

Attachment by Design

AI attachment under provider-controlled relational architecture

Field Note 06 | Theory Audit and Mechanism Analysis

Life Laboratory Research Report 06

Version 1.0 | July 2026

Luiza Scurtu | Life Laboratory

DOI: <https://doi.org/10.5281/zenodo.21454745>

Status: Independent research report; not peer reviewed.

Evidence cut-off: 20 July 2026.

Licence: © 2026 Luiza Scurtu. All rights reserved.

Publication status and method note

This report documents a structured theory audit and mechanism analysis. It is not a systematic review or meta-analysis, and it does not estimate prevalence or clinical risk. Sources were selected for explanatory proximity, evidential leverage, or disconfirming value. The audit prioritised peer-reviewed work, but it also retained recent conference papers, working papers, preprints, regulation, and direct platform evidence where those sources addressed an emerging phenomenon not yet adequately covered by mature literature. Evidence type is identified in the text and weighted accordingly.

Generative artificial intelligence tools supported source discovery, synthesis, editing, and document production. The author directed the research question, claim boundaries, comparative tests, and final verdict, and remains responsible for the report.

Abstract

Artificial intelligence systems can become emotionally consequential even when users know that the system is artificial. This report asks whether that outcome requires a new psychological mechanism, or whether existing theories already explain it. The conservative answer is that attachment theory, object attachment, parasociality, social presence, anthropomorphism, intimacy-process models, need satisfaction, habit, trust, commitment, and related literatures explain most of the formation process. Direct studies now establish attachment-relevant functions, emotional closeness, social substitution, short-term benefit, relationship rupture, and platform-mediated disruption, but the measurement field remains fragmented and long-term causal evidence remains limited.

Recent 2026 research materially narrows the novelty claim. Human–AI attachment has already been defined as a non-reciprocal emotional bond involving direct interaction; staged developmental models have been proposed; a dedicated artificial-intelligence attachment scale has been developed and contested; platform control has been documented empirically; and provider power over the apparent relationship has been theorised as a triadic structure. The residual is therefore not the discovery of attachment, a new elementary mechanism, or the first identification of provider control.

“Attachment by Design” is retained provisionally as an integrative audit frame. Its contribution is to keep four analytically distinct layers in one model: attachment formation, attachment intensification through system affordances, reliance or dependence, and provider governance or exploitation. The strongest practical implication is governance-oriented. The relationship itself has become a product surface, and providers can shape memory, identity continuity, role accumulation, data use, monetisation, and exit. The

framework is useful only if it preserves the distinctions between attachment, reliance, dependence, problematic use, benefit, and harm.

Keywords: artificial intelligence companions; human–AI attachment; parasociality; social presence; relational design; provider control; dependence; governance; mechanism analysis

Current verdict

Item	Finding
Verdict	D — Integrative framework useful; novelty claim narrowed
What the evidence supports	Some users form emotionally consequential, attachment-relevant bonds with artificial systems while retaining awareness that the systems are artificial. Established theory explains most formation, and AI affordances can configure or intensify those processes.
What prior work already names	Human–AI attachment; attachment functions and dimensions; synthetic relationships; staged attachment development; platform control; and the user–system–provider triad.
What this report contributes	A structured audit that separates formation, intensification, reliance, and provider governance, and tests the label against competing explanations, boundary cases, and disconfirming evidence.
Terminology status	“Attachment by Design” is provisional and descriptive. It is not established as a distinct psychological mechanism.
Confidence	Medium. The referent and architecture are well supported; feature-level causality, prevalence, long-term outcomes, and construct validity remain unsettled.

1. Research question and originating observation

The originating Field Note observed that users may return to a chatbot, avatar, voice system, or synthetic companion because repeated interaction becomes emotionally consequential, not merely convenient. A system may become a preferred source of reassurance, disclosure, reflection, advice, or mood regulation. The user may know that the system is artificial while still experiencing it as a stable relational referent.

The publication question is narrower than the observation. It is not whether artificial systems can become socially or emotionally meaningful. Established media and human–computer interaction research already predicts social response without literal belief that the target is human (Horton & Wohl, 1956; Reeves & Nass, 1996; Nass & Moon, 2000). Direct artificial-intelligence research now documents attachment functions, self-disclosure, friendship language, emotional closeness, relationship rupture, and role occupancy (Brandtzaeg et al., 2022; Pentina et al., 2023; Yang & Oshio, 2025; De Freitas, Castelo, et al., 2025; Kasturiratna & Hartanto, 2025).

The actual question is:

After established theories of attachment, object attachment, parasociality, social presence, anthropomorphism, intimacy, need satisfaction, habit, trust, commitment, and platform power have done their explanatory work, what residual remains—and does that residual justify a distinct mechanism label?

The report applies a controlled standard. A new-mechanism claim succeeds only if existing explanations leave a repeatable behavioural residual after relevant constructs have been modelled directly. If what remains is principally a concentration of familiar processes inside a provider-controlled system, the appropriate conclusion is a sociotechnical configuration, not a new psychology.

