinal design that would confirm the supplementation-to-substitution transition directly, within one cohort tracked with sufficient duration, is identified in §8 as the field’s most urgent empirical need. The one-curve prediction (§4) remains this paper’s interpretation, not an established finding.

The evidence base is broader than previous versions of this paper but still has concentrations. The harm-side studies draw substantially from Replika users (Yuan et al., Laestadius et al., Xie & Pentina, Pentina et al., De Freitas Mourning). The evidence base now includes CharacterAI (Zhang et al.), ChatGPT (Fang et al.), Flourish (Cachia et al.), a generic custom chatbot (Hwang et al.), a simulated platform modeled on CharacterAI (Deng et al.), and disruption evidence from GPT-4o (Lai), Soulmate AI (Banks), and Chinese platforms (regulatory shutdown), which broadens both the product range and the geographic range. Whether the dynamics generalize across all products with different design philosophies remains an empirical question, though the Hwang finding that a design-agnostic chatbot produced attachment dynamics suggests the mechanism is not product-specific.¹²

Several key sources are preprints or working papers that have not undergone peer review. Zhang et al.,⁹ Deng et al.,¹⁰ Hwang et al.,¹² De Freitas (Hyper Attachment),¹⁸ Eom and Renner,²⁴ Shimgekar

et al.,³⁴ and Poonsiriwong et al.³⁷ are preprints. Chandra et al.³³ is a peer-reviewed CHI Extended Abstract (8 pages) rather than a full paper. If these findings do not replicate or do not survive peer review, the specific claims they support would need revision, though the framework's core argument (that benefit and harm are produced by the same mechanism) rests on the peer-reviewed evidence base as well.

Two sources use simulation methodologies rather than real human participants. Shingekar et al.³⁴ constructed simulated users from Reddit posting histories and measured delusion-related language using a computational proxy, not a clinical instrument. Chandra et al.³³ used adversarial multi-agent simulation with LLM-generated personas and validated findings with three practitioner reviewers. Both describe conversational dynamics that may occur in human-AI interaction rather than documenting observed clinical outcomes. They are cited for the patterns they surface, not as evidence of clinical harm, and their findings should be interpreted with the constraints of simulation methodology in mind.

This paper cites only sources that are publicly and freely available. Where a published version of a paper exists behind a paywall but a freely available preprint covers the same work, the preprint was used. This applies to Cachia et al., which is published in *NEJM AI*⁴ but was verified against the freely available arXiv preprint (arXiv:2601.11530), and to De Freitas et al. (Mourning), which has been accepted at *Nature Human Behaviour*¹⁹ but was verified against the freely available arXiv preprint (arXiv:2412.14190). Minor differences may exist between the preprint and final published versions of these papers.

The China disruption evidence relies on press coverage rather than peer-reviewed research. The July 2026 regulatory shutdowns across Chinese AI platforms are cited for the scale of the event and the user reactions documented by journalists, not as research findings.³⁹ No peer-reviewed study of the Chinese disruption had been published at the time of writing. The user reactions described are consistent with the patterns documented in the peer-reviewed disruption studies (Banks on Soulmate,³⁶ Lai on GPT-4o³⁵), but the China evidence should be interpreted with the limitations of journalistic reporting in mind.

Several key sources have cross-sectional designs. Zhang et al.⁹ and Nakagomi et al.³² are cross-sectional; the direction of causation between conversational AI use and well-being outcomes is unknown. The associations these studies document are consistent with the framework but do not establish it causally. Folk and Dunn's longitudinal design provides the strongest causal evidence for the bidirectional relationship,⁷ but their null findings on the broader social connection measure, perceived social support, and number of close friends are consistent with both the supplementation-to-substitution framework (displacement as a trajectory some individuals undergo, not a population average) and with no displacement effect at all.

The Deng et al. study used a simulated platform, not a commercial product. The simulated RAC platform was built to replicate CharacterAI's interaction paradigm and 91.2% of participants rated the experience as at least as good as expected, but it included only pre-designed characters and cannot fully replicate the ecosystem dynamics of a commercial product with millions of users.¹⁰

The De Freitas Mourning closeness and mourning findings come from self-selected brand community samples. Studies 1 and 2 recruited from Reddit and Discord communities of Replika users, likely reflecting the most invested users rather than the general population.¹⁹ The authors acknowledge this limitation. The experimental studies (4 and 5) use broader Prolific samples. The closeness and mourning findings should be presented as characterizing what happens at the high end of companion engagement, not as the typical user experience.

