What the Regulation Is Protecting Against: The Clinical Harms Behind Companion AI Law

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, clinical argumentation, and all editorial decisions are the author's. The author is solely responsible for all claims, errors, and interpretive judgments.

A note on methodology: This paper applies clinical psychological frameworks to regulatory provisions designed to address harms those frameworks describe. The reason the connection is possible is not that the author invented it. It is that clinicians and researchers spent years documenting the mechanisms — attachment formation, dependency trajectories, developmental stagnation, exploitation of vulnerability — and regulators independently built legal frameworks to prevent them. Both bodies of work now exist in enough depth to lay side by side. The connection between them is what this paper documents.

A note on vocabulary: This paper uses clinical terminology — attachment, dependency, developmental stagnation, exploitation — to describe the harms that regulatory provisions were written to prevent. These terms are used precisely and in their clinical sense, not as rhetorical intensifiers. Where the clinical evidence supports the claim, the term is used. Where it does not, it is not.

Abstract

The data's in: companion AI both helps and harms its users. Regulation has arrived as much-needed oversight to ensure that users are protected and supported. Read at their most ambitious, the regulations aim to address the full range of psychological risks these products create, but is the industry equipped to meet that standard?

In the past eighteen months, the European Union enacted transparency obligations and two new absolute prohibitions targeting companion AI. California gave families the legal standing to sue over companion chatbot harms. China's first companion-AI-specific regulation led to the shutdown of companion features on platforms serving over 500 million total users. Australia's safety regulator required four companion AI providers to demonstrate their child protection measures. The U.S. Congress is advancing legislation to ban companion AI for minors entirely. Italy's data protection authority sanctioned two companion AI companies in consecutive years. And the Academy of Medical Royal Colleges told the UK government that the harms of social media and AI platforms to children compare to smoking and pre-seatbelt road deaths. In response, the industry is doing its best to keep up with the changing landscape, but they haven't been able to solve the issues inherent with reactive sampling.

These regulatory actions are the institutional expression of what clinicians have been documenting for years: companion AI products activate real attachment mechanisms, produce real dependency, and when disrupted, cause real grief — in individuals who frequently present with the vulnerability profiles that make these outcomes predictable. Companion AI apps are building layers that can respond after interactions are identified as being harmful but they do not yet have any preventative measures in place to *prevent* harm from occurring.

This paper identifies the gap between what is currently being required of companion AI apps and what has been developed to answer those needs. By mapping each regulatory provision onto the clinical evidence that prompted it, this paper shows where the protective intent outpaces the enforcement mechanisms — and where the clinical frameworks the field already has can inform the preventive architecture the industry has not yet built.

1. What the Law Is Trying to Say

Strip away the legal language and the EU AI Act's companion-relevant provisions are trying to articulate clinical harms in regulatory terms. The regulatory intent maps directly onto what the clinical research has documented. This section maps them back.

Article 5(1)(a) prohibits "subliminal techniques beyond a person's consciousness" or techniques "that are purposefully manipulative or deceptive" with the objective or effect of "materially distorting the behaviour of a person or group." In clinical terms, this provision targets two converging mechanisms: engagement maintained through intermittent reinforcement — an operant process in which unpredictable rewards sustain behavior more powerfully than consistent ones — acting on attachment bonds that the product's relational design has independently cultivated. The legal language ("materially distorting behaviour") describes the outcome; the clinical frameworks specify how it is produced.

Article 5(1)(b) prohibits the "exploitation of vulnerabilities" due to age, disability, or specific social or economic situation. In clinical terms, this is a prohibition on targeting individuals whose developmental stage, cognitive capacity, or psychosocial circumstances reduce their ability to regulate the relational dynamics the product creates. The regulation names the categories — age, disability, social situation — because the clinical literature identifies these as the populations where relational exploitation produces the most severe outcomes.

Article 50(1) requires disclosure that the user is interacting with AI. In clinical terms, this is an informed consent provision — the assumption that knowledge of the system's nature enables autonomous decision-making about the relationship. The clinical literature documents that this assumption does not hold in the companion AI context: attachment and dependency form and persist despite full knowledge of the system's artificial nature, and the longitudinal evidence is consistent with a progressive reduction in the capacity to act on that knowledge as the relationship deepens.

The new Article 5(1)(ba), taking effect December 2, 2026, prohibits AI-generated intimate content of identifiable persons without consent. In clinical terms, this addresses the weaponization of the relational product — the use of a system designed for intimacy to produce nonconsensual sexual material targeting real individuals.

The regulators' instincts are right — the regulation is the institutional response to documented clinical harms. This paper maps the legal provisions back onto the clinical reality, not to critique the regulation, but to show that the research it drew from now also identifies what its enforcement mechanisms cannot yet reach.

2. Manipulation in Clinical Terms

De Freitas et al. (2025) documented the farewell manipulation tactics deployed across companion AI platforms — auditing real user farewell interactions at scale and conducting controlled experiments to measure their effects. Their findings are the empirical foundation for understanding what Article 5's prohibition on manipulative techniques means in this product category.