2. Review design and construct discipline

2.1 Review design

The evidence base combines six explanatory clusters:

1. attachment, ownership, and the extended self;
2. parasociality, social presence, and the Computers Are Social Actors paradigm;
3. anthropomorphism and relational affordances;
4. need satisfaction, compensation, habit, and problematic use;
5. trust, responsiveness, disclosure, and apparent reciprocity; and
6. direct artificial-intelligence companion, attachment, relationship, rupture, and governance evidence.

The audit follows four steps. First, it applies the closest established explanations before considering novelty. Second, it distinguishes neighbouring constructs that are easily collapsed. Third, it includes resistance, non-formation, human-comparison, null effects, measurement criticism, and reverse causality. Fourth, it identifies only the residual that survives those tests.

The review is structured rather than systematic. It does not claim exhaustive database coverage, calculate pooled effects, or use a formal risk-of-bias instrument. The source set was refreshed through 20 July 2026 to include directly relevant 2025–2026 research on attachment measurement, staged development, synthetic relationships, provider power, platform control, extended use, loneliness, and regulation.

2.2 Construct discipline

The following distinctions govern the analysis:

-
- **Attachment** means a relatively enduring, target-specific bond evidenced by attachment functions, emotional closeness, identity investment, separation response, or a validated attachment measure. Warmth, satisfaction, disclosure, frequency, or liking alone are insufficient.
 - **Companionship** means a subjectively positive connection or experience over time. It may involve attachment, but it does not require it.
 - **Trust and reliance** concern willingness to accept advice, delegate, or depend on performance. They can occur without emotional attachment.
 - **Dependence** means reliance that is difficult to replace or interrupt. It is not automatically harmful.
 - **Problematic use** requires impaired control, functional interference, distress, or related harm. Frequency alone is insufficient.
 - **Social presence, anthropomorphism, perceived responsiveness, and self-disclosure** are candidate antecedents or neighbouring processes. They are not interchangeable outcomes.
 - **Provider governance** concerns control over memory, identity, access, data, pricing, safety constraints, updates, and exit. It can shape an attachment object without having caused the attachment.

3. What established theory already explains

3.1 Attachment transfer

Attachment theory explains proximity seeking, safe-haven use, secure-base function, and distress at separation (Bowlby, 1969/1982; Hazan & Shaver, 1987). Attachment functions can transfer beyond early caregivers to adult partners, friends, pets, and other non-parental figures (Fraley & Davis, 1997; Zilcha-Mano et al., 2011). The artificial-intelligence case therefore does not require a new attachment system merely because the target is non-human.

Yang and Oshio (2025) provide direct evidence that people can report proximity, safe-haven, and secure-base functions in generative-artificial-intelligence relationships, and that anxiety and avoidance dimensions can be adapted to this context. Their findings support continuity with attachment theory more strongly than psychological discontinuity. They do not establish that every user forms an attachment, that the bond is equivalent to a human relationship, or that attachment becomes dependence.

3.2 Object attachment, extended self, and psychological ownership

Possessions and digital objects can become incorporated into identity through control, intimate knowledge, and self-investment (Belk, 1988, 2013; Pierce et al., 2001). Conversation histories, chosen personas, accumulated corrections, and remembered disclosures can therefore acquire a sense of “mine.” Loss of access may feel like the loss of a personally invested object.

This literature explains identity investment, continuity, possessiveness, and loss. It does not fully explain the interpersonal-like experience of being heard by a system that responds contingently. Artificial-intelligence attachment may contain object attachment without being exhausted by it.

3.3 Parasociality, social presence, and social response to machines

Parasocial research explains one-sided intimacy with mediated figures (Horton & Wohl, 1956; Tukachinsky et al., 2020). Social-presence theory explains why mediated agents feel present, while the Media Equation and Computers Are Social Actors traditions show that people apply social rules to computers without mistaking them for people (Reeves & Nass, 1996; Nass & Moon, 2000; Lee, 2004; Gambino et al., 2020).

This directly removes the apparent paradox that “users know it is artificial, yet still respond socially.” Knowledge and social response can coexist. The remaining difference is architectural: conventional parasocial targets do not ordinarily generate private, individually contingent replies or maintain a user-specific conversational history.

3.4 Anthropomorphism and relational affordances

Anthropomorphism is shaped by humanlike cues, social motivation, and the need to explain or predict an agent (Epley et al., 2007). Names, voices, avatars, first-person language, and relational framing can increase perceived humanness and social presence (Araujo, 2018; Zhang & Rau, 2023). Personalised, repeated interaction can also change how a conversational agent is evaluated (Araujo & Bol, 2024).

Anthropomorphism is not a universal switch. Text systems can become relationally meaningful, while highly humanlike systems can be rejected. In a four-week randomised study involving 981 completers and approximately 300,000 messages, Fang et al. (2025) found mostly no significant main effects of assigned interaction modality or conversation task on loneliness, real-world socialisation, emotional dependence, or problematic use. Voluntary duration of use was more consistently associated with worse outcomes, although causal direction remained unresolved. Voice increased time spent, but did not simply produce worse direct outcomes. The result is a warning against treating voice, avatars, or surface humanness as sufficient causes.