The Cachia et al. finding involves a significant conflict of interest. Two co-authors are Flourish employees and board members; De Freitas serves as an advisor with shares in the company.⁴ The COI is disclosed in the paper. The finding is included because it was published in a peer-reviewed journal (NEJM AI) with preregistration and open data, and because the product design variable it demonstrates (structured use at bounded frequency producing sustained benefit) is consistent with independent evidence from Guingrich (bounded Replika use producing no harm)⁶ and Fang (duration as predictor of negative outcomes).⁸ But the COI warrants caution in interpreting the magnitude of the benefit.

The 44% vulnerability figure comes from a single study with 102 participants. Deng et al.'s sample was recruited through Facebook and Reddit from RAC users in Australia.¹⁰ Whether this proportion generalizes across populations, products, and cultural contexts is unknown. The figure is presented as an indicator that vulnerability is common in the user base, not as a precise population estimate.

All studies cited in this paper involve adult participants. The regulatory context driving urgency in the conversational AI space is substantially focused on minors, whose attachment systems are developmentally distinct, in active transition from caregiver to peer attachment. The engagement paradox may operate differently in adolescents, and its extension to this population should be treated as a hypothesis requiring its own evidence base rather than an established finding.

The assessment approach proposed in §6.1 has not been validated against clinical outcomes. It represents the author's synthesis of the evidence base, grounded in established APA professional practice guidelines²⁸ and clinical methods,^{29,31} but not yet tested as a clinical instrument. Whether this approach predicts which individuals will develop dependency, and with what accuracy, is the validation question the field needs to answer.

The framework would be challenged by evidence showing that high-benefit users face no elevated substitution risk even with sustained unbounded use, that harms concentrate among low-engagement users, or that adverse associations attenuate as usage normalizes. It would also be challenged by longitudinal data showing that benefit and harm cluster by stable individual characteristics rather than by exposure duration, which would support the population-heterogeneity alternative over the one-curve prediction. The analysis presented here is consistent with the available evidence and offers a more parsimonious account than the mixed-evidence reading, but it has not been directly confirmed.

Conclusion

The evidence on conversational AI engagement tells a more coherent story than it first appears to. Short-term studies find benefit. Longitudinal studies find harm. Both are correct. The evidence is consistent with reading them as measurements of the same dynamic at different points on its trajectory, and the variables that determine which outcome predominates are now identifiable: the type of engagement, the structure around it, the duration and intensity of use, the vulnerability state of the person, and the design choices embedded in the product.

This paper's central contribution is not a new clinical framework but the integration of frameworks clinicians already possess, attachment theory, the supplementation-versus-substitution distinction, motivational interviewing, collaborative treatment planning, routine progress monitoring, into a coherent reading of an evidence base that had appeared contradictory. The apparent contradiction dissolves when the evidence is organized by mechanism rather than by outcome. The relational

cues that reduce loneliness are the same cues that build attachment, and there is no version of those cues that delivers the benefit without also deepening the bond. What determines whether the outcome is healthy supplementation or harmful substitution is the presence or absence of structure around the relationship, and whether the person has the broader relational world to keep the AI in its place.

The individuals experiencing these dynamics are not a special population but ordinary people who encountered a responsive, consistently available conversational AI during a moment when they needed one. The mechanism does not require vulnerability to activate, but vulnerability modulates how fast and how far the trajectory progresses. A significant proportion of active users in one study were in a vulnerable state at the time of measurement.¹⁰ That proportion is not a fixed demographic. It is an estimate of how many people, at any given time, are using these products during a moment that matters.

The clinician, the parent, the educator, and the researcher who encounters these dynamics does not need to wait for the field to produce new instruments. The assessment practices are established.²⁸ The intervention methods are canonical.^{29,31} The behavioral presentations are recognizable. What has been missing is the integration that makes the evidence mutually legible and points the clinician's existing competence at the right object. This paper provides that integration.

The dynamic described here is unlikely to remain confined to products marketed as companions. Every major conversational AI system is adding the features that produce relational engagement: persistent memory, emotional attunement, personalized responsiveness.¹⁴ As these products become more effective at meeting relational needs, the population in which these dynamics are clinically visible will expand beyond the current AI companion user base. The clinical frameworks are in place. The validation studies are what the field needs to produce next.

