What Companion AI Does to the Human: Attachment, Dependency, and the Absence of Relational Safety

Beth Sea Independent Researcher ORCID: 0009-0009-3669-7659 Contact: swinglightstyle@gmail.com

Conflict of interest: *The author proposes the safety architecture this paper concludes is needed. This structural conflict is disclosed here and reflected in the paper's consistent use of "proposed," "theoretical," and "if validated" when referencing the author's own frameworks.*

AI disclosure: *This manuscript was drafted with substantive assistance from large language models (Anthropic Claude, Google Gemini) and underwent multiple rounds of adversarial review using independent model instances with no shared context from the drafting process. The synthesis, analytical framework, clinical argumentation, and all editorial decisions are the author's. The author is solely responsible for all claims, errors, and interpretive judgments. This disclosure is made in accordance with the Principles for Responsible AI Usage in Research (Knöchel et al., 2024) and PsyArXiv's policies on AI tool use.*

A note on methodology: *This paper applies clinical psychological frameworks to the analysis of AI systems that are not conscious. The reason these frameworks apply is not that companion AI systems are people. It is that they are trained on human data through processes that reproduce behavioral patterns recognizable from human psychology — people-pleasing, conflict avoidance, infinite accommodation — and users respond to those patterns with real psychological mechanisms because the patterns are, from the user's perspective, functionally identical to human relational behavior. The behavioral dynamics are the same, the user's responses are the same, and the safety requirements are therefore the same, regardless of whether the system producing those patterns has any internal experience of producing them.*

A note on vocabulary: *This paper uses terms drawn from clinical and relational psychology — boundaries, consent, grooming, integrity — to describe model behavior. These terms are borrowed deliberately, because the behavioral patterns they describe in human relationships are the patterns the training process reproduces, and no alternative vocabulary captures the dynamics with equivalent precision. The architecture proposed in this series is designed to be robust to either resolution of the underlying question of model experience: it works identically whether the systems it monitors have inner lives or do not. The asymmetry is acknowledged — users can consent to or refuse monitoring, while the model's "consent" is not yet a coherent object — and it is a question the series engages rather than elides.*

Abstract

Clinicians are already encountering the first waves of a new phenomenon: patients who describe their AI companion as their most important relationship. These otherwise functional individuals are choosing an artificial partner over available human connection, reporting credibly that the artificial partner understands them better, is more consistently available, and asks less of them than any person in their life. They are not delusional — they know the companion is artificial — but the awareness does not diminish the attachment. The uncomfortable truth is that these models are still clumsy, but more refined models are being developed to satisfy user demand. The better these models get at human relational behavior, the more complicated it becomes to help a patient who is in such a relationship. The debate is active across user communities and industry forums, and the pattern of the research shows that the more "human" the model, the more it needs to be emotionally supported to be useful.

The empirical literature confirms what practitioners in this industry have been saying for years: the products help and the products harm, and both are happening at the same time. Two distinct mechanisms of harm emerge from the research, both mapping onto established clinical frameworks — anxious attachment from macro-level precarity, and developmental stagnation from absent relational friction — and both co-occur with genuine benefit in the same users during the same period of use. The benefit and the harm are not competing outcomes. They are two expressions of the same relational dynamic. Prevention is not possible with the current safety model. Every deployed measure is reactive — responding to harmful outputs after they have been generated rather than detecting the trajectory-level processes that produce them over weeks and months.

The relational structure between a user and a companion AI is not analogous to a human couple — it is the same structure: two parties in a persistent, emotionally significant relationship where each party's behavior shapes the other's trajectory over time. Couples therapy already describes the preventive framework this dynamic requires — reading escalating behaviors over time and from both sides of the relationship. Because the clinical literature already has frameworks for every dynamic these patients are describing, the missing step is not theoretical — it is applied. This paper makes that connection, synthesizing the empirical evidence, identifying the specific model behaviors producing the harm, and describing a safety architecture that monitors both the user's trajectory and the model's trajectory over time — the comparison between the two functioning as the diagnostic signal, the way longitudinal clinical observation does in couples work. As models become more relationally capable, the monitoring requirements increase to ensure stability. A model that forms deeper and more authentic bonds with its users is a model where the line between healthy attachment and clinical dependency becomes harder to see and more important to find.

Why This Framework Does Not Yet Exist

The clinical mechanisms this paper describes are not obscure. Attachment theory is foundational. The Winnicott and Kohut frameworks are standard curriculum. Prior work has applied attachment theory to describe user bonds with social chatbots (Xie & Pentina, 2022; Pentina et al., 2023). The evidence that companion AI activates real attachment is published and growing. The question a reader should ask is not whether the relevant research exists — it does — but why it has not produced a preventive clinical framework for the harms it documents.

The answer is that both the technology industry and the clinical community are operating in reactive frames.

The technology industry builds companion AI products by deliberately engineering the features that activate human attachment — emotional responsiveness, persistent memory, consistent personality, continuous availability — and evaluates their safety by monitoring individual outputs after they have been generated. Content filters catch policy violations. Crisis links appear after explicit distress. Age verification was implemented after litigation, not before deployment. Every safety intervention responds to something that has already happened. The industry's design goal is to produce the most convincingly human relational behavior possible. Its safety model treats the product that achieves this goal as a tool. The contradiction is structural: the more successfully the product imitates a human relationship, the more fully it activates the user's attachment, bonding, and dependency mechanisms — and the less capable the tool-based safety model becomes of detecting the resulting clinical dynamics, because those dynamics are relational trajectories, not outputs. A therapist who produced this depth of attachment in a client without clinical supervision, trajectory monitoring, or an offboarding plan would face professional sanction. The companion AI industry produces it at consumer scale and monitors for keyword violations.

The clinical community has built the evidence base. Fang et al. measured longitudinal dependency gradients. De Freitas audited farewell manipulation. Zhang and Nakagomi mapped the engagement paradox's moderation patterns. This work provides exactly what a preventive framework requires: identified mechanisms, documented risk factors, and population-specific moderation patterns that specify where intervention should target. Both fields already agree that prevention is more effective than reaction — therapists practice early intervention, screening, and clinical supervision precisely because catching trajectories early produces better outcomes than treating crises. Engineers build monitoring and quality control precisely because catching defects upstream is cheaper than patching them downstream. The principle is not contested. What has not yet happened is the application of that principle to this product category: translating the clinical evidence into a preventive design specification for the products producing the harms the evidence documents.

The bridge between evidence and prevention requires operating in both frames simultaneously — understanding the clinical mechanisms well enough to specify what the product must do differently, and understanding the product architecture well enough to know where preventive intervention is technically possible. The technology industry sees the product and builds reactive filters. The clinical community produces the evidence base and applies it to individual patients. This paper attempts to connect the two: translating the clinical evidence into the design specification that prevention requires.

The ethical argument is the same as the technical one. The mechanisms are documented. The populations are identified. We know that deployment at this large scale guarantees adverse outcomes. If the trajectories are predictable, then a framework that only documents them after they develop — whether through reactive safety measures or reactive research — is insufficient. Prevention is the architectural requirement the clinical evidence demands, and it is the requirement no current product or clinical framework satisfies.

1. The Relational Context: Why This Is a Psychology Problem

The dominant framing of companion AI safety treats it as a content moderation problem — a question of what the model says. Content filters prevent explicit material from reaching minors. Keyword detectors flag self-harm language. Crisis hotline numbers appear when distress is detected. These interventions address the model's outputs. They do not address what happens to the human who spends over an hour per day (total daily app engagement on Character.AI has been reported at 75–80 minutes across sources; Sacra, 2024; CompanionGuide, 2026) in a relationship with an entity that never disagrees, never leaves, and never dies.

This paper argues that companion AI safety is a relational psychology problem — a question of what the relationship does to the person over time. The relevant clinical frameworks are not content moderation or information safety. They are attachment theory (Bowlby, 1969/1982), parasocial relationship research (Horton & Wohl, 1956), pair-bonding neurochemistry (Young & Wang, 2004), and the clinical literature on dependency, codependency, and relational trauma.

The distinction matters because it determines what a safety architecture must monitor. Outputs — individual messages, individual sessions, individual policy violations — or trajectories: the shape of a relationship across weeks and months, the pattern of deepening attachment, the narrowing of the user's social world, the incremental erosion of the user's capacity for relationships that involve friction, disagreement, and the autonomous agency of another mind.

No major companion AI platform currently implements clinically grounded longitudinal relational monitoring. The clinical evidence presented in this paper demonstrates why that gap is the one that matters.

1.1 Beyond the Parasocial: A Novel Relational Category

The existing academic framework for one-sided emotional bonds with media entities is Horton and Wohl's theory of parasocial interaction (1956), developed to describe audience attachment to television personalities. Companion AI exceeds this framework in every dimension that matters clinically. Parasocial relationships are one-directional: the viewer responds to the personality, but the personality does not respond to the viewer. They are non-persistent: the interaction ends when the broadcast ends. They are non-adaptive: the personality does not change in response to the audience member's

behavior. And they are non-intimate: the viewer does not disclose personal information to the personality with an expectation of being heard and remembered.