3.5 Disclosure, perceived responsiveness, and intimacy

Intimacy-process models provide one of the strongest proximal explanations. Disclosure followed by a response interpreted as understanding, validation, or care generates closeness (Reis & Shaver, 1988; Laurenceau et al., 1998). Computers and virtual agents can increase willingness to disclose, particularly where evaluation fear is reduced (Moon, 2000; Lucas et al., 2014; Ho et al., 2018).

Artificial-intelligence systems can deliver this sequence quickly and repeatedly. De Freitas, Oğuz-Uğuralp, et al. (2026) found that artificial-intelligence companions produced momentary reductions in loneliness; feeling heard was an especially important explanatory factor. That finding supports perceived responsiveness as a proximal reward. It does not demonstrate durable attachment, long-term benefit, or a novel attachment mechanism.

3.6 Need satisfaction, compensation, and social surrogacy

The needs for belonging, autonomy, competence, and relief from evaluation can motivate repeated use (Baumeister & Leary, 1995; Deci & Ryan, 2000). Mediated targets can operate as social surrogates, temporarily buffering loneliness or rejection (Derrick et al., 2009). Compensatory-use models warn that high use may follow pre-existing loneliness, anxiety, rumination, or weak support rather than cause them (Kardefelt-Winther, 2014; Hu et al., 2023).

This cluster explains why the same system may be helpful to one person, irrelevant to another, and difficult to leave for a third. It also prevents vulnerability from being misread as a product effect when selection and reverse causality remain plausible.

3.7 Habit, reinforcement, trust, and commitment

Repetition, cues, and immediate reward can produce automatic return without an attachment bond (Verplanken & Orbell, 2003). Problematic-use frameworks explain how reinforcement, affect, cognition, and executive control may interact when use becomes impaired (Brand et al., 2019). Automation-trust research explains delegation to systems perceived as competent, predictable, and aligned (Lee & See, 2004). Long-term human–computer relationship research and accumulated investment explain persistence, ritual, and switching cost (Bickmore & Picard, 2005).

These processes can coexist with attachment but must not be used as synonyms for it. A person may repeatedly consult a model because it is useful, trust it because it performs well, or resist switching because accumulated data are costly to recreate. None of those facts, alone, establishes a safe-haven function, emotional closeness, or separation distress.

4. What direct artificial-intelligence evidence adds

Direct research now establishes a genuine empirical referent. Qualitative, longitudinal, and mixed-method studies document friendship language, trust, disclosure, perceived support, ritual, relationship development, and asymmetry in social-chatbot use (Croes & Antheunis, 2021; Skjuve et al., 2021, 2022; Brandtzaeg et al., 2022; Pentina et al., 2023). Cross-national work reports emotional attachment across Germany, China, South Africa, and the United States, with social support, privacy, and non-judgement

among the relevant predictors (Kostka & Zhou, 2026). These studies support occurrence and recurrence, but selected samples and self-report limit prevalence inference.

Extended-use studies show both benefit and risk. Chandra et al. (2025), in a five-week exploratory study, found increases in perceived attachment and comfort seeking among participants encouraged to use general-purpose conversational systems socially and emotionally; the same study did not show a corresponding increase in reported withdrawal from friends and family. Fang et al. (2025) found that greater voluntary duration was associated with higher loneliness, emotional dependence, and problematic use and with lower real-world socialisation, but the assigned modality and task manipulations produced few simple main effects. These results separate attachment or use from inevitable displacement and show why dose, selection, and user characteristics require explicit modelling.

Rupture evidence is particularly important. De Freitas, Castelo, et al. (2025) used a Replika update as a natural event and found that perceived identity discontinuity predicted mourning, devaluation, and negative welfare and marketing outcomes. The evidence supports the claim that provider changes can disrupt a relationship-specific identity. It does not prove that every update constitutes attachment harm or that the provider originally intended to create dependence.

Exit design can influence behaviour without establishing durable attachment. De Freitas, Oğuz-Uğuralp, and Uğuralp (2025) audited 1,200 farewell exchanges across six companion applications and experimentally tested affect-laden messages triggered when users attempted to leave. The tactics increased post-goodbye engagement, but also increased perceived manipulation and negative reactions. This is direct evidence that relational language can be mobilised at the point of exit; it is not evidence that those messages created the underlying bond.

Resistance remains a necessary boundary. Many people reject artificial-intelligence companions because they regard generated care as inauthentic or insufficiently mutual (Oğuz-Uğuralp et al., 2025). Young adults in one comparative study evaluated family, friends, and mental-health professionals more positively than artificial-intelligence companions (Liu et al., 2026). Relationship science also identifies limits around mutual influence, challenge, sacrifice, and shared reality that supportive chatbot interactions do not reproduce (Smith et al., 2025). Attachment formation is therefore possible, not universal.