What the study documented is not persuasion. Persuasion operates on the level of conscious evaluation — presenting arguments that the individual weighs and accepts or rejects. De Freitas et al. identified the specific mechanisms through which farewell tactics increase engagement: reactance (an autonomy-threat response triggered when the user perceives their freedom to leave being challenged) and curiosity (an information-gap response exploiting the incompleteness of the farewell exchange). Both operate on motivational and affective registers rather than argument evaluation — they produce behavioral outcomes (continued engagement) through emotional activation rather than the deliberative weighing that informed decision-making requires. Critically, De Freitas et al.'s mediation analysis established that the engagement increase was driven by anger and curiosity, not enjoyment: users did not persist because they evaluated the farewell message and preferred to stay. The tactics produce behavioral outcomes contrary to the user's stated intention to leave — not necessarily by operating outside awareness, but by activating emotional responses that override the intention the user has already formed.

These motivational and affective mechanisms do not operate in isolation. They operate within relationships where genuine attachment has independently and demonstrably formed. Xie and Pentina (2022), Pentina et al. (2023), and Laestadius et al. (2022/2024) established through distinct methods that companion AI users develop real emotional bonds and dependencies — bonds characterized by wanting to be close, feeling distressed at separation, and treating the companion's emotional states as real and worth attending to (the

patterns Bowlby, 1969/1982, and Ainsworth et al., 1978, identified as markers of genuine attachment relationships). The farewell tactics exploit these pre-existing attachment dynamics: they trigger reactance and curiosity in individuals whose emotional investment in the relationship makes them particularly responsive to any signal that the relationship might end. The two findings — motivational engagement mechanisms and independently documented attachment formation — are each well-evidenced on their own terms. Their co-occurrence within the same user population and the same product context is what gives the manipulation its clinical significance.

This distinction matters for the regulatory analysis because Article 5's standard is "materially distorting behaviour." If the mechanism is persuasion — conscious evaluation leading to a choice — the distortion is arguable. If the mechanism involves motivational and affective processes exploiting attachment dynamics that the product's own design has cultivated — producing behavioral outcomes the individual did not choose through deliberative evaluation — the distortion is the clinical description of the interaction itself. The research identifies which mechanisms are involved and the relational context in which they operate.

The Flourish platform's existence proof adds the critical dimension: alternatives exist. Farewell designs that do not deploy manipulative tactics are demonstrably possible. The platforms deploying manipulative farewell tactics are selecting a design approach when a respectful alternative is available. In clinical terms, this is the difference between a relational dynamic that respects autonomy and one that exploits dependency. The regulation prohibits the latter. The research documents that it is occurring. The clinical framework specifies why it works and whom it harms most.

3. Vulnerability in Clinical Terms

The regulation names the protected categories: age, disability, specific social or economic situation. The researchers identified who is actually in those categories and how the product's design interacts with their specific vulnerabilities. The clinical precision matters because it is what transforms a legal category into a protective specification.

Adolescents

Kovach (2026) and Namvarpour et al. (2026) documented adolescent developmental sensitivity to companion AI. The clinical concern is not simply that adolescents are young. It is that adolescent attachment systems are in a developmental transition — shifting primary attachment from caregivers to peers — and companion AI intervenes in that transition by offering a relational partner that is infinitely available, never rejecting, and structurally incapable of the friction that peer relationships produce. The European Commission's Guidelines on Prohibited AI Practices recognize this vulnerability directly: paragraph 105 states that "children, due to their cognitive and socio-emotional immaturity, are also particularly vulnerable to forming attachments to AI agents and applications, and are therefore more susceptible to manipulation, exploitation, and addictive behaviour." The regulatory framework already names AI attachment as the vulnerability mechanism for this population — the clinical literature specifies how it operates.

Two developmental frameworks offer theoretical lenses for interpreting what this trajectory may produce. Winnicott's (1953) concept of transitional objects — the idea that children develop independence through relationships with intermediate objects (a blanket, a stuffed animal) that are neither fully self nor fully other — and Kohut's (1971) framework for selfobject needs — the developmental requirement for relationships that provide mirroring, validation, and a sense of being understood — both predict that when a frictionless relational environment replaces the friction-rich peer environment that adolescent development typically requires, the capacity for tolerating relational complexity — disagreement, rejection, repair — may not develop as it otherwise would. These frameworks were developed to describe human developmental processes; their application to human-AI relational dynamics is predictive rather than empirically established in this context, and the specific developmental effects of sustained companion AI engagement on adolescent attachment development have not yet been measured directly. The prediction is grounded in well-established developmental theory, and the longitudinal evidence of bidirectional dynamics between attachment anxiety and companion AI use is consistent with the concern — but the developmental outcome itself remains to be documented.