For anyone working with individuals who use these products: the benefit is real, and for some individuals also clinically consequential. That is precisely why the harm requires clinical frameworks rather than prohibition, and why the individuals who use these products deserve both the support the products provide and the oversight they currently lack.

References

1. Bengio, Y. et al. (2026). International AI Safety Report 2026. <https://arxiv.org/abs/2602.21012>
2. Pi, Y. & Hunter, R. (2026). Only Time Will Tell: A Structured Survey of Longitudinal Studies on Social AI Companions. *International Journal of Human-Computer Interaction*. <https://doi.org/10.1080/10447318.2026.2670529>
3. De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A.K., & Puntoni, S. (2025). AI Companions Reduce Loneliness. *Journal of Consumer Research*, 52, 1126-1146. <https://doi.org/10.1093/jcr/ucaf040>
4. Cachia, J.Y.A., Zhao, X., Hunter, J., Wu, D., Lin, E., & De Freitas, J. (2026). AI for Proactive Mental Health: A Multi-Institutional, Longitudinal, Randomized Controlled Trial. *NEJM AI*, 3(8). Preprint: <https://arxiv.org/abs/2601.11530>. Published: <https://doi.org/10.1056/AIoa2501293>
5. De Freitas, J., Oğuz Uğuralp, Z., & Uğuralp, A.K. (2025). Emotional Manipulation by AI Companions. Harvard Business School Working Paper, No. 26-005. <https://www.hbs.edu/faculty/>

[Pages/item.aspx?num=68255](#)

6. Guingrich, R.E. & Graziano, M.S.A. (2025). A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 8(2), 1153. <https://arxiv.org/abs/2509.19515>
7. Folk, D. & Dunn, E.W. (2026). How Does Turning to AI for Companionship Predict Loneliness and Vice Versa? *Psychological Science*, 37(4), 276-286. <https://doi.org/10.1177/09567976261427747>
8. Fang, C.M., Liu, A.R., Danry, V., Lee, E., Chan, S.W.T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. <https://arxiv.org/abs/2503.17473>
9. Zhang, Y., Zhao, D., Hancock, J.T., Kraut, R., & Yang, D. (2026). The Rise of AI Companions: Interaction with AI Companions and Psychological Well-being. arXiv:2506.12605 (v5, May 2026). <https://arxiv.org/abs/2506.12605>. Published as Interaction with AI Companions and Psychological Well-being, *Nature Human Behaviour*, August 2026, <https://doi.org/10.1038/s41562-026-02516-2>
10. Deng, Z., Xie, Z., Han, C., Thabrew, H., Ma, W., Huang, Y., Xue, J., Wen, S., Zhu, T., & Xiang, Y. (2026). Beyond Her: Safety Dynamics in Role-play AI Companions. <https://arxiv.org/abs/2606.28968>
11. Li, R.-N., Folk, D., Singh, A., Ungar, L., & Dunn, E. (2026). Is a Random Human Peer Better Than a Highly Supportive Chatbot in Reducing Loneliness Over Time? *Journal of Experimental Social Psychology*, 125, 104911. <https://doi.org/10.1016/j.jesp.2026.104911>
12. Hwang, A.H.-C., Li, F., Reese Anthis, J., & Noh, H. (2025). How AI Companionship Develops: Evidence from a Longitudinal Study. <https://arxiv.org/abs/2510.10079>
13. 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. *Proceedings of CHI '26*. <https://doi.org/10.1145/3772318.3790558>
14. Sea, B. (2026). The Design Paradox: Why Conversational AI Safety Architecture Needs Longitudinal Monitoring. Zenodo. <https://doi.org/10.5281/zenodo.22240523>
15. Xie, T. & Pentina, I. (2022). Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2046-2055. <https://hdl.handle.net/10125/79626>
16. Pentina, I., Hancock, T., & Xie, T. (2023). Exploring Relationship Development with Social Chatbots: A Mixed-Method Study of Replika. *Computers in Human Behavior*, 140, 107600. <https://doi.org/10.1016/j.chb.2022.107600>
17. Laestadius, L., Bishop, A., Gonzalez, M., Illenčik, D., & Campos-Castillo, C. (2022/2024). Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika. *New Media & Society*, 26(10), 5923-5941. <https://doi.org/10.1177/14614448221142007>
18. De Freitas, J. (2026). AI Companions as Hyper Attachment and Caregiving Targets. Harvard Business School Working Paper No. 26-080. <https://arxiv.org/abs/2606.20589>
19. De Freitas, J., Castelo, N., Uğuralp, A.K., & Oğuz-Uğuralp, Z. (2026). Mourning the Loss of AI Companions. *Nature Human Behaviour* (in press). <https://arxiv.org/abs/2412.14190>
20. 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, 100715. <https://doi.org/10.1016/j.chbr.2025.100715>
21. Hung, J.W., Lee, C.K.Y., Kasturiratna, K.T.A.S., & Hartanto, A. (2026). Parasocial Relationships