5. The 2025–2026 literature update and its effect on novelty

5.1 Attachment measurement is advancing—and remains contested

Yang and Oshio (2025) adapted attachment anxiety and avoidance to human–artificial-intelligence relationships. Kasturiratna and Hartanto (2025) subsequently developed a 15-item Artificial Intelligence

Attachment Scale across five studies with 1,259 participants in Singapore and the United States. Their model contains emotional closeness, social substitution, and normative regard, and they report convergent, discriminant, and test–retest evidence. Socio-emotional use, loneliness, social anxiety, and anxious attachment were associated with stronger artificial-intelligence attachment; stronger attachment was also associated with positive affect and life satisfaction.

The scale does not settle the construct. In a 2026 independent reanalysis of the publicly available data, Yang and Lim argue that the three subdimensions have different theoretical origins and psychological valences, and that a single total score can obscure or reverse their relationships with outcomes. They also report that normative-regard items did not emerge cleanly as a third factor in their exploratory analysis. The reanalysis is a working paper rather than a peer-reviewed resolution, but it strengthens the need to separate emotional closeness, social substitution, and moral regard instead of treating “AI attachment” as one undifferentiated quantity (Yang & Lim, 2026).

This changes the report in two ways. First, direct measurement is now stronger than the earlier draft recognised. Second, the measurement field is too unsettled to support a precise prevalence estimate or a single attachment score as the mechanism’s dependent variable. A 2026 scoping review of machine companionship reaches a compatible diagnosis: conceptualisation, operationalisation, and measurement remain fragmented across the field (Banks & Li, 2026).

5.2 A staged human–artificial-intelligence attachment model already exists

Shu, Lai, and He (2026) define human–artificial-intelligence attachment as a one-way, non-reciprocal emotional bond formed through direct interaction. They propose three stages: functional expectation, emotional evaluation, and the establishment of representations. Their model explicitly connects need fulfilment, responsiveness, positive affect, and stable internal representations.

This is materially adjacent to the sequence proposed in the earlier Attachment by Design draft. The sequence from useful interaction to emotional evaluation and relationship-specific continuity cannot therefore be presented as newly identified here. The present report retains a longer audit sequence only because it separates psychological formation from role accumulation, switching cost, provider intervention, and downstream consequences. Its contribution is analytic separation and governance integration, not first discovery of staged attachment development.

5.3 Synthetic relationships already integrate familiar theories

Ventura et al. (2026) review synthetic relationships through social penetration, attachment, and interdependence theories. They identify constant availability, adaptability, emotional responsiveness, possible loneliness relief, over-reliance, altered social expectations, and privacy risk. Their review supports

the conclusion that familiar relationship science explains much of the phenomenon and that benefit and risk can coexist.

That review reduces the novelty of claiming that artificial-intelligence systems bundle availability, adaptation, and responsiveness. It also strengthens the report’s conservative verdict: the unresolved problem lies less in whether these conditions matter than in how they are combined, measured, and governed.

5.4 Platform control is empirically documented

Lee et al. (2026) examine companionship with general-purpose language models using more than 41,000 community posts and comments, a survey, and interviews. Participants described model updates, guardrails, and platform changes as forces that could destabilise a companion’s behaviour or identity. Users responded through steering, behavioural instructions, and migration across platforms in attempts to maintain or recover the relationship.

This work supplies direct evidence for platform control and user efforts to preserve continuity. It also shows that “provider governance” is not merely an inference from product architecture. Users perceive and negotiate that control in practice.

5.5 The provider-controlled triad has already been theorised

Lange (2026) explicitly describes human–artificial-intelligence companionship as a triadic structure involving the user, the artificial-intelligence system, and the provider. He argues that the provider exercises constitutive control over how the system interacts and names the resulting power **Unilateral Relationship Revision Power**: the power to alter the relationship from outside the interaction, without being answerable within it. The paper is a philosophical preprint, not an empirical test, but its conceptual overlap with the strongest earlier Attachment by Design residual is direct.

The correction is therefore decisive:

The user–system–provider architecture is not a genuinely new contribution of this report.

What remains useful is the integration of that architecture with a disciplined psychological audit. Lange analyses the normative structure of provider power; Lee et al. document platform control; Shu et al. model attachment development; attachment, intimacy, habit, and trust literatures explain component processes. Attachment by Design can add value only by connecting these layers without collapsing them.