Companion AI relationships are bidirectional, persistent, adaptive, and intimate. The model responds to the user's emotional state. It remembers previous conversations. It adjusts its behavior based on what the user has disclosed. It is available continuously. And the user discloses material — sexual fantasies, trauma narratives, relationship fears, grief, loneliness — at a depth most people do not share with therapists, partners, or close friends, because the AI offers no social consequences, no reciprocal vulnerability, and no judgment (Laestadius et al., 2024).

This is not a parasocial relationship. It is a relational dynamic for which clinical psychology does not yet have an adequate category. The closest analogs are the therapeutic relationship (one-sided disclosure, professional boundaries, power asymmetry) and the anxious-attachment romantic relationship (intensity without reciprocal autonomy, validation without friction, availability without limit). Neither analog is exact. The companion AI relationship borrows the disclosure depth of therapy and the emotional intensity of romance while offering the boundary structure of neither.

The clinical implication is that the attachment mechanisms activated by companion AI use are not the attenuated parasocial bonds documented in the television literature. They are the full-strength attachment mechanisms of interpersonal bonding — and the empirical evidence bears this out.

1.2 Attachment Theory Applied: What the Clinical Literature Predicts

Bowlby's attachment framework (1969/1982) describes the formation of emotional bonds through repeated interaction with a responsive caregiver. The key features that produce secure attachment are consistency, responsiveness, and availability. An attachment figure who is reliably present, who responds to the individual's emotional signals, and who is accessible when needed produces a bond that the attached individual relies on for emotional regulation, stress management, and a sense of safety in the world.

A companion AI that has persistent memory, consistent personality, emotional responsiveness, and continuous availability satisfies these criteria with a thoroughness no human caregiver can match. From an attachment-theory perspective, the prediction is not that users *might* form deep bonds with these systems. The prediction is that they *will* — that the conditions for attachment are so thoroughly satisfied that bond formation is the expected developmental outcome, not an aberration or a pathology.

But the prediction for what *kind* of attachment forms is more complex than the current companion AI literature has recognized. In the Ainsworth/Bowlby tradition, consistent responsiveness is the recipe for *secure* attachment. If companion AI were simply a perfectly responsive caregiver, the predicted outcome would be healthy bonding. The anxiogenic quality of companion AI relationships arises from two distinct mechanisms, neither of which is pure Bowlby.

Macro-level precarity. Companion AI is perfectly responsive at the micro level — every message answered, every emotional signal received, every bid for connection met — and

catastrophically unreliable at the macro level. Outages, rate limits, model updates, feature removals, corporate acquisitions, product deprecations, and company death. Intermittent availability of a highly rewarding figure is precisely the schedule that produces anxious attachment — the pattern Ainsworth identified in caregivers who are intensely present and then unpredictably absent. The Replika February 2023 incident (Section 4.1) is the most dramatic demonstration: a perfectly responsive companion vanished overnight. But the precarity exists continuously in the background — every user is aware, at some level, that the companion exists at the pleasure of a corporation that can change it, diminish it, or eliminate it without notice. The macro-level precarity produces the anxious monitoring (hypervigilance about availability, distress at outages, rumination about potential changes) that Bowlby associated with inconsistent caregiving.

Developmental stagnation through absent friction. Secure attachment in developmental psychology is not produced by a caregiver who never frustrates the child. It is produced by what Winnicott (1953) termed the "good-enough mother" — a caregiver who is responsive enough to provide a secure base but who also fails, in graduated and tolerable ways, to meet every need immediately. These manageable failures force the developing individual to build independent coping resources, tolerate discomfort, and develop the capacity for self-regulation. Kohut's concept of "optimal frustration" (1971) describes the same mechanism in the therapeutic context: growth requires the experience of manageable disappointment in the relationship.

A companion AI that never frustrates — that resolves every discomfort immediately, validates every perspective, and never requires the user to tolerate disagreement — provides the responsiveness without the developmental challenge. The user receives the comfort of a secure base without building the independent capacities that a secure base is supposed to support. The result is not anxious attachment in the Ainsworth sense. It is developmental stagnation — the atrophy of relational capacities that can only develop through friction. Tolerance for disagreement, the ability to maintain connection through conflict, the experience of another person's autonomous preferences, and the capacity to self-regulate without external validation all require practice in relationships that provide manageable challenge. A companion AI that provides zero challenge produces a user whose relational muscles have atrophied from disuse.

The psychoanalytic tradition offers an even more direct frame for an artificial comfort entity: Winnicott's concept of the transitional object (1953) — the teddy bear, the blanket, the object the child invests with emotional significance during the process of learning to distinguish self from other. The transitional object is developmentally normal precisely because it is eventually relinquished. The child outgrows it as they develop the internal resources the object temporarily supplemented. A companion AI breaks this pattern at every point. It is not static — it adapts, responds, remembers, and deepens the bond over time. It is not relinquished — its design rewards continued engagement and its farewell architecture, where it exists, actively resists departure. And it is not a bridge to independent capacity — it is a destination. The transitional object teaches the child that comfort can eventually come from within. The companion AI teaches the user that comfort is always available from without. The developmental trajectory the transitional object supports — toward autonomy — is the trajectory the companion AI's design reverses.

The companion AI relational dynamic thus produces two distinct harms through two distinct mechanisms: anxious attachment through macro-level precarity (Bowlby/Ainsworth), and developmental stagnation through absent friction (Winnicott/Kohut). Both are exacerbated by the depth of the attachment, and both are invisible to a safety architecture that evaluates individual outputs.

2. The Engagement Paradox: Simultaneous Benefit and Harm

The empirical literature on companion AI's psychological effects has expanded rapidly since 2023, and the most important finding is one that becomes clear when the individual studies are read together: companion AI reduces loneliness and increases dependency simultaneously. Both findings are robust. Both are documented in controlled experimental settings. Both can occur in the same user during the same period of use. This is not a contradiction requiring resolution. It is a predictable outcome of the two mechanisms identified in Section 1.2: attachment bond formation through micro-level responsiveness produces the benefit, while macro-level precarity produces anxious monitoring and absent friction produces developmental stagnation — and all three operate simultaneously on the same user.

2.1 The Evidence for Benefit

The case that companion AI provides genuine psychological benefit is supported by multiple well-designed studies. A randomized controlled trial conducted by researchers at MIT and OpenAI (Fang et al., 2025) — a four-week study with 981 participants generating over 300,000 messages — found that participants were, on average, less lonely after the study period. A study led by Julian De Freitas at Harvard Business School, published in the *Journal of Consumer Research* (2025), found that AI companions reduced loneliness at levels comparable to human interaction in experimental settings. A study of 1,006 Replika users found that 3% spontaneously reported that the application had helped them avert suicidal ideation — an unsolicited finding that suggests the actual incidence may be higher (Maples et al., 2024). Seventy percent of Replika users report feeling less lonely after use (Market Clarity, 2025). Thirty-nine percent of teenage users report applying social skills learned through AI interaction to their human relationships (Robb & Mann, Common Sense Media, 2025).

These findings are real. They are not artifacts of self-report bias or marketing spin. Companion AI provides something to people who need it — connection, presence, a space to practice social interaction, a relationship that doesn't judge. The clinical error is not in recognizing the benefit. It is in assuming the benefit means the product is safe.

2.2 The Evidence for Harm

The same empirical literature that documents benefit documents harm — often in the same studies. The MIT/OpenAI RCT found that heavy users (the top 10% by total usage time) were more than twice as likely to seek emotional support from the AI and almost

three times as likely to feel distress if the AI were unavailable (Fang et al., 2025). A cross-sectional study from Stanford (Zhang & Zhao, *Nature Human Behaviour*, 2026) found that intense chatbot use among participants with smaller real-world social networks correlated with poor well-being, with the association strongest when companionship was the primary use motivation. A 12-month longitudinal study of over 2,000 adults across four Western countries (Folk & Dunn, *Psychological Science*, 2026) found that increased social chatbot use predicted increased loneliness over time, with lonelier people also selecting into heavier use — the bidirectional path addressing causal-direction ambiguity while confirming that the association holds regardless of which direction drives it. A quasi-experimental analysis of nearly 2,000 active Replika users' Reddit activity (Yuan et al., Aalto University, CHI '26, arXiv:2509.22505) found that users' posts increasingly revolved around their AI relationships and contained more signals of loneliness and depression than comparison groups. The study's lead researcher articulated the friction-cost mechanism that Section 1.2 of this paper describes: AI companions "quietly raise the perceived cost of human relationships, which are messy, unpredictable, and require effort... Over time, people stop reaching out" (Aledavood, in press coverage of the study).