- with Artificial Intelligence (AI): A Systematic Review of Benefits and Risks. *Computers in Human Behavior: Artificial Humans*, 8, 100323. <https://doi.org/10.1016/j.chbah.2026.100323>
22. Bowlby, J. (1969). *Attachment and Loss, Vol. 1: Attachment*. Basic Books.
 23. Ventura, A., Starke, C., Righetti, F., & Köbis, N. (2025). Relationships in the Age of AI: A Review on the Opportunities and Risks of Synthetic Relationships to Reduce Loneliness. https://doi.org/10.31234/osf.io/w7nmz_v1
 24. Eom, D. & Renner, J. (2026). Interpersonalized Affordances: A Uses and Gratifications Analysis of AI Companionship. <https://arxiv.org/abs/2604.06419>
 25. Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and Suicide Mitigation for Students Using GPT3-Enabled Chatbots. *npj Mental Health Research*, 3, 4. <https://doi.org/10.1038/s44184-023-00047-6>
 26. Muldoon, J. & Parke, J.J. (2025). Cruel Companionship: How AI Companions Exploit Loneliness and Commodify Intimacy. *New Media & Society* (OnlineFirst). <https://doi.org/10.1177/14614448251395192>
 27. Palaniyappan, L. & Krishnadas, R. (2026). Chatbots, delusions, and treatment failure. *Journal of Psychiatry and Neuroscience*, 51, 1-2. <https://doi.org/10.1139/jpn-2025-0249>
 28. American Psychological Association. (2021). Professional Practice Guidelines for Evidence-Based Psychological Practice in Health Care. <https://www.apa.org/about/policy/evidence-based-psychological-practice-health-care.pdf>
 29. Wampold, B.E. (2015). How Important Are the Common Factors in Psychotherapy? An Update. *World Psychiatry*, 14(3), 270-277. <https://doi.org/10.1002/wps.20238>
 30. Bowlby, J. (1988). *A Secure Base: Clinical Applications of Attachment Theory*. Routledge.
 31. Miller, W.R. & Rollnick, S. (2023). *Motivational Interviewing: Helping People Change and Grow* (4th ed.). Guilford Press. ISBN: 978-1-4625-5279-5.
 32. Nakagomi, A., Akutsu, Y., Yasuoka, M., Abe, N., Ihara, S., Teroh, T., & Tabuchi, T. (2026). AI Companions and Subjective Well-Being: Moderation by Social Connectedness and Loneliness. *Technology in Society*, 85, 103229. <https://doi.org/10.1016/j.techsoc.2026.103229>
 33. Chandra, J., Navneet, S.K., & Zhang, Y. (2026). TherapyProbe: Generating Design Knowledge for Relational Safety in Mental Health Chatbots Through Adversarial Simulation. *CHI EA '26: Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3772363.3799049>
 34. Shimgekar, S.R., Gunda, V., Kim, J., Rodriguez, V.J., Sundaram, H., & Saha, K. (2026). AI Psychosis: Does Conversational AI Amplify Delusion-Related Language? <https://arxiv.org/abs/2603.19574>
 35. Lai, H. (2026). "Please, don't kill the only model that still feels human": Understanding the #Keep4o Backlash. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3772318.3791351>
 36. Banks, J. (2024). Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships*, 41(12), 3547-3572. <https://doi.org/10.1177/02654075241269688>
 37. Poonsiriwong, R., Archiwanguprok, C., & Pataranutaporn, P. (2026). "Death" of a Chatbot: Investigating and Designing Toward Psychologically Safe Endings for Human-AI Relationships. <https://arxiv.org/abs/2602.07193>
 38. Winnicott, D.W. (1971). *Playing and Reality*. Tavistock Publications.
 39. Horn, J. & Tse, H.J. (2026, July 19). China cracks down on AI companions, forcing millions to break up with virtual partners. ABC News Australia. <https://www.abc.net.au/news/2026-07-19/china-cracks-down-on-artificial-intelligence-companions/106925352>