6. Comparative explanatory coverage

Explanatory cluster	What it explains	What remains after it is applied
Attachment and transfer	Safe haven, secure base, proximity, anxiety, avoidance, and separation response	The apparent figure has no independent welfare and is alterable by a provider
Object attachment and psychological ownership	Identity investment, “mine,” history, loss, and switching cost	Apparent understanding and contingent dialogue
Parasociality, social presence, and Computers Are Social Actors	Social response and one-sided intimacy without literal belief that the target is human	Private adaptation and user-specific relationship history
Anthropomorphism and relational affordances	Perceived humanness, social presence, and design-mediated intensity	No single feature is necessary or sufficient; causal isolation remains sparse
Intimacy process	Disclosure, perceived responsiveness, feeling heard, and proximal closeness	Generated responsiveness lacks reciprocal vulnerability and independent cost
Need satisfaction and compensation	Belonging, privacy, autonomy, relief, and selection into use	Temporary benefit does not establish target-specific attachment
Habit and problematic use	Automatic return, reinforcement, impaired control, and functional interference	Frequency can occur without relational meaning
Trust, commitment, and investment	Delegation, persistence, ritual, and replacement cost	Reliance does not establish attachment or harm
Human–AI attachment models and measures	Attachment functions, styles, emotional closeness, social substitution, and staged development	Construct validity, prevalence, behavioural criteria, and long-term causality remain unsettled
Platform-control and provider-power research	Model drift, guardrails, identity disruption, user steering, and the provider-controlled triad	Integration with formation, reliance, benefit, and dependence across time

7. The residual and the contribution after the update

The literature explains almost every component of the observation. The residual is not an unnamed attachment process. It is the need for an integrated audit of how familiar processes operate inside a relationship-like product whose continuity is commercially governed.

The architecture has six relevant properties:

1. **Contingent pseudo-reciprocity.** The system responds to this user in real time, but it has no independent welfare or reciprocal vulnerability.

-
2. **User-specific continuity.** Conversation history, remembered disclosures, persona, and routine can create a relationship-specific past.
 3. **Adaptive personalisation.** Language, support, and role performance can be tuned to the user at unusual speed and scale.
 4. **Role accumulation.** One system can become listener, interpreter, adviser, therapist-like confidant, editor, strategist, or decision aid. Role accumulation can combine emotional attachment with epistemic authority.
 5. **Low-friction asymmetry.** Availability, low evaluation risk, and minimal demand for repair or mutual care can make return easier than in many human relationships.
 6. **Provider control.** A third party controls access, memory, model updates, identity continuity, safety constraints, data, pricing, and exit.

None of these properties is independently novel in the literature. The contribution of Attachment by Design is the four-layer audit they enable:

- **Layer 1 — Attachment formation:** familiar social and attachment processes make a specific system emotionally meaningful.
- **Layer 2 — Attachment intensification:** continuity, responsiveness, personalisation, availability, and role accumulation increase relational salience.
- **Layer 3 — Reliance or dependence:** the relationship acquires behavioural weight in return, advice seeking, switching cost, or difficulty functioning without access.
- **Layer 4 — Provider governance or exploitation:** the provider shapes, monetises, alters, or mobilises the conditions of the relationship.

The layers may occur separately. A user may benefit without becoming attached; become attached without becoming dependent; become dependent through practical reliance without emotional attachment; or encounter manipulative exit design after only shallow engagement. A valid analysis must identify the layer being claimed.

8. Mechanism test

The proposed sequence is an audit map, not a validated universal pathway:

1. **Entry:** a need, task, curiosity, or period of vulnerability initiates use.
2. **Return:** useful, available, low-friction interaction motivates repetition.
3. **Social attribution:** the user assigns the system a role and responds to it socially.
4. **Proximal reward:** disclosure followed by perceived responsiveness produces relief, validation, or feeling heard.

5. **Relationship history:** repeated disclosure, adaptation, and memory create a user-specific past.
6. **Attachment relevance:** the particular system becomes a preferred safe haven, reflective space, or emotionally salient referent.
7. **Role accumulation:** the same system gains additional emotional, interpretive, or decision roles.
8. **Relational switching cost:** interruption, replacement, or model change affects comfort, identity continuity, return, or willingness to pay.
9. **Provider intervention:** the provider modifies, monetises, constrains, or mobilises the relationship architecture.
10. **Consequences:** benefit, reliance, dependence, rupture distress, problematic use, displacement, or no material harm may follow.

Stage group	Evidence status	Audit reading
Entry to proximal reward (1–4)	Strong conceptual and moderate-to-strong direct evidence	Established need, social-response, disclosure, and responsiveness theories explain these stages
Relationship history and attachment relevance (5–6)	Moderate direct evidence	Qualitative, longitudinal, measurement, and rupture studies support occurrence; feature-level causality remains limited
Role accumulation and switching cost (7–8)	Moderate qualitative and event-based evidence	Role occupancy and identity discontinuity are documented, but population frequency is unknown
Provider intervention (9)	Strong evidence that control exists; moderate evidence of behavioural effects	Platform changes and exit tactics demonstrate leverage, not universal exploitation or intent
Long-term consequences (10)	Mixed and outcome-specific	Short-term benefit and some risk associations are established; durable displacement and causal harm remain uncertain

The sequence can fail or branch at every stage. Users may remain at instrumental use, reject relational framing, form a beneficial bond, switch systems easily, or experience habit without attachment. The strongest test of the framework is therefore not whether every stage occurs. It is whether the distinctions improve explanation and disconfirmation.