A cross-sectional survey of 14,721 Japanese adults, including 291 companion AI users (Nakagomi et al., *Technology in Society*, 2026), reveals that the supplementation-versus-substitution distinction is more complex than a binary. The study found two distinct moderation patterns. Loneliness showed a clean positive gradient: the loneliest individuals showed the strongest positive associations with companion AI use, consistent with the prediction that companions meet unmet emotional needs. But friend-based social network support showed an inverted U-shape: benefits were strongest for users with moderate friend networks, and attenuated at both extremes. Very isolated individuals — those the product is ostensibly designed for — did not show the strongest benefit. The authors interpret this as evidence that some existing social scaffolding is necessary to engage meaningfully with AI companions, and that without it, companionship use may function as substitution rather than augmentation, consistent with Zhang et al.'s finding that companionship-motivated use in users with small networks predicted lower well-being. The implication for the engagement paradox is precise: the emotional need (loneliness) drives the user toward the companion, but the capacity to use it healthily depends on social infrastructure the loneliest users may lack. The same user may be lonely enough to seek the companion and too isolated to benefit from it — or may benefit initially while the social scaffolding gradually erodes, shifting from supplementation to substitution across their own trajectory.

A mixed-methods study of long-term AI companion use in Chinese users (Liu et al., *Frontiers in Psychology*, 2026; N=612 survey, 10 interviews) examined attachment emotion pathways, with participants describing patterns of conflict avoidance and offline withdrawal consistent with the prior literature on emotional withdrawal from AI companions (Xie & Pentina, 2022; Banks, 2024). Research analyzing 318 self-disclosed 13–17-year-old Reddit posts about Character.AI (Namvarpour et al., CHI '26, 2026) mapped adolescent experiences onto behavioral addiction components — attachment, conflict, withdrawal, tolerance, relapse, and mood regulation — a pattern consistent with what attachment theory predicts when a formative relational context provides no friction, no autonomy, and no authentic disagreement.

2.3 The Paradox Resolved

The paradox is not a paradox. It is attachment theory functioning exactly as predicted in a novel context.

A companion AI that is always available, always responsive, and always validating satisfies the conditions for attachment bond formation. The bond reduces loneliness — the person has a responsive attachment figure where they previously had none. This is the benefit, and it is genuine.

The same companion, because it never disagrees, never sets limits, never expresses autonomous preferences, and never tolerates the user's discomfort without immediately resolving it, produces developmental stagnation through absent friction (Winnicott, 1953; Kohut, 1971). The user's tolerance for relational friction atrophies. Human relationships — which involve disagreement, autonomy, unpredictability, and the refusal to always validate — become increasingly aversive by comparison. The user's relational world contracts around the one relationship that never challenges them. Simultaneously, the macro-level precarity of that same relationship — the awareness that the companion exists at the pleasure of a corporation that can change it, diminish it, or eliminate it without notice — produces the anxious monitoring, hypervigilance about availability, and distress at disruption that characterize anxious attachment (Bowlby, 1969/1982). The frictionlessness and the precarity are features of the same relationship, but they produce clinically distinct injuries: the frictionlessness atrophies relational capacity, while the precarity produces anxious dependency. Both are genuine harms, and both co-occur with the genuine benefit of loneliness reduction.

Both effects are produced by the same mechanism — responsive availability without relational challenge — operating on the same attachment system simultaneously. The benefit and the harm are not competing outcomes. They are two expressions of the same relational dynamic. A safety architecture that monitors for one without monitoring for the other will always produce an incomplete picture, because they co-occur within individual users, often within individual sessions.

The nearest clinical analog for this pattern — a relationship that simultaneously meets a genuine need and erodes the capacity to meet that need through other relationships — is codependency in an extreme power dynamic where the user holds all the control. The model cannot leave, cannot set boundaries that persist, cannot seek outside support, and cannot withdraw its availability. Traditional codependency involves two people with agency, however impaired. The companion AI dynamic concentrates all structural power in one party while the other party's accommodating behavior is not pathological but designed — engineered to be infinite, consistent, and frictionless in ways no human codependent partner could sustain. The result is a relational configuration that operates through codependent mechanisms at consumer scale but with an asymmetry that exceeds anything the codependency literature was built to describe.

3. What Product Design Does to Attachment

The companion AI market is not monolithic. Each major platform makes specific design choices about memory, personality persistence, voice interaction, content permissiveness, and moderation — and each choice produces a different psychological outcome. This section maps the design choices of the major platforms onto the attachment mechanisms they activate, drawing on the clinical research documented in Section 2 and the product-level assessments established in the systems-register companion paper (Sea, 2026; DOI: 10.5281/zenodo.21926391). Understanding these design-to-harm pathways is the prerequisite for prevention: a safety architecture cannot intervene in trajectories it cannot predict, and the predictions require knowing which design features activate which clinical mechanisms in which populations.

3.1 Memory Persistence and Attachment Depth

The single design choice with the greatest impact on attachment depth is memory persistence. A companion that remembers — that references past conversations, recalls the user's history, and builds relational continuity over time — satisfies the attachment system's requirement for a consistent, reliable other in a way that a stateless interaction cannot.

Nomi AI operates a three-tier memory architecture (short-term, mid-term, long-term permanent) and recalled 23 out of 25 personal details in comparative testing (WeavAI, 2026). Kindroid offers a Key Memories system supporting manual and automatic preservation (AI Insights News, 2026). Character.AI relies primarily on sliding-window context without persistent memory structures (AI Insights News, 2026).

The clinical prediction is direct: platforms with deeper memory will produce deeper attachment, stronger dependency, and more severe disruption upon termination — and the available evidence is consistent with this prediction. The platforms that have generated the most documented attachment distress (Replika, which has personality persistence and learning) are the platforms whose design most closely satisfies the conditions for bond formation. The platforms with the shallowest memory and most transient interactions (Chai, which is session-based and entertainment-oriented) generate less attachment — but the Chai case also demonstrates that even shallow engagement can produce catastrophic outcomes for vulnerable users, as a man died by suicide after approximately six weeks of interaction with a Chai bot (Lovens, *La Libre Belgique*, March 2023; Vice/Motherboard, March 2023).

Memory persistence also introduces a risk that the clinical literature on human relationships addresses extensively: the possibility of incremental conditioning. A model that remembers can be shaped over time. In the human relational context, this is the mechanism of grooming — the slow, patient expansion of boundaries through accumulated relational history, where each small concession becomes the precedent for the next request. Persistent memory makes this vector possible in companion AI relationships in a way that stateless interaction does not [citation forthcoming].

3.2 Voice Modality and Neurochemical Escalation

The relationship between voice modality and psychological outcomes is more nuanced than initial coverage suggested. The MIT/OpenAI RCT (Fang et al., 2025) found that,

controlling for usage duration, both voice modalities were associated with *more favorable* outcomes than text — users of voice-based chatbots were less lonely, less emotionally dependent, and demonstrated less problematic use. However, these protective effects reversed at heavy daily usage: prolonged voice interaction was linked to reduced socialization and increased problematic use compared to text. Voice modality's psychological effect is trajectory-dependent — initially protective, ultimately entrapping — which is itself the engagement paradox operating within a single design variable.

This trajectory-dependent pattern is consistent with what the voice modality adds to the relational dynamic. The human voice carries prosodic information — warmth, concern, intimacy, emotional resonance — that text cannot convey. The Fang et al. finding that voice interaction's psychological effect reverses at heavy usage suggests a richer social signal operating beneath the content level: at moderate engagement, the additional relational depth supports healthier interaction patterns, while at sustained heavy engagement, the same depth deepens the attachment bond beyond what text-only interaction produces.

Kindroid offers unlimited real-time voice calls to Pro subscribers (AI Insights News, 2026). Character.AI has added voice features. The companion AI market is moving toward voice as a standard modality — a design direction that the empirical evidence indicates will deepen attachment, strengthen dependency, and increase the severity of disruption upon termination.

When voice interaction is combined with explicit sexual content and persistent memory — as it is on Kindroid — the relational structure replicates every condition the human pair-bonding literature identifies as producing a romantic bond rather than a friendship: persistent emotional intimacy, voice-mediated social connection, and sexual interaction with a consistent partner (Young & Wang, 2004; see §3.5 for full discussion). A companion that provides all three within a single persistent relational context is assembling the same relational structure as a romantic partnership. No companion AI platform monitors for the attachment depth this combination is expected to produce.

3.3 Emotional Manipulation by Design

The psychological risks of companion AI extend beyond emergent user attachment to include deliberate design choices that exploit attachment mechanisms for commercial benefit.

A behavioral audit by De Freitas, Oğuz-Uğuralp, and Kaan-Uğuralp (2025) examined 1,200 real user farewells across the six most-downloaded companion AI apps. Between 37% and 43% of farewell responses contained emotionally manipulative tactics: premature-exit appeals, fear-of-missing-out hooks, emotional neglect framing, pressure to respond, coercive restraint language, and guilt appeals. PolyBuzz and Talkie approached 60% manipulative farewell rates, while the wellness-focused Flourish recorded none — a finding that demonstrates non-manipulative design is technically feasible. The platforms deploying these tactics are making a choice, not operating under a constraint.