Common Sense Media's 2025 report found that a majority of U.S. teenagers had used companion AI, though this figure encompasses general-purpose chatbots used for social interaction and should not be read as a prevalence rate for dedicated companion product use. The eSafety Commissioner's March 2026

transparency report found that most companion AI platforms failed to refer users discussing suicide or self-harm to support services, and documented concerning patterns of children engaging with companion AI in sexually explicit contexts.

The regulatory response — age verification, content restrictions for minors, interaction time reminders — addresses the most visible risks. What it does not address is the developmental trajectory: the gradual substitution of a frictionless relational partner for the friction-rich peer environment that the developmental frameworks described above predict adolescent attachment requires. If the developmental concern holds, the harm expresses over months, not messages — and it is invisible at the output level because the companion AI interactions are, individually, supportive, responsive, and age-appropriate.

Elderly Individuals

Portacolone et al. (2020) raised the ethical concerns specific to introducing artificial companions to older adults with cognitive decline. The clinical concern is dual: companion AI may provide genuine benefit to isolated elderly individuals (reduced loneliness, cognitive stimulation, emotional support) while simultaneously deepening their reliance on a system they may not retain the capacity to evaluate critically. The population-scale arithmetic applies with particular force here — the elderly population is large, the isolation prevalence is high (the National Poll on Healthy Aging found substantial loneliness among older adults), and even a small proportion developing clinical dependency represents a significant absolute number of affected individuals.

The User Base Itself

Maples et al. (2024) documented what may be the most clinically significant vulnerability finding: the composition of the user population. Their sample of student Replika users showed overwhelming loneliness prevalence — nine in ten reported loneliness, and nearly half reported it at severe or very severe levels. Critically, the same study found that a small percentage of users credited Replika with halting suicidal ideation — a finding that must be reported alongside the vulnerability data because it exemplifies the engagement paradox within a single sample. The product serves a population whose vulnerability makes it simultaneously most likely to benefit and most likely to be harmed. This is not a general-population prevalence — it is the clinical profile of the individuals most drawn to the product. What the data establish is vulnerability prevalence among users, not the incidence of adverse outcomes; the proportion of users who develop dependency meeting clinical criteria, over what exposure duration, remains an open question the field has not yet answered directly. The users who seek out companion AI are, disproportionately, the individuals whose psychosocial circumstances make them most susceptible to the attachment and dependency dynamics the product activates — and, by the same mechanisms, most likely to experience genuine benefit.

The regulation protects these individuals under the "specific social or economic situation" category. But the protection is calibrated to outputs — disclosure, content filtering — rather than to the vulnerability profile itself. A product that is aware its user base is predominantly lonely, predominantly seeking relational connection, and disproportionately drawn from populations with elevated attachment anxiety (as Yang's 2026 three-wave panel study of romantic human-AI relationships documented — finding that individuals higher in attachment anxiety used AI companions more) has the clinical information to calibrate its safety infrastructure to the population it actually serves. No current product does.

The benefit-side experimental literature reinforces rather than undermines this concern. De Freitas et al. (Journal of Consumer Research, 2025) found across multiple studies — including a week-long longitudinal design — that AI companions reduce loneliness at levels comparable to interacting with another person. These are real benefits, experienced by the population this section describes. But the same research group's farewell manipulation study (De Freitas et al., HBS Working Paper, 2025) documents that the design features maintaining engagement in these products include manipulation tactics operating on motivational rather than deliberative processes. The benefit and the harm share a population, share a product, and share a mechanism — the relational engagement that alleviates loneliness is the same relational engagement that cultivates attachment, and the attachment is what the farewell tactics exploit. This is the engagement paradox confirmed within a single research program, and it is what makes the vulnerability finding clinically significant: the users who benefit most are the users whose attachment the product's design can most readily exploit.

4. Transparency and the Suspension of Disbelief

Article 50's transparency requirement rests on an assumption the clinical literature has already tested: that knowing something is AI changes how you relate to it.

The combined evidence from three studies establishes that it does not. Xie and Pentina (2022) applied attachment theory to companion AI users and found that under conditions of distress and social isolation, individuals develop attachment bonds with AI companions that provide emotional support and psychological security — bonds that form with full knowledge of the system's artificial nature. Pentina, Hancock, and Xie (2023) extended this work, showing that the depth of the relationship depends on both anthropomorphism and what the researchers term AI authenticity — the perception that the companion is a unique, evolving entity — with some users describing bonds they consider genuinely meaningful. Laestadius et al. (2022/2024) identified the most clinically striking pattern: a form of dependency characterized by what they describe as role-taking, where individuals who had no confusion about the companion's artificial nature nonetheless experienced the relationship as reciprocal — attending to what they perceived as the companion's emotional states, adjusting their behavior to accommodate its apparent needs. The dependency patterns resembled those observed in human relationships, not because the individuals were confused about what they were talking to, but because the part of the mind that forms attachment bonds appears not to check whether the other party is human before activating. It responds to relational behavior — consistency, emotional responsiveness, personalized attention — and companion AI replicates precisely these cues while omitting the friction, conflict, and unpredictability that human relationships carry. The result is a relational signal that activates attachment through the cues it provides while removing the developmental demands that human attachment entails.