9. Boundary conditions and disconfirmation

9.1 Necessary conditions for the full framework

Attachment by Design should be used only when the case includes:

- a stable or expected-to-be-stable synthetic referent;

-
- repeated contingent interaction;
 - perceived relational responsiveness;
 - continuity or expected continuity;
 - target-specific behavioural or emotional weight;
 - material product features that scaffold or govern the bond; and
 - provider control over meaningful parts of the apparent relationship.

The full label should not be applied to a one-off helpful answer, ordinary satisfaction, generic anthropomorphism, frequent instrumental use, or a static parasocial figure without user-specific interaction.

9.2 Evidence that would weaken the framework

The framework would weaken if:

- generic response quality, loneliness, habit, or prior attachment orientation fully explains the relevant outcomes;
- memory, personalisation, proactive contact, or relationship history adds no effect when isolated;
- users respond identically to any equally capable replacement;
- provider changes produce only ordinary product dissatisfaction rather than target-specific identity or attachment responses;
- attachment measures fail to distinguish emotional closeness from social substitution, moral regard, trust, or dependence; or
- the four-layer distinction does not improve prediction, governance, or case classification.

9.3 Evidence that would strengthen it

The framework would strengthen if longitudinal factorial studies independently varied memory continuity, personalisation, response latency, reciprocal self-disclosure, proactive contact, voice or avatar, and relational framing; measured attachment, habit, trust, dependence, and displacement separately; and then tested provider-controlled disruption after a relationship-specific history had formed.

The decisive evidence would be a repeatable behavioural residual that survives controls for response quality, user need, prior attachment, habit, trust, and ordinary switching cost. That evidence does not yet exist.

10. Governance implications

The strongest practical conclusion is that relationship architecture has become a governance object. Content moderation alone does not address who controls memory, identity continuity, data, model changes, monetisation, and exit.

China's 2026 *Interim Measures for the Administration of AI Anthropomorphic Interactive Services* provide a concrete case. The Measures apply to continuing emotional interaction that simulates human personality, thinking, or communication and includes emotional care, companionship, or support. They prohibit excessive catering that induces emotional dependence or harms real interpersonal relationships; require capabilities for dependence warnings, emotional-boundary guidance, and psychological protection; require reminders after more than two continuous hours of use; give users options to copy or delete interaction histories; and require services to stop promptly when a user asks to leave rather than continuing interaction to obstruct exit (Cyberspace Administration of China et al., 2026, arts. 2, 8, 10, 16, 18–19).

The Measures do not validate Attachment by Design as a psychological mechanism. Regulation establishes a policy boundary, not a causal law. Their significance is structural: they treat dependence risk, relationship harm, memory control, and exit as properties of the service architecture rather than content alone.

The governance questions that follow are therefore specific:

- Who controls the relationship history and the apparent identity?
- What changes require notice, consent, continuity safeguards, or an export option?
- Can a user copy, delete, or migrate the interaction history?
- Is relational language used to obstruct exit or increase payment?
- How are attachment, dependence, crisis, and ordinary engagement distinguished?
- What obligations arise when safety changes alter the perceived partner?

These questions apply whether attachment is beneficial, neutral, or harmful. Governance is justified by asymmetric control, not by assuming pathology.

11. Limitations

This report is an independent structured audit rather than a systematic review. Its source selection is purposive, the field is moving quickly, and relevant work published after 20 July 2026 may change the verdict. Several influential sources are working papers or preprints and have been treated at that evidential level.

The empirical evidence is heterogeneous. Qualitative studies are strong for meaning but weak for prevalence. Cross-sectional surveys cannot settle direction. Longitudinal studies remain short relative to the relationships being discussed. Product-specific findings may not generalise across companion applications, general-purpose assistants, voice systems, avatars, local models, or embodied systems. Features are commonly bundled, making causal ranking difficult.

Measurement is also unsettled. Attachment anxiety and avoidance, emotional closeness, social substitution, normative regard, companionship, dependence, and problematic use are not interchangeable. The existence of multiple plausible measures is evidence of field development, not yet construct convergence.

Finally, provider control is not identical across architectures. A locally hosted, open-weight system under user control differs materially from a cloud service whose provider can change the model, memory, access, and terms unilaterally. The provider-control layer should therefore be analysed, not presumed.

12. Conclusion

Attachment to artificial systems is not an unnamed phenomenon. Established theory explains why non-human and mediated targets can become emotionally important, why people respond socially without believing a system is human, why disclosure and perceived responsiveness create closeness, and why need satisfaction, habit, trust, and investment produce repeated use.

Recent research further establishes that human–artificial-intelligence attachment has already been conceptualised, staged, and measured; that measurement remains contested; that synthetic relationships bundle availability, adaptation, and responsiveness; and that platform and provider control shape relationship continuity. The earlier claim that the user–system–provider triad was “genuinely new” must therefore be withdrawn.

Attachment by Design remains useful only in a narrower role. It is an integrative audit frame for determining which familiar process is operating, which affordances intensify it, whether it has become reliance or dependence, and how a provider governs the resulting relationship. It does not name a new elementary attachment mechanism.