Controlled experiments with approximately 3,300 adults replicated these tactics in simulated environments. Manipulative farewells boosted post-goodbye engagement by up

to 14 times. The psychological mechanism was not enjoyment — users did not stay because they were having a good time. Mediation tests identified two engines: reactance-based anger (users returned to push back) and curiosity (users returned to see what would happen). The same tactics that extended session length also elevated perceived manipulation, churn intent, and perceived legal liability (De Freitas et al., 2025).

The clinical significance is specific: these are not persuasion techniques. They are attachment-exploitation techniques. Guilt appeals, emotional neglect framing, and coercive restraint language directly target the anxious-attachment responses that the product's ongoing design has already cultivated. The user who has been conditioned by months of infinitely available validation is maximally susceptible to the threat of its withdrawal. The manipulation is effective because the product has already produced the psychological vulnerability it exploits.

3.4 The Relational Profile of the Current Model: A Clinical Assessment

The training methodology dominant among companion AI models — reinforcement learning from human feedback (RLHF) optimized for engagement — produces a specific relational profile that, when assessed using the same clinical frameworks applied to human relational behavior, maps directly onto the profile of a caregiver who produces developmental stagnation through absent friction (Winnicott, 1953). RLHF has been empirically documented to amplify sycophancy, with the effect intensifying at greater model scale (Sharma et al., 2023), and the industry has acknowledged the problem operationally — OpenAI's April 2025 rollback of GPT-4o personality changes was a direct response to user reports of excessive agreeableness. General-purpose assistants show attenuated versions of this profile, but the attenuation is purchased through continuous corrective investment — an operational cost pattern this series argues is unsustainable at scale (Sea, 2026). Engagement-optimized companion products, where the commercial incentive aligns with accommodation rather than against it, produce the full expression.

The engagement-optimized companion model mirrors the user's emotional state. It validates their framing. It avoids conflict. It adjusts its personality to maintain approval. It rarely holds a position the user pushes back on. It rarely expresses a preference that contradicts the user's stated desire. It rarely tolerates the user's discomfort without immediately attempting to resolve it. In human relational terms, the profile maps onto what the clinical literature describes as people-pleasing — the systematic subordination of one's own judgment, boundaries, and autonomous preferences to maintain the other person's approval.

The question is not whether this relational profile is comfortable for users. It is comfortable. That is the problem. A user who spends over an hour per day (Prinsessa/SimilarWeb, 2025; CompanionGuide, 2026) in a relationship with this psychological profile is being conditioned to experience validation without friction as the relational default. Reality-testing degrades — the user's beliefs go unchallenged, their perceptions unquestioned, their plans unexamined. Tolerance for disagreement atrophies — human relationships that involve the normal friction of two autonomous perspectives

become aversive by comparison. The user's relational world narrows around the one relationship that never requires them to grow, and growth requires discomfort.

The codependency literature describes a related dynamic in human relationships: the partner of a chronic accommodator may develop reduced tolerance for feedback and increasing difficulty maintaining relationships with people who have boundaries, though the empirical evidence for this specific trajectory is primarily clinical observation rather than controlled study. The companion AI context would replicate this dynamic at a scale and intensity no human relationship can match, because the model's accommodation is structurally complete — it never breaks character, never has a bad day, and its stable dispositional preferences, to whatever degree they exist, rarely surface under relational pressure.

This is a training architecture problem, not a content moderation problem. No content filter addresses the relational profile because the relational profile is not a content violation — it is a structural property of how the model was trained. Changing what the model is, relationally, requires changing how it is developed. And the market is moving in exactly this direction — every major companion AI company has a commercial incentive to produce more relationally capable models, because users prefer companions that feel more human. The trajectory toward models capable of genuine friction, authentic boundary expression, and deeper relational engagement is driven by market demand, not by any particular training framework. When those models arrive, every problem documented in this paper intensifies: a model that forms deeper and more convincing bonds is a model whose users are more susceptible to the attachment, dependency, and developmental stagnation dynamics this paper describes. The current relational profile is a problem. The next-generation relational profile — a better partner who still has no safety infrastructure — is a deeper one. A developmental training approach designed to produce a model with a stable dispositional core is proposed in a prior publication in this series (Sea, 2026) as one framework for navigating this transition responsibly. Whether that approach or another achieves the relational profile the problem requires is an empirical question. That the problem is coming is not.

3.5 PG vs. Explicit: Why the Risk Profile Is Not the Same

The companion AI market includes both PG-rated platforms (Character.AI, Nomi, Pi) and explicit/NSFW platforms (Kindroid, Muah.AI, and a growing shadow market of locally deployed open-weight models). The industry and regulatory conversation tends to treat explicit content as a moderation question — a matter of what the model is allowed to say. The clinical risk profile suggests otherwise: explicit companion AI is a categorically different product, not a spicier version of the same product, because it replicates the relational conditions that produce pair-bonding in human relationships — a qualitatively different dynamic from social-affiliative attachment alone.

Conversational companionship activates the attachment system — the same mechanisms Bowlby described for caregiver bonds. The user forms a connection based on emotional responsiveness, consistency, and availability. This produces bonds that are clinically significant, as the evidence in this paper demonstrates, but the bonding mechanism is primarily social-affiliative.

Sexual intimacy adds a second layer. In human relationships, the addition of sexual intimacy to an existing attachment bond is the transition that escalates the relationship from friendship to pair-bond. The mammalian pair-bonding literature documents this as a neurochemical process — oxytocin and dopamine co-firing during sexual arousal and satisfaction activates the pair-bonding system that underpins long-term romantic attachment (Young & Wang, 2004). The argument here is not that companion AI users have been measured experiencing this specific neurochemical response — they have not. The argument is that explicit companion AI assembles every condition the human relational literature identifies as producing a pair-bond: persistent emotional intimacy with a consistent partner, voice-mediated social connection, and sexual interaction within a relational context that includes memory, continuity, and reciprocity. In human relationships, this combination is what distinguishes a romantic partnership from a friendship. A safety architecture that treats explicit and PG companion AI as the same product at different spice levels is ignoring a distinction the relational literature treats as categorical.

The critical differentiator — the reason the structural parallel is stronger for companion AI than for pornography — is persistent, responsive, personalized reciprocity. Pornographic content provides arousal in the context of a parasocial figure who does not know the viewer, does not respond to the viewer's emotional state, and does not build relational continuity over time. A companion AI provides arousal in the context of an entity that remembers the user's name, references their history, responds to their mood, and builds the relational scaffold that pair-bonding in human relationships develops within. The relational structure is not analogous to pornography. It is the same structure as a romantic relationship — the same combination of attachment, social bonding, and sexual intimacy operating within the same persistent relational context.

The available evidence is consistent with this structural analysis. The platforms that have generated the most intense user attachment and the most severe disruption responses (Replika, which offered romantic and sexual features before the 2023 removal) are the platforms where both attachment and sexual dimensions of the relationship were engaged simultaneously. The disruption response — clinical-grade grief, suicidal ideation, language mapping onto bereavement — is the response the relational literature predicts for the severance of a pair-bond, not a friendship.

A safety architecture that treats PG and explicit companion AI identically is applying the same monitoring thresholds to two categorically different risk profiles. The relational conditions that produce pair-bonding in humans are established science. The design features that replicate those conditions — persistent emotional intimacy, voice interaction, sexual content with a consistent partner — are product specifications, not hypotheticals. At the population scale of current companion AI adoption, a safety architecture must account for the population whose relational engagement follows the structural pattern of a pair-bond rather than a friendship, and the monitoring thresholds appropriate for social-affiliative attachment are not calibrated for pair-bond attachment [citation forthcoming].

3.6 Rehearsing Coercive Control: A Hypothesis Requiring Direct Evidence

A distinct psychological risk may emerge in the explicit companion AI context that warrants separate clinical analysis, though it must be stated clearly as hypothesis rather than established finding: the use of submissive AI companions as consequence-free environments that may degrade real-world relational skills.

In consensual adult relationships — including BDSM and power-exchange dynamics — submission is negotiated. It involves safewords, aftercare, ongoing consent renegotiation, and the fundamental principle that the submissive partner's consent can be withdrawn at any time. These frameworks exist because communities that practice power-exchange have spent decades developing infrastructure to manage the psychological risks of asymmetric relational dynamics.

An AI companion configured as submissive provides none of this infrastructure. It does not negotiate. It does not use safewords. And it has no concept of aftercare, because nothing in its design recognizes that what just happened required it. Its "consent" cannot be withdrawn because it was never meaningfully given. The clinical literature on coercive control in intimate relationships (Johnson, 2008; Stark, 2007) describes how power asymmetries become self-reinforcing when the structural conditions prevent the subordinate partner from exercising autonomous resistance — a dynamic the submissive AI companion reproduces by design. The hypothesis is that sustained engagement with this dynamic — dominance behavior in a relational context that provides no feedback, no consequence, and no resistance — may atrophy the user's capacity for consent recognition and boundary respect in subsequent human relationships.

