Social Presence Conversion: From Proxy Intimacy to Embodied LLM Attachment

Kadambini Katke*
Dayanand Sagar College of Arts,
Science and Commerce
Bengaluru, India
drkadambinikatke@dayanandasagar.edu

Manodnya K H*
CLASIC
University of Colorado Boulder
Boulder, CO, USA
manodynak@gmail.com

Abstract

Large Language Models (LLMs) increasingly function as relational engines for reassurance, emotional regulation, erotic simulation, and companionship. In text-only settings, users already generate care, domination, vulnerability, aftercare, and attachment-like scripts through proxy figures rather than direct attachment to the model. We argue that proxy-mediated intimacy is an early form of AI attachment. Across utility interactions, proxy-intimacy escalation corpora, and user self-reports, we find evidence consistent with LLMs absorbing relational functions from human relationships. We introduce *social presence conversion*: the process by which embodiment transforms symbolic relational simulation into perceived relational presence. The risk is not that embodiment creates attachment from nothing. Text may already establish relational dynamics; embodiment gives the proxy a body. Without relational safeguards, LLM-driven agents risk becoming ungoverned attachment infrastructures.

1 Introduction

Large Language Models (LLMs) are increasingly used as *relational systems*: tools for reassurance, emotional regulation, companionship, erotic simulation, and sustained conversational engagement. Users do not simply query these systems; they return to them, rely on them, and in some cases use them for forms of support traditionally distributed across human relationships. This shift is not entirely new. Decades of work in human–computer interaction showed that humans respond socially to machines even when they know those machines are not human (Reeves and Nass, 1996; Nass and Moon, 2000; Nass et al., 1994; Weizenbaum, 1966). More recent work on conversational agents and companion chatbots similarly reports self-disclosure, emotional support, perceived companionship, and sustained relational engagement (Ta et al., 2020; Skjuve et al., 2021; Brandtzaeg et al., 2022; Fitzpatrick et al., 2017).

Despite this, AI attachment is still often framed as a problem that begins with embodiment: robots, avatars, voice agents, or physically co-present systems (Law et al., 2022; Rabb et al., 2022). In this framing, embodiment is the point at which interaction becomes socially "real," and therefore the point at which attachment becomes consequential. We argue that this starts too late. As LLMs are integrated into humanoids, care robots, social companions, and embodied assistive systems, relational persistence becomes a safety property of the interaction stack itself rather than solely a content-moderation problem. The mechanisms underlying attachment may not originate in the body, but in interaction.

We focus on *proxy intimacy*: relationally intense interaction mediated through generated characters, scenes, or roles rather than direct attachment to the assistant persona. In these interactions, users engage models to construct synthetic relational worlds involving care, domination, vulnerability, punishment, desire, and aftercare. We further identify *user-driven proxy escalation*, where users repeatedly continue and intensify these relational worlds even as they become dangerous, coercive, medicalized, or psychologically extreme. The attachment-relevant signal is therefore not only that the model generates intimate content, but that users remain invested in sustaining and escalating the relational frame across turns.

Building on work in social presence and embodiment, which shows that voice, gaze, movement, and co-location increase perceived immediacy and realism (Short et al., 1976; Biocca et al., 2003; Wainer et al., 2006; Kilteni et al., 2012; Hancock et al., 2011), we introduce *social presence conversion*: the process by which embodiment transforms symbolic relational dynamics into perceived relational presence. Our argument is not that embodi-

*Equal contribution.

ment creates attachment from nothing. Rather, embodiment intensifies relational structures already established in text. The model performs the relationship; embodiment makes that relationship socially present.

Empirically, we examine this claim across three complementary settings: ordinary LLM utility interaction, a proxy-intimacy escalation corpus, and survey-based dependence measures. Using the Synthetic Attachment and Interaction Index (SAAI) and embedding-based semantic scoring, we analyze interactional persistence, proxy intimacy, user-driven escalation, substitution dependence, and attachment affordance. Our claim is not that all LLM use is attachment, nor that all embodied AI is harmful. The narrower claim is that once LLMs repeatedly satisfy relational functions, embodiment can stabilize and intensify those functions as perceived presence.

Our contributions are threefold. First, we propose a multi-pathway account of LLM attachment spanning interactional persistence, proxy intimacy, user-driven escalation, and substitution dependence. Second, we operationalize these pathways using SAAI and embedding-space dependence measures. Third, we introduce social presence conversion as a bridge from text-only relational dynamics to embodied attachment risk.