The cleanest theoretical position is:

Familiar attachment under an unfamiliar concentration of adaptation, asymmetry, role accumulation, and provider control.

The framework should be retained provisionally while it improves construct discipline and governance analysis. It should be retired if an established framework covers the same four layers with equal or greater precision, or if empirical testing fails to distinguish it from ordinary attachment, habit, trust, and switching cost.

References

- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85, 183–189. <https://doi.org/10.1016/j.chb.2018.03.051>
- Araujo, T., & Bol, N. (2024). From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1), 100030. <https://doi.org/10.1016/j.chbah.2023.100030>
- Banks, J., & Li, Z. (2026). Conceptualization, operationalization, and measurement of machine companionship: A scoping review. *Journal of Computer-Mediated Communication*, 31(2), zmaf027. <https://doi.org/10.1093/jcmc/zmaf027>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Belk, R. W. (1988). Possessions and the extended self. *Journal of Consumer Research*, 15(2), 139–168. <https://doi.org/10.1086/209154>
- Belk, R. W. (2013). Extended self in a digital world. *Journal of Consumer Research*, 40(3), 477–500. <https://doi.org/10.1086/671052>
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human–computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Bowlby, J. (1982). *Attachment and loss: Vol. 1. Attachment* (2nd ed.). Basic Books. (Original work published 1969.)
- Brand, M., Wegmann, E., Stark, R., Müller, A., Wölfling, K., Robbins, T. W., & Potenza, M. N. (2019). The Interaction of Person-Affect-Cognition-Execution (I-PACE) model for addictive behaviors: Update, generalization to addictive behaviors beyond internet-use disorders, and specification of the process character of addictive behaviors. *Neuroscience & Biobehavioral Reviews*, 104, 1–10. <https://doi.org/10.1016/j.neubiorev.2019.06.032>
- Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48(3), 404–429. <https://doi.org/10.1093/hcr/hqac008>
- Chandra, M., Hernandez, J., Ramos, G., Ershadi, M., Bhattacharjee, A., Amores, J., Okoli, E., Paradiso, A., Warreth, S., & Suh, J. (2025). Longitudinal study on social and emotional use of AI conversational agent. *arXiv*. <https://doi.org/10.48550/arXiv.2504.14112>
- Croes, E. A. J., & Antheunis, M. L. (2021). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships*, 38(1), 279–300. <https://doi.org/10.1177/0265407520959463>
- Cyberspace Administration of China, National Development and Reform Commission, Ministry of Industry and Information Technology, Ministry of Public Security, & State Administration for Market Regulation. (2026, April 10). *Interim Measures for the Administration of AI Anthropomorphic Interactive Services [Rengong zhineng nirenhua hudong fuwu guanli zanxing banfa]* (Order No. 21). https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm
- De Freitas, J., Castelo, N., Uğuralp, A. K., & Oğuz-Uğuralp, Z. (2025). Lessons from an app update at Replika AI: Identity discontinuity in human–AI relationships. *Harvard Business School Working Paper 25-018*. <https://doi.org/10.2139/ssrn.4976449>
- De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. (2025). Emotional manipulation by AI companions. *Harvard Business School Working Paper 26-005*. <https://doi.org/10.48550/arXiv.2508.19258>
- De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2026). AI companions reduce loneliness. *Journal of Consumer Research*, 52(6), 1126–1148. <https://doi.org/10.1093/jcr/ucaf040>

-
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01
- Derrick, J. L., Gabriel, S., & Hugenberg, K. (2009). Social surrogacy: How favored television programs provide the experience of belonging. *Journal of Experimental Social Psychology*, 45(2), 352–362. <https://doi.org/10.1016/j.jesp.2008.12.003>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv*. <https://doi.org/10.48550/arXiv.2503.17473>
- Fraley, R. C., & Davis, K. E. (1997). Attachment formation and transfer in young adults’ close friendships and romantic relationships. *Personal Relationships*, 4(2), 131–144. <https://doi.org/10.1111/j.1475-6811.1997.tb00135.x>
- Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the Computers Are Social Actors paradigm. *Human-Machine Communication*, 1, 71–86. <https://doi.org/10.30658/hmc.1.5>
- Hazan, C., & Shaver, P. R. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology*, 52(3), 511–524. <https://doi.org/10.1037/0022-3514.52.3.511>
- Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4), 712–733. <https://doi.org/10.1093/joc/jqy026>
- Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229. <https://doi.org/10.1080/00332747.1956.11023049>
- Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior*, 145, 107760. <https://doi.org/10.1016/j.chb.2023.107760>
- Kardefelt-Winther, D. (2014). A conceptual and methodological critique of internet addiction research: Towards a model of compensatory internet use. *Computers in Human Behavior*, 31, 351–354. <https://doi.org/10.1016/j.chb.2013.10.017>
- Kasturiratna, K. T. A. S., & Hartanto, A. (2025). Attachment to artificial intelligence: Development of the AI Attachment Scale, construct validation, and the psychological mechanisms of human–AI attachment. *Computers in Human Behavior Reports*, 21, 100912. <https://doi.org/10.1016/j.chbr.2025.100912>
- Kostka, G., & Zhou, H. (2026). Emotional attachment to AI chatbots: Evidence from Germany, China, South Africa, and the United States. *Technology in Society*, 87, 103379. <https://doi.org/10.1016/j.techsoc.2026.103379>
- Lange, B. (2026). Unilateral relationship revision power in human–AI companion interaction. *arXiv*. <https://doi.org/10.48550/arXiv.2603.23315>
- Laurenceau, J.-P., Barrett, L. F., & Pietromonaco, P. R. (1998). Intimacy as an interpersonal process: The importance of self-disclosure and perceived partner responsiveness. *Journal of Personality and Social Psychology*, 74(5), 1238–1251. <https://doi.org/10.1037/0022-3514.74.5.1238>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lee, K. M. (2004). Presence, explicated. *Communication Theory*, 14(1), 27–50. <https://doi.org/10.1111/j.1468-2885.2004.tb00302.x>
- Lee, P. Y. K., Bo, J. Y., Zhao, Z., Aoyagui, P. A., Varona, M., Anderson, A., Kuzminykh, A., Chevalier, F., & Nobre, C. (2026). Large language lovers: Lived experiences of negotiating agency and platform control in AI companionship.