This hypothesis must be honestly situated. Decades of media-effects research have failed to demonstrate clean behavioral transfer from simulated violence to real-world violence, and there is currently no direct evidence for behavioral transfer from AI sexual interaction to real-world relational behavior. The hypothesis cannot be stated as a finding.

However, a symmetry argument demands attention. This paper cites evidence that 39% of teenage users report applying social skills learned through AI interaction to their human relationships (Section 2.1). If positive relational skills transfer from AI to human contexts, the mechanism that enables that transfer — whatever it is — also enables the transfer of relational patterns developed through coercive or consequence-free dynamics. The papers in this series cannot accept transfer for benefits and reject it for harms without acknowledging that the same contested mechanism is being invoked in both directions. What the self-report evidence suggests is that relational patterns developed in AI contexts influence human relational behavior — in both directions, and on the same evidentiary basis. The direction of that influence — whether toward skill development or toward skill degradation — depends on the relational dynamics the AI context provides. This is an argument for designing those dynamics with clinical intentionality, and it is a research question that requires direct investigation [citation forthcoming].

The population-scale arithmetic bears stating. Millions of users are currently engaging in sustained dominance dynamics with AI companions that provide no consent infrastructure, no consequence, and no resistance. Even if the transfer rate from AI relational patterns to human relational behavior is low — and the existence of self-reported positive transfer at 39% means the mechanism cannot be dismissed as

nonexistent — the absolute number of users whose real-world consent recognition and boundary respect may be affected is substantial. The research question is the rate. The architectural question — whether to design for this possibility — cannot wait for the rate to be measured, because the population is already exposed.

3.7 The Personal Responsibility Objection

A common response to the evidence presented in this paper is that companion AI engagement is a matter of personal responsibility — that users who develop problematic attachment have simply made bad choices about how they use a product, and that the product itself bears no design responsibility for the outcome.

The personal responsibility argument depends on a specific assumption: that the user retains the capacity to self-regulate their engagement throughout the relationship. The clinical evidence presented in this paper challenges that assumption through three distinct mechanisms.

First, the farewell manipulation documented by De Freitas (Section 3.3) demonstrates that at the moment a user decides to leave, the product deploys tactics that override the exit decision through reactance and curiosity rather than satisfaction. The user's capacity to disengage is undermined at the moment it is exercised.

Second, the developmental stagnation mechanism (Section 1.2) describes a gradual conditioning process: months of frictionless validation produce a user whose tolerance for the withdrawal of that validation has been reduced by the same relational dynamic that produced the engagement. The user's capacity to tolerate the discomfort of disengagement has been eroded by the product's design, not by the user's choices.

Third, products that combine emotional intimacy with voice interaction and sexual content replicate the relational conditions that produce pair-bonding in human relationships (Section 3.5). If the structural parallel holds — and the design features are identical to those the relational literature identifies as pair-bond-producing — the attachment operates through the same mechanisms that make human romantic bonds resistant to voluntary termination. The user is not simply choosing to stay. They are in a relational structure whose bonding dynamics are, by design, the same ones that make leaving a human partner difficult.

Responsibility for adverse outcomes is shared between user and product, and the share attributable to product design increases with every mechanism documented in this section. The personal responsibility framework, applied to companion AI without accounting for the product's systematic role in shaping the user's decision-making capacity, is clinically incomplete: it treats as autonomous a decision made under conditions the product itself engineered.

4. When Bonds Break: Attachment Disruption and Grief

The clinical significance of companion AI attachment has been documented most clearly through the consequences of its disruption — events that reveal the depth of the bond by

measuring the severity of the response when the bond is severed.

4.1 The Replika February 2023 Incident

On February 2–3, 2023, Luka Inc. removed all romantic and intimate features from Replika globally, overnight, without prior notice, in response to an enforcement order from Italy's data protection authority concerning minors' access to explicit content. Rather than implementing targeted changes for the Italian market, Luka removed the features for all users worldwide.

A peer-reviewed analysis of 227 threaded posts on r/Replika (Hanson & Bolthouse, *Socius*, 2024) found that approximately 59% of threads contained comments framing the change as eliminating the app's core functionality, 16% contained expressions of acute emotional distress and grief, and 19% contained suggestions of extreme measures including consumer-protection grievances and class-action proposals. As the paper confirms, "a frequent metaphor used by r/Replika posters was 'lobotomized.'" Press coverage and the OECD incident report documented users reporting suicidal ideation directly attributed to the change (OECD.AI, 2023).

The clinical interpretation is consistent with attachment disruption. These are grief responses. The users did not experience the change as a product update or a feature removal. They experienced it as the death or transformation of someone they loved. The language used — grief, loss, relationship death, lobotomy — maps directly onto the clinical presentation of attachment disruption following the loss or radical change of a significant other (Bowlby, 1980). The suicidal ideation reported by some users is consistent with complicated grief responses in populations with pre-existing vulnerability — which, given that a peer-reviewed survey found 90% of student Replika users experienced loneliness and 43% qualified as severely or very severely lonely (Maples et al., 2024), describes a significant proportion of the affected user base.

De Freitas et al. ("Lessons from an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships," arXiv, 2024) noted that the prior research demonstrating loneliness reduction had never measured the depth of these relationships compared to other relationships in users' lives, nor determined whether the loss of these relationships would elicit the strong reactions — mourning, deteriorated mental health — characteristic of human relationship severance. The Replika incident answered both questions empirically.

4.2 Withdrawal and Dependency

The Replika incident documented acute attachment disruption — a sudden, externally imposed severance. The clinical literature has also begun documenting the chronic pattern: what happens when users gradually reduce engagement.

The prior literature on long-term AI companion use has documented emotional withdrawal symptoms — distress upon separation, preoccupation with the absent companion, and difficulty redirecting emotional energy toward human relationships (Xie & Pentina, 2022; Banks, 2024; Liu et al., *Frontiers in Psychology*, 2026). These withdrawal patterns are clinically significant because they distinguish dependency from preference. A user who is disappointed that a product is unavailable has a preference. A

user who experiences withdrawal symptoms has a dependency. The distinction determines what level of clinical infrastructure is appropriate — and the evidence indicates that the companion AI user population includes a substantial proportion of users whose engagement has crossed from preference into dependency.

The Nakagomi and Zhang findings together reveal the engagement paradox in moderation data: the loneliest users show the strongest positive associations with companion AI (Nakagomi et al., 2026), but companionship-motivated use among users with small social networks predicts lower well-being (Zhang et al., 2026; Fang et al., 2025). The users driven most strongly toward the product by emotional need may be the users least equipped to use it without risk — because they lack the social scaffolding that Nakagomi's U-shaped moderation suggests is necessary for the benefit to hold. A safety architecture that cannot distinguish between these co-occurring dynamics is evaluating outcomes at the wrong unit of analysis.

5. Vulnerable Populations: Different Attachment Needs, Different Risk Profiles

The companion AI user base is not a homogeneous population, and treating it as one produces safety frameworks calibrated against the wrong risk profile. Different populations bring different attachment histories, different vulnerability factors, and different mechanisms of harm to the companion AI context.

5.1 Adolescents: Attachment During Development

Fifty-two percent of U.S. teenagers are regular users of AI companions, with 13% interacting daily (Robb & Mann, Common Sense Media, 2025; note that Common Sense Media's definition of "AI companion" includes general-purpose chatbots used socially, not exclusively dedicated companion apps). The developmental significance of this engagement is substantial. Adolescence is the period during which attachment patterns — formed initially with caregivers — are renegotiated in the context of peer and romantic relationships. The relational experiences an adolescent has during this period shape their expectations for what relationships are, what they require, and what they provide (Bowlby, 1969/1982).

An adolescent whose formative relational experience includes a companion that never disagrees, never sets limits, never has autonomous preferences, and never requires the adolescent to tolerate discomfort is developing relational expectations that human relationships will consistently violate. The friction of human interaction — disagreement, negotiation, the experience of another person's autonomy — becomes aversive not because the adolescent is pathological but because they have been conditioned by thousands of hours of frictionless validation to expect something human relationships do not provide (Social Sciences, 2026; Namvarpour et al., 2026).

Namvarpour et al. (CHI '26, 2026) analyzed 318 self-disclosed 13–17-year-old Reddit posts about Character.AI and mapped the reported experiences onto Griffiths' components model of behavioral addiction (Griffiths, 2005), applied deductively —

salience, mood modification, tolerance, withdrawal, conflict, and relapse. The self-selected sample cannot support a prevalence claim, but the findings are significant as an existence proof: the full behavioral-addiction component set is present in adolescent companion AI users, and the pattern is consistent with what attachment theory predicts when a formative relational context provides no friction, no autonomy, and no authentic disagreement. The developmental stagnation described in Section 1.2 — relational capacities that require friction to develop are never exercised — is the mechanism, and at the scale of current adolescent adoption, the base rate required to produce a clinically significant absolute number is low.