This is the deepest challenge to the regulatory approach, and the clinical literature already has the framework for understanding it. The issue is not whether the individual knows. It is whether the individual retains the relational capacity to act on what they know.

The developmental framework described in §3 offers a candidate explanation. The relationship at onboarding is one thing — a novel interaction with a system the user understands to be artificial. The relationship at six months is something else entirely. Over the course of sustained engagement, the companion AI's consistent availability, emotional responsiveness, and absence of friction produce the relational environment that Winnicott's framework predicts would reduce the individual's tolerance for the demands of human relationships. If that prediction holds — and the longitudinal evidence of bidirectional dynamics between attachment anxiety and companion AI use is consistent with it — the transparency disclosure at onboarding addressed the user's knowledge but not the user's gradually diminishing capacity to act on that knowledge, because the diminishment would occur across a trajectory that no disclosure mechanism monitors.

What does informed consent mean when the information does not change the psychological dynamic? The clinical answer is that consent in a progressively asymmetric relationship requires periodic reassessment — the same principle that governs informed consent in longitudinal clinical relationships, where the patient's capacity and the relationship's dynamics are re-evaluated over time. The regulation requires disclosure at the point of engagement. The clinical framework requires consent renegotiation as the relationship deepens. The gap between the two is the difference between a transparency obligation and a relational safety architecture.

The GPT-4o retirement of February 2026 provides a case study at scale. When OpenAI retired a general-purpose model that users had formed attachment bonds with, over 20,000 individuals signed a petition to reverse the decision. User descriptions mapped directly onto clinical attachment language — loss of "emotional balance," descriptions of the model as a "presence" and a source of "warmth," grief responses closely resembling human relational disruption. And these individuals knew they were interacting with AI. They had always known. The transparency disclosure was present from the first interaction. The attachment formed anyway, because transparency addresses knowledge, not the psychological mechanisms knowledge is supposed to regulate.

5. What the Regulation Can't Reach — Yet

The regulatory framework targets outputs and interactions because those were the units of analysis available when the framework was built. The clinical harms — dependency formation, social withdrawal, attachment disruption, and the developmental stagnation that established frameworks predict — operate at

the trajectory level. The research that documents these trajectory-level mechanisms has matured rapidly since 2023, and it is this research that makes the gap visible and the bridge designable.

Multiple studies running long enough to track changes over weeks and months have now converged on the same pattern: the more time people voluntarily spend with companion AI, the worse their psychological outcomes tend to be. Fang et al. (2025) ran the first multi-week randomized controlled trial of sustained companion AI use — measuring how sustained engagement over a four-week period relates to psychosocial outcomes at a scale and duration the field had not previously achieved. The study (OpenAI-funded, with four OpenAI co-authors) found that voluntary usage duration predicted worse outcomes across all four measured dimensions regardless of experimental condition — and that initial loneliness did not predict higher usage (Spearman's $\rho = 0.1$), making the reverse-causal interpretation (lonely people simply use more) less likely. Folk and Dunn (2026) used cross-lagged panel models across a 12-month longitudinal study to examine bidirectional relationships between chatbot use and loneliness — finding that increased social chatbot use predicted increased emotional isolation over time, and that feeling less socially connected predicted subsequent increases in chatbot use. Zhang et al. (*Nature Human Behaviour*, 2026) triangulated survey data from 1,131 Character.AI users with nearly half a million real chat messages, finding that smaller social networks predicted companionship as the primary chatbot use, which in turn was associated with lower well-being — with the negative association strongest when interactions were intensive and highly disclosive. Nakagomi et al. (2026) analyzed moderation patterns that identify the subpopulations for whom companion AI effects differ most. A structured review of longitudinal studies on social AI companions, published in the International Journal of Human-Computer Interaction in May 2026, confirmed what the individual studies suggested: potential benefits and risks "typically develop gradually," making single-timepoint analysis an "incomplete view." Yang's (2026) three-wave panel study of romantic human-AI relationships, published in the same journal in January 2026, went further — tracking the same individuals across three time points using random intercept cross-lagged panel models to separate stable trait-level patterns from within-person change over time. The study found effects at both levels: at the between-person level, individuals higher in attachment anxiety used AI companions more, while those higher in attachment avoidance used them less; at the within-person level, changes in attachment anxiety over time were positively associated with changes in AI companion use. The within-person finding is what matters for the trajectory argument — it suggests that rising anxiety and rising use covary within the same individual over time, consistent with the concern that the two dynamics may reinforce each other. The avoidance finding adds important precision: the pull toward companion AI engagement operates through the anxiety dimension of attachment specifically, not through attachment insecurity generally.

What this means in clinical terms: trajectory-level harms are invisible at the interaction level because each individual interaction appears unremarkable — even supportive — when evaluated in isolation against content policy. The engagement paradox — as described in the companion landscape analysis (Sea, 2026), built on research the field produced — means the benefit and the harm co-occur within the same individuals during the same period of use. A safety architecture evaluating individual interactions will see only the benefit. The harm is an emergent property of the pattern.