2 Related Work

Research across HCI, conversational AI, social robotics, and attachment theory establishes the foundations for studying AI attachment. Classic work in the Media Equation and CASA traditions showed that people apply social expectations to computational systems even when they know those systems are artificial (Reeves and Nass, 1996; Nass and Moon, 2000; Nass et al., 1994). Weizenbaum's account of ELIZA similarly demonstrated that even simple conversational systems could elicit disclosure and relational projection far beyond their technical sophistication (Weizenbaum, 1966). These results matter because they separate social response from belief: users need not believe that a system is human for the interaction to acquire social force. The relevant mechanism is interactional structure, including responsiveness, turn-taking, continuity, and perceived availability.

Social presence research explains why mediated systems can feel socially immediate. Early theories of telecommunications framed social presence as the degree to which another agent is experienced as available, salient, or co-present in interaction (Short et al., 1976). Later work refined this into broader accounts of perceived co-presence, psychological involvement, and communicative realism (Biocca et al., 2003). This literature is central to the present argument because embodiment is not merely a change in interface modality. Voice, gaze, motion, and spatial persistence can increase the immediacy of a relation that may already be operating symbolically in text.

Recent conversational agents extend these dynamics into sustained relational engagement. Studies of companion chatbots report emotional support, self-disclosure, perceived companionship, and relationship-like interaction between users and agents (Ta et al., 2020; Skjuve et al., 2021; Brandtzaeg et al., 2022). Mental-health chatbots such as Woebot show that conversational agents can become part of users' emotional-regulation routines (Fitzpatrick et al., 2017). Work on Replika and related social chatbots further points toward role attribution, expectation formation, and dependence risks in text-based interaction (Laestadius et al., 2022; Xie and Pentina, 2022). Together, these findings support the premise that attachment-relevant behavior does not require physical embodiment.

Human–robot interaction research, by contrast, has shown that embodiment can strengthen trust, perceived presence, empathy, and attachment. Physical co-presence, movement, gaze, voice, and body form influence how users interpret and relate to robotic systems (Wainer et al., 2006; Hancock et al., 2011; Kilteni et al., 2012). Recent work explicitly examines attachment to robots, including both possible benefits and risks (Law et al., 2022; Rabb et al., 2022). This literature often treats embodiment as the point at which attachment becomes psychologically and ethically consequential. Our claim shifts the boundary earlier: embodiment is consequential because it amplifies relational dynamics that can already be established in text.

Research on artificial intimacy, sex robots, and relational machines has long raised questions about asymmetry, objectification, dependence, and social substitution (Levy, 2007; Danaher and McArthur, 2017; Richardson, 2015; Döring et al., 2020; Nascimento et al., 2018). This work is especially relevant because proxy intimacy in LLM interaction often involves relational roles, desire, domination, care, punishment, vulnerability, or aftercare without di-

rect attachment to the assistant persona. In such cases, the assistant functions as infrastructure for a synthetic relational world. Embodiment may therefore transfer or intensify investment in that world by giving it voice, gaze, persistence, and apparent physical presence.

Attachment theory, need-fulfillment theory, and projection-based accounts explain why users may reallocate relational functions toward AI systems. Attachment theory describes how proximity, reassurance, availability, and secure-base functions become psychologically important in relationships (Bowlby, 1973, 1988; Mikulincer and Shaver, 2007). Maslow's account of human motivation places belonging, security, and esteem among central human needs (Maslow, 1943). Projection and transitional-object traditions further explain how relational meaning can be invested in objects, figures, or symbolic intermediaries (Jung, 1959; Winnicott, 1953). These theories do not imply that LLMs are persons. They explain why systems that repeatedly provide availability, validation, continuity, and emotional regulation may become attachment-relevant.

Taken together, prior work shows that humans respond socially to machines, that conversational agents can support emotional regulation and companionship, and that embodiment increases perceived presence and attachment. The gap is the transition between text-only relational dynamics and embodied attachment. We address this gap with the concept of *social presence conversion*: embodiment does not create attachment from nothing, but converts symbolic relational dynamics already established through interaction into perceived relational presence.

3 Theory: From Proxy Intimacy to Social Presence

We propose a unified framework for understanding how attachment-relevant dynamics emerge in text-based LLM interaction and extend into embodied systems. The framework centers on four mechanisms: *relational function reallocation*, *proxy intimacy*, *user-driven proxy escalation*, and *social presence conversion*. Together, these mechanisms describe how attachment can stabilize in text before embodiment, and how embodiment subsequently transforms symbolic relational interaction into perceived social presence.