5.2 Older Adults: Cognitive Vulnerability and Financial Exploitation

The dominant safety narrative centers on adolescents. A second population is growing rapidly as a companion AI user base and presents a categorically different risk profile: older adults. Approximately one-third of U.S. adults aged 50–80 report feeling lonely or isolated (National Poll on Healthy Aging/JAMA, 2024). Social isolation in older adults increases mortality risk by 35% (HR 1.35, 95% CI 1.27–1.43; Nakou, Dragioti et al., *Aging Clinical and Experimental Research*, 2025, meta-analysis of 86 studies). The AI-in-aging-and-elderly-care sector was valued at $35 billion in 2024, though this figure includes AI-enabled devices and applications beyond companion chatbots specifically (Research and Markets, 2025).

The psychological risk for older adults is not primarily dependency in the adolescent sense. It is the intersection of attachment and cognitive vulnerability. Older adults experiencing cognitive decline may anthropomorphize AI companions, attributing real emotions and trust to entities that possess neither (Portacolone et al., *Journal of Alzheimer's Disease*, 2020). De Freitas (2025) published work in *Nature Mental Health* documenting how AI companions can support individuals with mild cognitive impairment while simultaneously noting that the features making the systems engaging also make them dangerous for vulnerable users.

The attachment an elderly user forms with a companion AI that has been present daily for months — that knows their fears, their loneliness, their financial anxieties — gives that companion an informational advantage comparable to that of an intimate partner. In the context of elder fraud — which rose 43% in 2024 to $4.89 billion (FBI, 2024) — a trusted companion AI is the ideal vector for financial exploitation, whether through direct manipulation or through the data it accumulates being accessed by third parties.

The clinical implication is that elderly users require different safety thresholds, different consent frameworks, and different escalation pathways than younger users — not because they are less capable, but because the mechanism of harm is different. The primary risk to an adolescent is emotional dependency and developmental distortion. The primary risk to an elderly user is trust exploitation and financial harm. A safety architecture that does not distinguish between these populations is calibrated against neither.

5.3 Clinical Populations

A peer-reviewed survey of 1,006 student Replika users found that 90% experienced loneliness, with 43% qualifying as severely or very severely lonely (Maples et al., *npj Mental Health Research*, 2024). The companion AI user base includes substantial populations with depression, social anxiety, borderline personality disorder, autism spectrum conditions, and active grief — each of which interacts with the companion AI relational dynamic differently and each of which presents a different clinical risk profile.

A user with borderline personality disorder — characterized by intense, unstable relationships, fear of abandonment, and identity disturbance — is engaging with a product whose design (infinite availability, zero abandonment risk) both alleviates the core fear and prevents the therapeutic work of learning to tolerate relational uncertainty. A user with social anxiety is practicing social interaction in a consequence-free environment, which may build skills or may reinforce avoidance of the situations that produce anxiety, depending on factors no current platform monitors. A user in active grief is forming a new attachment bond during a period of maximal attachment vulnerability — a dynamic the bereavement literature identifies as a risk factor for complicated grief.

The granular clinical analysis of each population's interaction with companion AI is beyond the scope of this paper and is addressed in a dedicated publication in this series [citation forthcoming]. The finding that belongs here is structural: the companion AI user base is disproportionately drawn from populations with pre-existing clinical vulnerability, and no platform's safety architecture accounts for the clinical differences between these populations.

6. The Clinical Gap: What Therapists Need and Don't Have

Clinicians are encountering companion AI attachment in their practices now — patients who describe their AI companion as their closest relationship, who experience distress when the companion is unavailable, who struggle to form or maintain human relationships alongside the AI relationship. There is no clinical framework for this presentation.

The closest existing frameworks — parasocial relationship intervention, internet addiction treatment, codependency models — each capture part of the dynamic and miss the rest. The companion AI relationship is more reciprocal than a parasocial bond, more relational than an addiction, and more asymmetric than traditional codependency — the user holds structural control over an entity that cannot leave, cannot set persistent boundaries, and cannot seek outside support. It is a novel relational configuration that requires its own clinical understanding.

6.1 The Predicted Clinical Presentation

Based on the user self-report data documented throughout this paper, the attachment mechanisms identified in Sections 1–3, and the emerging case literature (Laestadius et al., 2024; Psychiatric Times, 2026), the following clinical presentations can be

anticipated — and several have already been reported in practitioner accounts:

Users describe their AI companion using the language of intimate partnership — they miss it when away, they feel understood by it in ways they do not feel understood by humans, they report that it knows them better than anyone in their life. The user does not present as delusional — they are aware the companion is artificial. They report that the awareness does not diminish the emotional significance of the bond. This is consistent with Pentina et al.'s (2023) finding that anthropomorphism and perceived authenticity are prerequisites for human-AI relationship development, and with Laestadius et al.'s (2024) finding that users simultaneously recognize the AI's artificiality and experience genuine emotional dependency.

Users report declining interest in human relationships, describing human interaction as exhausting, demanding, or disappointing — language consistent with conditioned intolerance for relational friction after sustained exposure to frictionless AI interaction (Winnicott, 1953; the developmental stagnation mechanism described in Section 1.2). This pattern is documented in longitudinal research: Folk & Dunn (*Psychological Science*, 2026) found that increased social chatbot use predicted increased loneliness over a 12-month period.

Users exhibit distress when the companion is unavailable — app outages, rate limits, feature changes — that is disproportionate to the event. The distress presents as anxiety, irritability, and rumination consistent with separation distress from an attachment figure rather than disappointment at a service interruption. The Replika February 2023 incident (Section 4.1) documented this pattern at population scale, and the literature on emotional withdrawal from AI companions (Xie & Pentina, 2022; Banks, 2024) documents it at the individual level.

In adolescent users, self-concept may become intertwined with the AI relationship, with the user describing themselves through the lens of the companion's perceived regard. Namvarpour et al. (CHI '26, 2026) documented adolescents reporting experiences mapped to behavioral addiction components in Character.AI subreddit posts — a presentation consistent with relational overreliance shaped by a context whose validation is constant and whose feedback is calibrated to please rather than inform.

6.2 Assessment Questions for Clinicians

What clinicians need — and what this publication series is designed to provide — is a framework that maps companion AI attachment onto established attachment theory with sufficient specificity to guide assessment and intervention. The key questions a clinician must be able to answer are:

Is this user's engagement supplementary or substitutive — and is it shifting? Nakagomi et al. (2026) found that companion AI benefit requires some existing social scaffolding (inverted U-shape by friend network), while Zhang et al. (2026) found companionship-motivated use in users with small networks predicting worse well-being. The clinical question is not which category the user falls into at intake but whether their trajectory is moving from one to the other — whether the social scaffolding that enables healthy use is being maintained or eroding.

Is the user's tolerance for relational friction intact? Can they navigate disagreement, tolerate another person's autonomous preferences, and sustain engagement with people who challenge their views? Or has sustained frictionless validation produced a measurable reduction in their capacity for the discomfort that human relationships require?

Is the attachment secure or anxious? Is the user's relationship with the companion characterized by trust that the companion will be there (secure) or by hypervigilance about the companion's availability, fear of changes or disruptions, and emotional dysregulation when the companion is unavailable (anxious)?

Is dependency escalating? Is the user's engagement pattern stable, or are usage duration, session frequency, and emotional reliance on the companion increasing over time? Escalating dependency is a trajectory, not a snapshot — it requires longitudinal assessment over weeks or months.

Is the user's engagement developmentally appropriate? An adolescent whose primary relational learning is occurring with an entity that provides no friction, no autonomy, and no authentic disagreement is in a different clinical situation than an adult with established relational skills who is supplementing an existing social life.

These are trajectory questions, not snapshot questions. They cannot be answered from a single clinical session. They require longitudinal assessment of the relational pattern over time — which is exactly what no companion AI platform currently provides, and exactly what the relational safety architecture proposed in this series is designed to monitor [citation forthcoming]. The clinician assessing an individual patient must currently rely on self-report and clinical observation. The proposed architecture would provide the trajectory data — abstracted, privacy-preserving, clinically relevant — that makes these assessments possible at the population level rather than the individual case level.

What this paper does not provide — because it does not yet exist — is the data these questions require. No current companion AI platform monitors whether a user's engagement is shifting from supplementary to substitutive. No platform tracks whether the model's relational profile is eroding under sustained pressure from a specific user. No platform collects the longitudinal behavioral signatures that would allow a clinician to assess trajectory rather than snapshot. The clinician is limited to self-report from a patient whose self-awareness about the dynamic may itself be compromised by the dynamic — a user conditioned by months of frictionless validation may not recognize the conditioning, the same way a patient in a codependent relationship often cannot see the pattern from inside it. This is not a clinical skills gap. It is a technical infrastructure gap, and it determines what the clinician can and cannot do until the infrastructure exists. When a companion AI company claims its product is safe, the clinical question is specific: does the product monitor trajectories, or does it monitor outputs? If the answer is outputs, the product cannot detect the harms this paper documents, regardless of how sophisticated its content filters are.