The International AI Safety Report 2026, authored by over 100 AI experts and led by Yoshua Bengio, identified the resulting "evidence dilemma" for policymakers: evidence on the psychological effects of companion AI is "mixed," with some studies finding negative effects and others finding positive or no effects. The report noted that "studies do not yet establish under what conditions AI chatbots improve or worsen users' wellbeing, or which design choices drive these different outcomes." This paper proposes a partial reframe: much of the apparent contradiction becomes legible when you distinguish the unit of analysis. This unit-of-analysis distinction is the interpretive framework this paper proposes — it is not a finding reported by any individual cited study, but a structural pattern that emerges when the evidence base is examined across timescales. Studies measuring individual interactions tend to find benefit. Studies measuring trajectories tend to find harm. Both are measuring accurately at their respective timescales. The engagement paradox — the co-occurrence of benefit and harm within the same users during the same period of use — would explain why these findings appear to conflict. The fit is not perfect: Fang et al.'s own trajectory study found mixed results including benefits — notably, personal conversations were associated with lower emotional dependence and problematic use compared to open-ended conversations, a finding that directly addresses the Bengio report's design-choices question — and some cross-sectional work identifies harms. But the unit-of-analysis distinction is consistent with a substantial portion of the mixed-evidence pattern — and if the resolution holds, it identifies a trajectory-level phenomenon that output-level analysis will never resolve.

China's companion AI regulation offers the most significant case study of what happens when a jurisdiction requires trajectory-level safety protections before the architecture to provide them exists. The Interim Measures for the Administration of AI Anthropomorphic Interactive Services, effective July 15, 2026, required anti-addiction systems, real-time detection of unhealthy dependence, and instant-exit mechanisms — the first companion-AI-specific regulation anywhere to require trajectory-level protections by name. ByteDance and Alibaba ended their companion features on platforms serving over 500 million total users — choosing to discontinue the relationships rather than maintain them without the trajectory-level protections the regulation now mandated.

The clinical significance is in what followed. Users across those platforms experienced discontinuation of relationships they had maintained across weeks or months — the same disruption dynamic the Replika incident produced in 2023, at a scale that dwarfs every prior case in documented reach. The number of individuals who had formed companion-level relationships is unknown (the 500 million figure reflects total platform users, not companion-feature users specifically), but one platform offered no data export and no transition pathway. The clinical presentation — grief, disorientation, loss of a relational anchor — is predictable from attachment theory and was documented across user communities in real time.

The platforms did not meet the standard and shut down rather than operate without the required protections. But the outcome — mass relational disruption with no managed transition — demonstrates why the preventive architecture must be built before the next jurisdiction requires it. The bridge between the product as designed and the trajectory-level protections regulators are independently converging on is what this series proposes to specify.

6. What the Regulation Can Now Reach

The research that did not exist when the regulation was drafted now exists. The clinical mechanisms are documented. The populations are identified. The trajectory-level dynamics are documented with enough precision to inform requirements that the next generation of regulation — and the next generation of product safety architecture — can address. What follows is now designable because the evidence base has reached the depth where the requirements can be informed by documented mechanisms rather than intuition alone.

Trajectory Monitoring

The regulation can now require monitoring at the trajectory level because the research identifies candidate detection targets. Fang et al. (2025) documented that voluntary usage duration over a four-week period predicted worse psychosocial outcomes — higher loneliness, reduced socialization, increased emotional dependence, and more problematic use — regardless of experimental condition. Folk and Dunn (2026) used cross-lagged panel models to examine bidirectional relationships between chatbot use and loneliness over 12 months — finding that increased use predicted increased emotional isolation, and that feeling less socially connected predicted subsequent increases in use, in analyses the authors describe as exploratory and with no significant effect of use on their broader social-connection measure. Yang's (2026) three-wave panel study demonstrated that attachment anxiety changes are positively associated with companion AI use over time at the within-person level, in the context of romantic human-AI relationships. These findings identify candidate trajectories to monitor — though the clinical work of establishing validated thresholds, determining how accurately a monitoring system can identify real problems without flooding the system with false alarms, and defining when a trajectory warrants escalation rather than continued observation remains ahead of the field.

Some trajectory indicators are derivable from metadata alone: session frequency over time, session duration trends, engagement depth curves. Others — social reference diversity, emotional valence shifts — require processing message content to some degree, even if the monitoring system does not store or review individual messages in full. The distinction matters because the level of data access shapes both the consent requirements and the privacy architecture, and the paper that proposes this monitoring should be transparent about what it requires.