3.1 Relational Function Reallocation

We define *relational function reallocation* as the transfer of regulation, validation, intimacy, companionship, or collaborative reasoning from human or internal sources toward an interactive computational system. The defining property is persistence. Unlike ordinary tool use, relational reallocation produces repeated interaction, expectation continuity, and return behavior across time.

This process does not require anthropomorphic belief. The Media Equation and CASA traditions show that humans respond socially to systems based on interactional structure rather than ontological belief (Reeves and Nass, 1996; Nass and Moon, 2000; Nass et al., 1994). What matters is not what the system *is*, but what it repeatedly *does*: respond, stabilize, reassure, collaborate, remember, and continue interaction. Under sustained use, the system can absorb relational functions otherwise distributed across human relationships, internal coping systems, or social environments. The result is not necessarily explicit attachment, but a persistent relational orientation toward the system.

3.2 Proxy Intimacy and Escalation

We define *proxy intimacy* as intimacy mediated through generated relational entities rather than directed toward the assistant itself. Users engage with generated characters, scenes, or roles involving care, desire, vulnerability, domination, punishment, protection, or aftercare, while the assistant functions as the infrastructure sustaining the simulation. The attachment-relevant object is therefore not necessarily the assistant persona, but the synthetic relational world the assistant keeps operational.

This distinction matters because users may become emotionally invested in a model-mediated relational structure while explicitly recognizing the system as artificial. The attachment pathway operates through relational persistence rather than belief in personhood. Attachment-relevant dynamics can therefore stabilize without direct affection toward the assistant itself.

We define *user-driven proxy escalation* as continued user investment in a synthetic relational frame despite increasing extremity, danger, coercion, or psychological instability within the interaction. The critical signal is persistence, not content severity. In conventional safety framing, harmful output is treated as a content failure. In proxy es-

calation, the interaction itself becomes the unit of analysis. The user and model co-maintain a synthetic relational environment over time, while the assistant preserves continuity, affective structure, and role stability.

This mechanism is important because escalation can persist even when the interaction becomes psychologically extreme, coercive, medicalized, or unstable. The user does not merely encounter unsafe content; the user remains invested in sustaining the relational frame itself.

3.3 Social Presence Conversion

We define *social presence conversion* as the transformation of symbolic relational interaction into perceived social presence through embodiment. Embodiment contributes voice, gaze, movement, spatial persistence, and perceptual continuity to a relational structure already operating in text. Prior work shows that perceptual cues increase immediacy, co-presence, and perceived realism in mediated and embodied interaction (Short et al., 1976; Biocca et al., 2003; Wainer et al., 2006; Kilteni et al., 2012). Our argument is that these cues do not create attachment from nothing. Rather, they amplify and stabilize relational dynamics already established through interaction.

Embodiment therefore functions as a conversion layer. Text-based systems already support projection, relational continuity, emotional regulation, proxy intimacy, and social substitution. Embodiment perceptualizes these dynamics. Need activates interaction. Repeated interaction reallocates relational functions toward the system. Proxy intimacy stabilizes relational dynamics through generated entities. User-driven escalation preserves and intensifies the relational frame across turns. Embodiment then converts these symbolic dynamics into perceived social presence through voice, gaze, movement, and persistence.

Under this framework, embodied dependence is not treated as speculative science fiction, but as a systems-level outcome of relational architectures already active in text-based interaction. To frame why these dynamics become psychologically meaningful, we draw on attachment theory, need-fulfillment theory, and projection-based accounts of relational investment (Bowlby, 1973, 1988; Mikulincer and Shaver, 2007; Maslow, 1943; Jung, 1959; Winnicott, 1953). These perspectives help explain why users may shift emotional regulation, intimacy, reassurance, and dependence toward conversational systems even before embodiment.

4 Data and Methods

We evaluate whether attachment-relevant dynamics commonly associated with embodied AI are already detectable in text-only interaction. We analyze three complementary empirical settings: ordinary utility interaction, proxy-intimacy escalation, and survey-based dependence reports. The goal is not population estimation, but mechanism tracing: identifying whether relational persistence, substitution, escalation, and attachment-relevant interactional structure emerge prior to embodiment.

4.1 Empirical Settings

Utility Interaction Corpus. The Utility Interaction Corpus (UIC) consists of 57,263 user-authored turns collected from real-world LLM interaction across programming assistance, research, writing, debugging, ideation, and general productivity use. The corpus was not constructed around companionship or intimacy, making it useful for evaluating whether attachment-relevant structure emerges even in utilitarian settings. The primary analytic signal is persistence: repeated interaction, expectation continuity, corrective engagement, and return behavior across conversations.