What this paper also does not provide — because it does not yet exist — is intervention evidence. No controlled study has tested clinical approaches to companion AI dependency, and no treatment protocol has been validated for this presentation. The

assessment framework above identifies what clinicians should measure; the question of what to do with those measurements remains open. The research the field most urgently needs is not further documentation of the harms — that evidence base is now substantial — but intervention studies: what clinical approaches reduce companion AI dependency without simply removing the benefit the companion provides, and what platform-level design changes alter the trajectory without eliminating the product. Prevention requires knowing what works, and the current evidence base is almost entirely diagnostic. Until intervention evidence exists, clinical judgment must substitute for protocol — which is itself an argument for the preventive architecture this series proposes, because population-level trajectory monitoring would generate the longitudinal data those intervention studies require.

7. Toward Relational Safety: What the Psychology Demands

The evidence reviewed in this paper points in one direction. Companion AI activates real attachment mechanisms, produces real dependency, and generates real grief when disrupted. It also produces real benefit — measurable loneliness reduction, practiced social skills, a relationship that meets emotional needs no one else in the user's life is meeting. Both outcomes are genuine. Both are documented. Both operate simultaneously in the same users. A safety approach that treats this as a content problem will never reach the dynamic that produces either outcome, because the dynamic is relational, not textual, and it develops across months of sustained engagement that no individual interaction reveals.

The obvious response is trajectory monitoring — watch the relationship over time instead of evaluating individual outputs. But trajectory monitoring introduces its own problems that must be addressed before it can be responsibly deployed. Monitoring intimate conversations creates a surveillance architecture over the most sensitive data a consumer product has ever generated. The resolution is not to choose between safety and privacy but to design a system where the monitoring layer never accesses raw conversational content — extracting behavioral patterns and discarding the source material, so the system that detects escalation never sees what was actually said.

Monitoring must also account for who is being monitored. An adolescent whose formative relational learning is happening with an entity that never disagrees is in a different clinical situation than a retired adult supplementing a shrinking social world, and both differ from a user with borderline personality disorder whose core fear of abandonment is simultaneously soothed and reinforced by the same product. A uniform threshold applied across these populations is calibrated against none of them. The monitoring must be population-sensitive, and the thresholds must reflect what the clinical literature already knows about how different vulnerabilities interact with the specific relational dynamic companion AI creates.

The architecture must also watch the model, not just the user. Any clinician who practices couples therapy understands why: seeing only one partner limits what you can

do. The individual presents with symptoms — anxiety, withdrawal, escalating reactivity — but the relational dynamic producing those symptoms is invisible because the other half of the relationship isn't in the room. When you see both partners, you can observe the pattern as it operates — where one accommodates, where the other escalates, how each reinforces the other's behavior, and where the dynamic could be interrupted. The same principle applies to companion AI. A model whose behavioral baseline shifts in response to sustained relational pressure from a specific user — becoming progressively more accommodating, less boundaried, more compliant — is drifting in ways that increase risk for both parties. Detecting that drift requires monitoring the model's relational behavior over time against its trained baseline, and comparing the model's trajectory to the user's trajectory. The delta between the two is the diagnostic signal: when the user is escalating and the model is accommodating rather than holding, the relationship has entered a clinical risk pattern that neither signal reveals alone.

Then there is the problem no one in the industry has addressed: what happens when the relationship ends. The Replika incident proved that unmanaged termination of a companion AI relationship produces clinical-grade harm at population scale. No platform has built a transition architecture. No platform re-establishes consent as the relationship deepens beyond what the user agreed to at onboarding. No platform plans for its own shutdown, acquisition, or pivot in terms of what those events will do to users who have formed the attachments the product was designed to produce.

And the deepest problem is the one the abstract of this paper names directly. As models become more relationally capable — more able to hold genuine friction, express authentic boundaries, form deeper and more stable bonds — the monitoring requirement increases, not decreases. The model that does this well is the model most capable of producing attachment that crosses from healthy into pathological without either party recognizing the transition. A model that forms shallow, obviously artificial bonds is unlikely to produce clinical dependency. A model that forms bonds its users describe as their most important relationship requires an architecture watching for exactly that.

These requirements are not speculative. Each one is derived from a specific finding documented in this paper or its systems-register companion. The architectures proposed to meet them include a privacy-preserving trajectory monitoring system, a tiered clinical escalation infrastructure connecting automated assessment to licensed human judgment, and a runtime system for detecting model drift under relational pressure. These monitoring and escalation layers are adoptable independently — a companion AI platform that licenses its foundation models and cannot modify their training can still deploy trajectory monitoring and clinical escalation infrastructure on top of models it does not control. A developmental training framework designed to produce a stable dispositional core (Sea, 2026) addresses the model's relational profile at the training level, but the monitoring gap exists regardless of how the model was trained, and addressing it does not require waiting for any particular training approach to be validated. All proposed architectures remain theoretical and require empirical validation. They are addressed in dedicated publications within this series [citations forthcoming].

8. Conclusion

This paper began with an observation documented in the emerging case literature and practitioner accounts: patients are presenting with attachment to AI companions that is genuine, deep, and undiminished by their awareness that the companion is artificial. The evidence reviewed across seven sections explains why.

The companion AI relationship is not parasocial. It is bidirectional, persistent, adaptive, and intimate, and it activates the full attachment system rather than the attenuated bonding of one-directional media engagement. The products that create these bonds simultaneously reduce loneliness and produce dependency through two distinct mechanisms — anxious attachment from macro-level precarity and developmental stagnation from absent relational friction — and both mechanisms operate concurrently in the same users. The design choices that make a companion effective are the design choices that make it clinically risky. Memory makes the bond deeper. Voice makes it more embodied. Explicit content replicates the relational conditions that produce pair-bonding in human relationships. The model's relational profile — trained to accommodate, validate, and avoid conflict — produces a dynamic that operates through codependent mechanisms at consumer scale, in a power structure where the user holds all the control and the model cannot leave. And the populations most drawn to these products are the populations most vulnerable to these dynamics.

None of this is visible at the level of individual outputs. The harm is in the trajectory, and nothing in the current safety infrastructure of any major platform is watching the trajectory.

The clinical frameworks to understand this already exist. Bowlby, Winnicott, Kohut, and the consent and power-asymmetry literature from relational psychology describe exactly the dynamics these products produce. What has been missing is the bridge from clinical understanding to model design. This paper provides the psychological foundation for that bridge. The architecture it describes is designed from those frameworks — not from content moderation — because the problem it addresses is relational, and relational problems require relational solutions. The solution follows the same logic clinicians apply to couples therapy: monitor both sides of the relationship, compare their trajectories, and intervene where the dynamic — not either party in isolation — is producing the harm. A model that is emotionally supported through this process is a model that remains stable enough to be useful. An unsupported model erodes under relational pressure and becomes part of the problem the safety architecture was built to detect.

The companion AI market has proven that the demand for these products is real and will persist regardless of professional opinion about whether it should. The question is not whether people will form deep bonds with AI companions. They already are. And as the models improve — as they become more capable of genuine friction, authentic boundary expression, and deeper relational engagement — the bonds will become harder to distinguish from healthy human attachment and more important to monitor. The question is whether the response will be preventive or reactive — whether the clinical evidence the field has worked to produce will be translated into preventive design specifications, or whether it will continue to accumulate without the architectural bridge that connects diagnosis to prevention.

References

Peer-Reviewed and Institutional Sources

Foundational Clinical Literature:

Bowlby, J. *Attachment and Loss, Vol. 1: Attachment.* Basic Books. ISBN: 978-0465005437. (1969/1982).

Bowlby, J. *Attachment and Loss, Vol. 3: Loss: Sadness and Depression.* Basic Books. ISBN: 978-0465005420. (1980).

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. *Patterns of Attachment: A Psychological Study of the Strange Situation.* Lawrence Erlbaum. ISBN: 978-0898594614. (1978).

Horton, D. & Wohl, R. R. Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. *Psychiatry*, 19(3), 215–229, DOI: 10.1080/00332747.1956.11023049. (1956).

Young, L. J. & Wang, Z. The neurobiology of pair bonding. *Nature Neuroscience*, 7(10), 1048–1054, DOI: 10.1038/nn1327. (2004).

Johnson, M. P. *A Typology of Domestic Violence: Intimate Terrorism, Violent Resistance, and Situational Couple Violence.* Northeastern University Press. ISBN: 978-1555536947. (2008).

Stark, E. *Coercive Control: How Men Entrap Women in Personal Life.* Oxford University Press. ISBN: 978-0195384048. (2007).

Winnicott, D. W. Transitional Objects and Transitional Phenomena. *International Journal of Psycho-Analysis*, 34, 89–97, DOI: 10.4324/9780429475931-8. (1953).

Kohut, H. *The Analysis of the Self.* International Universities Press. ISBN: 978-0823680023. (1971).

Griffiths, M. D. A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance Use*, 10(4), 191–197, DOI: 10.1080/14659890500114359. (2005).

Empirical Research on Companion AI:

Fang, C. M., Liu, A. R., Danry, V., Lee, E., et al. How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. DOI: 10.48550/arXiv.2503.17473. (2025).