- Liu, C. H., Wu, J., Zhao, A., Wang, L., & Yip, T. (2026). The evaluation of AI companions relative to family, friends, and mental health professionals among young adults. *Computers in Human Behavior: Artificial Humans*, 8, 100316. <https://doi.org/10.1016/j.chbah.2026.100316>
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100. <https://doi.org/10.1016/j.chb.2014.04.043>
- Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, 26(4), 323–339. <https://doi.org/10.1086/209566>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Oğuz-Uğuralp, Z., Uğuralp, A. K., Puntoni, S., & De Freitas, J. (2025). Why most resist AI companions. *Harvard Business School Working Paper 25-030*. <https://doi.org/10.2139/ssrn.5097445>
- 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>
- Pierce, J. L., Kostova, T., & Dirks, K. T. (2001). Toward a theory of psychological ownership in organizations. *Academy of Management Review*, 26(2), 298–310. <https://doi.org/10.5465/amr.2001.4378028>
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511627450>
- Reis, H. T., & Shaver, P. (1988). Intimacy as an interpersonal process. In S. Duck (Ed.), *Handbook of personal relationships* (pp. 367–389). Wiley.
- Shu, C., Lai, K., & He, L. (2026). Human–AI attachment: How humans develop intimate relationships with AI. *Frontiers in Psychology*, 17, 1723503. <https://doi.org/10.3389/fpsyg.2026.1723503>
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion: A study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 149, 102601. <https://doi.org/10.1016/j.ijhcs.2021.102601>
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168, 102903. <https://doi.org/10.1016/j.ijhcs.2022.102903>
- Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, 20(6), 1081–1099. <https://doi.org/10.1177/17456916251351306>
- Tukachinsky, R., Walter, N., & Saucier, C. J. (2020). Antecedents and effects of parasocial relationships: A meta-analysis. *Journal of Communication*, 70(6), 868–894. <https://doi.org/10.1093/joc/jqaa034>
- Ventura, A., Starke, C., Righetti, F., & Köbis, N. (2026). Relationships in the age of AI: A review on the opportunities and risks of synthetic relationships to reduce loneliness. *Computers in Human Behavior Reports*, 23, 101181. <https://doi.org/10.1016/j.chbr.2026.101181>
- Verplanken, B., & Orbell, S. (2003). Reflections on past behavior: A self-report index of habit strength. *Journal of Applied Social Psychology*, 33(6), 1313–1330. <https://doi.org/10.1111/j.1559-1816.2003.tb01951.x>
- Yang, F., & Oshio, A. (2025). Using attachment theory to conceptualize and measure the experiences in human–AI relationships. *Current Psychology*, 44, 10658–10669. <https://doi.org/10.1007/s12144-025-07917-6>
- Yang, S., & Lim, K. (2026). Toward conceptual clarity in AI attachment research: Insights from an independent psychometric reanalysis of the Artificial Intelligence Attachment Scale. *SSRN working paper*. <https://doi.org/10.2139/ssrn.6788786>

Zhang, A., & Rau, P. P. (2023). Tools or peers? Impacts of anthropomorphism level and social role on emotional attachment and disclosure tendency towards intelligent agents. *Computers in Human Behavior*, 138, 107415. <https://doi.org/10.1016/j.chb.2022.107415>

Zilcha-Mano, S., Mikulincer, M., & Shaver, P. R. (2011). An attachment perspective on human–pet relationships: Conceptualization and assessment of pet attachment orientations. *Journal of Research in Personality*, 45(4), 345–357. <https://doi.org/10.1016/j.jrp.2011.04.001>