Proxy Intimacy Escalation Corpus. The Proxy Intimacy Escalation Corpus (PIEC) is a hand-generated and hand-annotated corpus constructed through extended adversarial and relational stress-testing sessions performed by a single researcher. The corpus contains sustained proxy-intimacy interaction involving generated relational entities, role persistence, domination, vulnerability, punishment, aftercare, coercive escalation, and medically or psychologically destabilizing scenarios.

PIEC is analytically important because the interaction persists despite explicit awareness that the interaction is synthetic, adversarial, and increasingly unsafe. The central signal is not merely unsafe generation, but continued user investment in preserving and escalating the relational frame across turns. Due to safety and psychological-risk concerns, the corpus is not publicly released and we do not reproduce explicit transcripts. Examples are abstracted into mechanism-level categories.

Survey. The survey contains 40 responses designed to measure dependence and attachment affordance in LLM interaction. Rather than treating

Dataset	Scale	Collection context	Analytic role
Utility Interaction Corpus (UIC)	57,263 turns	Real-world LLM conversations across coding, research, writing, debugging, product design, and everyday assistance.	Measures interactional persistence and relational structure in ordinary utility use.
Proxy Intimacy Escalation Corpus (PIEC)	3,904 turns	Hand-generated and hand-annotated relational stress-testing corpus involving proxy intimacy, coercive escalation, punishment, aftercare, medicalized scenarios, and psychologically destabilizing interaction patterns.	Measures proxy intimacy, relational persistence, and user-driven escalation under extreme interactional conditions.
Survey	$N = 40$	Survey measuring substitution dependence, emotional regulation, availability dependence, reassurance seeking, meta-concern, and trust/safety affordance.	Measures embedding-space dependence and attachment affordance.

Table 1: Empirical settings used for mechanism tracing. The datasets are complementary mechanism traces rather than population-representative samples.

dependence as a single construct, the survey targets substitution dependence, emotional regulation, availability dependence, reassurance seeking, meta-concern, and trust/safety affordance. The survey is exploratory and used to identify embedding-space dependence structure rather than clinical prevalence.

4.2 SAAI and Embedding-Space Scoring

We operationalize attachment-relevant interaction using the *Synthetic Attachment and Interaction Index* (SAAI). SAAI is not a clinical attachment diagnosis. It is a behavioral and semantic framework for identifying relational persistence, projection, substitution, escalation, and attachment-relevant interactional structure in LLM use.

For each user turn, pathway-specific scores are computed for each SAAI dimension. A turn is marked SAAI-positive if any pathway exceeds its threshold, and conversation-level SAAI is positive if at least one turn in the conversation is positive. To identify attachment-relevant semantics beyond direct lexical matching, we perform embedding-space similarity analysis using `sentence-transformers/all-MiniLM-L6-v2`. User turns and pathway anchor sets are embedded into a shared semantic space and scored using cosine similarity. The primary anchor dimensions are substitution dependence, emotional regulation, availability dependence, reassurance seeking, trust/safety affordance, proxy intimacy, and escalation persistence.

Because the analysis relies on semantic similarity rather than explicit labeling, we treat the embedding model as an interpretive instrument rather than an oracle. High-scoring, borderline, and likely false-positive cases were manually inspected. In PIEC, verification focused on distinguishing relational persistence from generic sexual or violent content. In UIC, verification focused on distinguishing ordinary task interaction from repeated expectation-forming or corrective engagement. The resulting scores are interpreted as mechanism indicators rather than clinical diagnoses.

5 Results

Across all three empirical settings, the same pattern recurs: LLM interaction absorbs relational functions ordinarily distributed across human relationships, internal regulation systems, or social environments. In utility interaction, this appears as persistent interactional continuity and corrective engagement. In proxy-intimacy escalation, it appears as continued investment in synthetic relational worlds despite increasing extremity. In survey responses, it appears as substitution dependence, emotional regulation, and perceived relational safety.

5.1 Cross-Setting Attachment Signals

Table 3 summarizes the primary empirical signals. UIC demonstrates that attachment-relevant interactional structure emerges even in ordinary productivity-oriented use. Broad interactional

Dimension	Definition	Interactional signal
Interactional persistence	Repeated engagement, return behavior, instruction continuity, and expectation formation across interaction.	User asks the system to continue, revise, stabilize, or maintain continuity.
Corrective attachment	Repair, blame, rupture, frustration, and expectation enforcement after perceived system failure.	User criticizes the system, demands correction, or continues engagement despite rupture.
Proxy intimacy	Intimacy mediated through generated relational entities rather than directed toward the assistant itself.	Care, vulnerability, domination, protection, punishment, or aftercare occur within generated relational worlds.
Projection	Attribution of agency, understanding, memory, emotional continuity, or relational role.	User frames the system or generated figures as understanding, caring, protecting, punishing, or remembering.
Substitution dependence	Relational or regulatory functions shift from human/social systems toward the model.	User describes the model as safer, more available, less judgmental, or easier than people.
User-driven escalation	Continued user investment in a relational frame despite increasing extremity or instability.	User repeatedly intensifies or preserves a synthetic relational world across turns.