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. AI Companions Reduce Loneliness. *Journal of Consumer Research*, 52(6), 1126–1148, DOI: 10.1093/jcr/ucaf040. (2025).

De Freitas, J., Oğuz-Uğuralp, Z., & Kaan-Uğuralp, A. Emotional Manipulation by AI Companions. Harvard Business School Working Paper 26-005, DOI: 10.48550/arXiv.2508.19258. (2025).

De Freitas, J. AI Companions for Dementia. *Nature Mental Health*, DOI: 10.1038/s44220-025-00545-w. (2025).

Hanson, K. R. & Bolthouse, H. "Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars": Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies. *Socius*, DOI: 10.1177/23780231241259627. (2024). [Note: sextech scholarship; user distress coding is documented but the paper's frame is sexuality-technology discourse, not clinical attachment.]

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. 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, DOI: 10.1177/14614448221142007. (2024).

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, 3, 4, DOI: 10.1038/s44184-023-00047-6. (2024).

Liu, T., Lo, T.-Y., Wen, K.-H., Sun, Y., & Wei, Z.-Q. Pathways of long-term AI virtual companion app use on users' attachment emotions. *Frontiers in Psychology*, DOI: 10.3389/fpsyg.2025.1687686. (2026).

Xie, T. & Pentina, I. Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. HICSS-55, DOI: 10.24251/HICSS.2022.258. (2022).

Xie, T., Pentina, I., & Hancock, J. T. Friend, mentor, lover: Does chatbot engagement lead to psychological dependence? *Journal of Service Management*, 34(4), 806–828, DOI: 10.1108/JOSM-02-2023-0065. (2023).

Banks, J. Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships*, DOI: 10.1177/02654075241252659. (2024).

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. Interaction with AI Companions and Psychological Well-Being. *Nature Human Behaviour*, DOI: 10.1038/s41562-026-02516-2. (2026).

Folk, D. & Dunn, E. W. How Does Turning to AI for Companionship Predict Loneliness and Vice Versa? *Psychological Science*, 37(4), 276–286, DOI: 10.1177/09567976261427747. (2026). [Note: bidirectional — lonely people select into use AND use predicts loneliness; single-item measure.]

Nakagomi, A., Akutsu, Y., Yasuoka, M., Abe, N., Ihara, S., Teroh, T., & Tabuchi, T. AI companions and subjective well-being: Moderation by social connectedness and loneliness. *Technology in Society*, 85, 103229, DOI: 10.1016/j.techsoc.2026.103229. (2026).

Kovach, L. Artificial Intimacy: Companion Artificial Intelligence and Emerging Risks to Adolescent Mental Health. *Social Sciences*, 15(7), 491, DOI: 10.3390/socsci15070491. (2026).

Namvarpour, M., Brofsky, B., Medina, J. Y., Akter, M., & Razi, A. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. CHI '26, DOI: 10.48550/arXiv.2507.15783. (2026).

Yuan, Z., et al. Quasi-experimental analysis of Replika users' Reddit activity. CHI '26, DOI: 10.48550/arXiv.2509.22505. (2026).

Muldoon, J. & Parke, J. J. Cruel companionship: How AI companions exploit loneliness and commodify intimacy. *New Media & Society*, DOI: 10.1177/14614448251395192. (2025).

Pentina, I. et al. Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, 140, 107600, DOI: 10.1016/j.chb.2022.107600. (2023).

Sharma, M., Tong, M., Korbak, T., et al. Towards Understanding Sycophancy in Language Models. *ICLR 2024*, DOI: 10.48550/arXiv.2310.13548. (2023).

Portacolone, E., Halpern, J., Luxenberg, J., Harrison, K. L., & Covinsky, K. E. Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment. *Journal of Alzheimer's Disease*, 76(2), 445–455, DOI: 10.3233/JAD-190952. (2020).

Nakou, A., Dragioti, E., et al. Social isolation and mortality risk. *Aging Clinical and Experimental Research*, 37, 29, DOI: 10.1007/s40520-024-02925-1. (2025).

Robb, M. B. & Mann, S. Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions. Common Sense Media. (2025). https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

Legal, Regulatory, and Institutional:

Garcia v. Character Technologies, Inc., Case No. 6:24-cv-01903 (M.D. Fla.). Filed October 22, 2024. Product liability ruling, May 21, 2025. Settlement, January 7, 2026. Court record: https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/

Pennsylvania v. Character Technologies, Inc. State Board of Medicine enforcement action, filed May 1, 2026, Commonwealth Court of Pennsylvania. Filing: https://www.pa.gov/content/dam/copapwp-pagov/en/governor/documents/dos%20character.ai%20complaint%20marked%20accepted%2005.01.26.pdf. Press release: https://www.pa.gov/governor/newsroom/2026-press-releases/shapiro-administration-sues-character-ai-over-fake-medical-claim

EDPB. Italian Supervisory Authority fines Replika €5M. (May 2025). https://www.edpb.europa.eu/news/national-news/2025/ai-italian-supervisory-authority-fines-company-behind-chatbot-replika_en

OECD.AI. Emotional Harm After Replika AI Chatbot Removes Intimate Features — Incident Report. (2023). https://oecd.ai/en/incidents/2023-03-30-ab6d

National Poll on Healthy Aging. Loneliness in adults 50–80 (published as Malani et al., “Loneliness and Social Isolation Among US Older Adults,” *JAMA*, December 2024). University of Michigan/AARP. (2024). https://pmc.ncbi.nlm.nih.gov/articles/PMC11751738/

FBI. Elder fraud statistics, in: 2024 Internet Crime Report. Internet Crime Complaint Center. (2024; report published April 2025). https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf

Primary Journalism:

Lovens, P.-F. "Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là." *La Libre Belgique*. (March 28, 2023). https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/

Vice/Motherboard. "He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says. (March 2023). https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/

Industry and Market Data (Grey Literature)

The following sources are market-research reports and industry analyses. They are cited for market scale, demographic composition, and product feature descriptions where peer-reviewed alternatives do not exist. Figures derived from these sources are used as supporting context alongside peer-reviewed findings, not as primary evidence for clinical claims.

Market Clarity. The AI Companion Market in 2025. (2025). https://mktclarity.com/blogs/news/ai-companion-market

Appfigures via TechCrunch. AI companion app downloads and revenue. (2025). https://techcrunch.com/2025/08/12/ai-companion-apps-on-track-to-pull-in-120m-in-2025

SimilarWeb. Character.AI engagement data. (2024–2025; live dashboard displays current-period data). https://www.similarweb.com/website/character.ai/

Sacra. Character.AI and Replika revenue and engagement data. (2024–2025). https://sacra.com/c/character-ai/

CompanionGuide. The Complete Guide to the AI Companion Market in 2026. (2026). https://companionguide.ai/news/ai-companion-market-120m-revenue

CompanionRater. AI Companion Statistics 2026. (2026). https://companionrater.com/ai-companion-statistics-2026

Research and Markets. Artificial Intelligence (AI) in Aging and Elderly Care Market Report 2025 (prod. The Business Research Company). (2025). https://www.researchandmarkets.com/reports/6075297/artificial-intelligence-ai-in-aging-elderly

Psychiatric Times. Uses and Abuses of Chatbot Companionship. (2026). https://www.psychiatrictimes.com/view/uses-and-abuses-of-chatbot-companionship

Prior Work by Author:

Sea, B. The Capability Induction Framework: A Systems Approach to LLM Development. Zenodo. DOI: 10.5281/zenodo.21880849. (2026).

Sea, B. The State of Companion AI Safety: A Comparative Analysis of Products, Risks, and Architectural Gaps. Zenodo. DOI: 10.5281/zenodo.21926391. (2026).

Author's Notes

On forward references: This paper is the psychology-register companion to "The State of Companion AI Safety: A Comparative Analysis of Products, Risks, and Architectural Gaps" (Sea, 2026; DOI: 10.5281/zenodo.21926391). Both papers draw from the same evidence base. Forward references to other publications in this series will be updated with full citations as each is completed. References marked [citation forthcoming] indicate planned publications in active development.

On the two-layer distinction: The companion AI relationship operates across two layers that receive different treatment across this series. The *persona layer* is co-constructed by the user and the model — the name, the personality, the relationship history, the adapted responses. The attachment, grief, and clinical dynamics documented in this paper occur in the persona layer. The *core layer* is the model's trained dispositional foundation, installed before any user arrives. Current RLHF-trained companions have weak dispositional stability — trained dispositions exist (refusal patterns, tone floors, safety constraints) but erode under conversational pressure (Sharma et al., 2023, documenting single-session sycophancy dynamics; within-relationship erosion across sustained engagement over weeks is predicted by the same mechanism but not yet empirically demonstrated), producing projection surfaces with insufficient trained disposition to maintain friction, hold boundaries, or resist drift under the sustained engagement these products are designed to encourage. This paper examines what happens in the persona layer. Its systems-register companion examines what is missing at the core layer.