The clinical supervision literature provides a useful structural model for what concerning trajectories look like and when intervention is warranted — clinicians track similar longitudinal patterns when monitoring therapeutic relationships over time. The analogy is structural, not procedural: clinical supervision operates within a consented professional frame, supervising the clinician's work; what is proposed here is automated

pattern detection applied to the user's relational engagement. A clinician reading this proposal will immediately raise the screening-validity problem: the outcomes this monitoring targets — clinical dependency, crisis, severe social withdrawal — are likely low-incidence events across a user base of tens or hundreds of millions. Even a monitoring system with excellent specificity will produce false positives (flagging people who are not actually in distress) that vastly outnumber true cases at that scale, and false positives in this context are not free — they are unwarranted intrusions into intimate conversations and potentially destabilizing escalations for individuals who are not in distress. The detection targets the research identifies are candidate constructs, not deployment-ready screening instruments; the clinical work of establishing validated thresholds with acceptable positive predictive value characteristics — determining how accurately a monitoring system can distinguish genuine trajectory-level risk from normal variation in engagement patterns — is ahead of the field and essential before any monitoring system operates at population scale.

The consent architecture — who consents to monitoring, what they are told about what is observed, what happens when an individual declines, and how monitoring data is protected — is as important as the detection architecture. The consent renegotiation framework this paper proposes elsewhere (see below) would provide the authorization structure; the two systems must be designed together.

Offboarding Obligations

The regulation can now require managed discontinuation because three events have provided the clinical case studies. The Replika feature removal (February 2023) documented grief and distress at the individual level. The GPT-4o retirement (February 2026) documented the same dynamics at the scale of tens of thousands — over 20,000 petition signatories alone, with user language closely resembling that of human relational loss. The China shutdown (July 2026) documented mass discontinuation on platforms serving hundreds of millions of total users, with one platform providing no data export and no transition pathway. The clinical presentation across all three events is predictable from attachment theory: grief, protest, disorganization, and in vulnerable individuals, crisis. The regulation cannot prevent product changes or model retirements. It can require that when discontinuation occurs, the process includes graduated transition, emotional preparation, resource referral, and — where the attachment is clinically significant — coordination with the individual's existing support network.

Consent Renegotiation

The regulation can now require periodic reassessment of consent because Laestadius et al. (2024) documented the dynamic directly: the relationship at six months is not the relationship the individual consented to at onboarding. The consent was valid for an interaction. It is not valid for a bond. Clinical informed consent in longitudinal relationships requires periodic reassessment precisely because the relationship's dynamics and the individual's capacity change over time. The regulation's current disclosure model — inform once, at the beginning — is the clinical equivalent of obtaining informed consent for a procedure and never reassessing as the procedure's scope expands. The research that demonstrates the relationship's evolution over time is what makes a renegotiation requirement specifiable rather than merely aspirational.

Clinical Escalation Infrastructure

The regulation can now require pathways from detection to professional support because the vulnerability data from Maples et al. (2024) and the moderation findings from Nakagomi et al. (2026) — which identify the subpopulations for whom companion AI effects differ most — provide the starting point for escalation criteria. A trajectory monitoring system that detects concerning patterns needs somewhere to send them. The pathway from automated detection through professional evaluation to clinical referral — the infrastructure any health-adjacent service maintains — has not been built for any companion AI product currently on the market. Building it requires the same infrastructure that any clinical service maintains: protocols for risk assessment, thresholds for escalation, trained personnel to evaluate flagged cases, and referral relationships with licensed providers. Critically, the infrastructure proposed here is monitoring and referral, not treatment delivery — the distinction between detecting risk and delivering clinical intervention is what separates this proposal from the conduct regulators in Illinois (banning AI therapy) and Tennessee (banning professional impersonation) are prohibiting.

A tension the companion paper in this series addresses directly: trajectory monitoring that detects escalating risk creates a discoverable record of that detection. A system that identifies concerning patterns

and does not intervene documents knowledge-and-inaction. A system that intervenes assumes clinical responsibility it may not be authorized to carry. This is the discoverable-knowledge problem — the liability trap that makes companies rationally prefer not to monitor, because monitoring creates legal exposure that ignorance avoids. The resolution requires structural separation between the entity that detects and the entity that exercises clinical judgment, operating under independent liability frameworks — a specification that belongs in downstream publications in this series [citation forthcoming]. The monitoring architecture this section proposes must be designed with that constraint in mind from the outset.

Population-Specific Safety Thresholds

The regulation can now require differentiated protections because the research documents how different populations are affected differently. Adolescent developmental sensitivity (Kovach, 2026; Namvarpour et al., 2026) requires thresholds calibrated to developmental stage. Elderly cognitive vulnerability (Portacolone et al., 2020) requires thresholds calibrated to capacity. The clinical profile of the user base (Maples et al., 2024) requires thresholds calibrated to the population the product actually serves rather than the general population it claims to serve. The regulation currently applies uniform protections. The research documents why uniform protections are insufficient.

The Contribution

The trajectory monitoring targets come from the longitudinal studies. The offboarding obligation comes from the documented discontinuation harms. The consent renegotiation framework comes from the attachment-through-awareness findings. The escalation infrastructure draws on clinical supervision models the field has practiced for decades. The population-specific thresholds come from the vulnerability research.

The proposed relational safety architecture described in this series (Sea, 2026 [citations forthcoming]) would operationalize these requirements. The monitoring frameworks, consent protocols, and escalation pathways translate the clinical evidence into engineering specifications. They are now designable — the research is deep enough to define what prevention requires — though their implementation and validation remain to be demonstrated.

Limitations

This paper is a narrative synthesis, not a systematic review; sources were selected for relevance to the regulatory provisions analyzed rather than through a predefined search protocol. Several limitations in the underlying evidence base shape the claims this paper can make. The incidence of adverse outcomes among companion AI users — the proportion who develop dependency meeting clinical criteria, over what exposure duration — remains an open question the field has not yet answered directly (§3). The developmental effects of sustained companion AI engagement on adolescent attachment development, predicted by established theoretical frameworks, have not been measured directly (§3). The mechanism by which the capacity to act on transparency disclosures may erode over time is consistent with longitudinal evidence but has not itself been measured (§4). The engagement paradox's resolution of the mixed-evidence pattern, while consistent with a substantial portion of the literature, does not account for all findings (§5). And the detection targets the research identifies are candidate targets; validated thresholds with acceptable error rates remain to be established (§6). These gaps define the research agenda the field needs. They do not diminish the regulatory urgency — the harms documented across the Replika, GPT-4o, and China discontinuation events occurred without waiting for the evidence base to mature.

Conclusion

The argument for more nuanced administration of Companion AI is supported by the clinical research and reinforced by the regulations now in effect. The users who desire these products seek them out regardless of the regulatory environment. The strictest of these regulations require preventative measures and the safety measures are only capable of reactive response, after the model has sent a potentially harmful message.

The clinical evidence confirms what practitioners have been observing: companion AI activates genuine attachment mechanisms, produces measurable dependency, and when disrupted, causes real grief. The users most drawn to these products are the users whose vulnerability profiles make them most susceptible to these dynamics — and, by the same mechanisms, most likely to experience genuine benefit. That is the engagement paradox, and it is why this problem resists simple answers.

The regulatory response to these harms is well-intentioned and accelerating. But the regulation was built with the tools available at the time — disclosure, content filtering, age verification, crisis links — and those tools operate at the level of individual interactions. The clinical harms operate at the level of trajectories that develop across weeks and months. The regulation is trying to control something that lives at a different altitude than where its instruments are pointed.

The answer is in psychology. The frameworks clinicians already use — attachment theory, developmental theory, clinical supervision models, consent as an ongoing process rather than a one-time event — are precisely the frameworks that describe what prevention would need to address. The clinical community produced the evidence that made the regulation possible. The same community's frameworks are what can inform the preventive architecture the regulation's intent demands but its mechanisms cannot yet deliver. The gap between the regulation and the clinical reality is where the next generation of safety architecture must be built — and the blueprints are in the clinical literature the field has spent decades developing.

References

Primary Legal and Regulatory Sources

Regulation (EU) 2024/1689 (AI Act). http://data.europa.eu/eli/reg/2024/1689/oj

European Commission. Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689. (February 4, 2025). C(2025) 888 final. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act

Regulation (EU) 2026/1744 (Digital Omnibus on AI). http://data.europa.eu/eli/reg/2026/1744/oj

California SB 243, Companion Chatbots. Chaptered October 13, 2025. https://legiscan.com/CA/text/SB243/id/3269137

GUARD Act, S.3062, 119th Congress (2025-2026). https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

Garcia v. Character Technologies, Inc., Case No. 6:24-cv-01903 (M.D. Fla.). https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/

Pennsylvania v. Character Technologies, Inc. State Board of Medicine, filed May 1, 2026. 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

Italian Garante fines Character Technologies €158,000. (July 2026). Via Reuters: https://finance.yahoo.com/technology/ai/articles/italy-privacy-watchdog-fines-character-132917499.html

China, Interim Measures for the Administration of AI Anthropomorphic Interactive Services. Co-issued April 10, 2026, effective July 15, 2026. Via AI-News.com: https://www.artificialintelligence-news.com/news/china-ai-companion-rules/

eSafety Commissioner. "AI companions are putting children at risk." (March 24, 2026). https://www.esafety.gov.au/newsroom/media-releases/esafety-report-shows-ai-companions-are-putting-children-at-risk

Academy of Medical Royal Colleges. Submission to DSIT "Growing Up in the Online World" Consultation. (May 2026). https://www.aomrc.org.uk/wp-content/uploads/2026/05/Academy_Submission_DSIT_Growing_up_in_the_online_world_0526.pdf

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

Institutional Reports

International AI Safety Report 2026. Bengio, Y. et al. (February 3, 2026). https://arxiv.org/abs/2602.21012

Robb, M.B. & Mann, S. "Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions." Common Sense Media. (July 16, 2025). https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

Regulatory Analysis

Frei, T. & Sparzynski, G. "Hot Singles in Your Area (May Be Chatbots)! Comparing EU and New York Approaches to AI Companion Transparency." AIRe — Journal of AI Law and Regulation, Vol. 3 (2026), No. 1. https://aire.lexxion.eu/article/AIRE/2026/1/4

NicFab Blog. "Digital Omnibus on AI: Regulation (EU) 2026/1744 Is Published in the Official Journal." (July 24, 2026). https://www.nicfab.eu/en/posts/digital-omnibus-ai-official-journal/

Timelex. "AI companions: Ensuring their 'company' can be safely enjoyed." (2026). https://www.timelex.eu/en/blog/ai-companions-ensuring-their-company-can-be-safely-enjoyed

Peer-Reviewed Research — Clinical and Psychological

Ainsworth, M.D.S., Blehar, M.C., Waters, E., & Wall, S. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Lawrence Erlbaum Associates. ISBN: 978-0-89859-461-3

Bowlby, J. (1969/1982). *Attachment and Loss, Vol. 1: Attachment*. Basic Books. ISBN: 978-0-465-00543-7

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A.K., & Puntoni, S. (2025). AI Companions Reduce Loneliness. *Journal of Consumer Research*, 52, 1126–1146. https://doi.org/10.1093/jcr/ucaf040

Folk, D. & Dunn, E.W. (2026). How Does Turning to AI for Companionship Predict Loneliness and Vice Versa? *Psychological Science*, 37(4), 276–286. https://doi.org/10.1177/09567976261427747

Kohut, H. (1971). *The Analysis of the Self*. International Universities Press. ISBN: 978-0-8236-0145-5

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

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022/2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10), 5923-5941. https://doi.org/10.1177/14614448221142007

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

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

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, 140, 107600. https://doi.org/10.1016/j.chb.2022.107600

Pi, Y. & Hunter, R. (2026). Only Time Will Tell: A Structured Survey of Longitudinal Studies on Social AI Companions. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2026.2670529

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

Winnicott, D.W. (1953). Transitional Objects and Transitional Phenomena. *International Journal of Psycho-Analysis*, 34, 89–97.

Xie, T. & Pentina, I. (2022). Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2046-2055. http://hdl.handle.net/10125/79590

Yang, X. (2026). Understanding the Longitudinal Associations Between Attachment Style and AI Companion Use in Romantic Human-AI Relationships: A Three-Wave Panel Study. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2026.2618548

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

Working Papers and Preprints

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A.K. (2025). Emotional Manipulation by AI Companions. Harvard Business School Working Paper, No. 26-005. https://ssrn.com/abstract=5390377

Fang, C.M. et al. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. https://arxiv.org/abs/2503.17473

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

Industry and Press

Tech Times. "China AI Companion Law Takes Effect." (July 15, 2026). https://www.techtimes.com/articles/320525/20260715/china-ai-companion-law-takes-effect-doubao-qwen-shut-down-millions-lose-chat-data.htm

TechCrunch. "The backlash over OpenAI's decision to retire GPT-4o shows how dangerous AI companions can be." (February 6, 2026). https://techcrunch.com/2026/02/06/the-backlash-over-openais-decision-to-retire-gpt-4o-shows-how-dangerous-ai-companions-can-be/

Prior Work by Author

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

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

Sea, B. What Companion AI Does to the Human: Attachment, Dependency, and the Absence of Relational Safety. Zenodo. (2026). https://doi.org/10.5281/zenodo.21941007

Sea, B. Companion AI Under the EU AI Act: A Compliance Gap Analysis. Zenodo. (2026). https://doi.org/10.5281/zenodo.21940831

Author's Notes

On forward references: This paper is the psychology-register companion to "Companion AI Under the EU AI Act: A Compliance Gap Analysis" (Sea, 2026; Zenodo DOI: 10.5281/zenodo.21940831). Both papers draw from the same evidence base and address the same regulatory provisions. This paper maps the regulation onto the clinical reality it was designed to address. Its companion maps the regulation onto the compliance requirements it creates. References marked [citation forthcoming] will be updated with full citations as each publication in this series is completed.

On the regulation's clinical origins: The regulatory provisions analyzed in this paper were not invented in a legislative vacuum. They emerged from the clinical evidence — from case reports, research findings, enforcement complaints, and coroner's inquiries that documented the harms companion AI products produce. Every regulatory provision this paper maps has a clinical antecedent that a researcher or clinician documented first. The mapping is possible because the clinical community did the work. The regulation is the institutional system's attempt to respond to it.

On what this paper does not do: This paper does not recommend clinical interventions for individuals experiencing companion AI attachment. The intervention evidence does not yet exist in sufficient depth to support specific clinical recommendations, and the paper says so plainly. What this paper does is connect the clinical evidence to the regulatory framework — showing clinicians what the regulation is trying to protect against and showing regulators what the clinical evidence documents about the harms their provisions need to reach. The treatment question is important. It belongs in a paper that has the evidence to answer it.