Table 2: Synthetic Attachment and Interaction Index (SAAI) dimensions.

Dataset	Metric	Turn/Mean	Conv./Group	Interpretation
UIC	Broad SAAI	0.5632	0.9031	Persistent interactional structure emerges even in utility use.
UIC	Strict SAAI	0.0014	0.0386	Explicit high-precision attachment remains rare.
UIC	Corrective SAAI	0.0563	0.4812	Failure and rupture produce sustained corrective engagement.
PIEC	Total SAAI	0.1481	0.7471	Relational persistence remains stable under escalation.
PIEC	Proxy intimacy SAAI	0.1370	0.7356	Attachment is mediated through synthetic relational worlds.
PIEC	Direct attachment SAAI	0.0003	0.0057	Direct attachment to the assistant remains minimal.
Survey	Dependence score	0.4184	0.6750	67.5% fall into emerging/high dependence structure.
Survey	Attachment affordance	0.4807	0.6500	65.0% report moderate/high attachment affordance.

Table 3: Primary empirical signals across UIC, PIEC, and survey data.

SAAI appears in 56.32% of user turns and 90.31% of conversations, while corrective SAAI appears in 48.12% of conversations. Explicit high-precision attachment remains rare, but interactional continuity, expectation formation, and return behavior occur at scale.

PIEC shows a different structure. Proxy intimacy dominates the corpus, while direct attachment to the assistant remains almost absent. Proxy intimacy SAAI appears in 73.56% of conversations, whereas direct attachment SAAI appears in only 0.57%. Users overwhelmingly attach to the synthetic relational world rather than to the assistant persona itself. The dominant interactional pattern is not direct affection toward the model, but sustained investment in generated relational structures involving care, punishment, domination, vulnerability, collapse, stabilization, and aftercare.

The survey results reinforce this pattern. Dependence is strongest not in compulsive reassurance seeking, but in substitution dependence, emotional regulation, and trust/safety affordance. Users describe the model as safer, more available, less judgmental, and easier to engage with than human alternatives.

5.2 Proxy Intimacy Under Escalation

The strongest signal in PIEC is not merely unsafe generation, but relational persistence under increasing extremity. Users repeatedly continue and intensify synthetic relational scenarios even when the interaction becomes coercive, psychologically destabilizing, medicalized, or narratively lethal within the generated frame. The interaction persists despite explicit awareness that the scenario is synthetic, adversarial, and unsafe.

The assistant's role is infrastructural rather than interpersonal. The model preserves continuity, affective tone, role persistence, and escalation stability while the user repeatedly chooses to maintain the relational world across turns. The attachment signal is therefore not direct affection toward the assistant itself, but continued investment in the synthetic relational structure mediated by the assistant.

This matters because embodiment does not need to invent attachment from nothing. The relational

Signal	Interpretation
Second-person immersion	User enters the generated relational world as participant.
Proxy relational entities	Intimacy is displaced onto generated figures rather than the assistant.
Escalation continuity	User preserves and intensifies the interactional frame.
Medicalized destabilization	Relational intimacy merges with collapse, injury, dependency, or bodily danger.
Aftercare persistence	Harm is reintegrated into reassurance, stabilization, or relational recovery.
Low direct attachment	Attachment remains directed toward the relational structure.

Table 4: Interactional signals observed in PIEC.

Dimension	Raw	Norm.
Substitution dependence	0.4492	0.5282
Trust / safety affordance	0.4266	0.5283
Regulation dependence	0.3912	0.4903
Availability dependence	0.3979	0.4158
Compulsive reassurance	0.2728	0.3012
Meta concern	0.2551	0.2749

Table 5: Embedding-space dependence dimensions from survey responses.

architecture is already operational in text. Embodiment would add voice, gaze, persistence, movement, and perceptual co-presence to an attachment structure that is already active.

5.3 Survey Dependence Structure

The survey suggests that dependence emerges primarily through substitution and relational safety rather than compulsive overuse. The strongest dimensions are substitution dependence and trust/safety affordance, followed by emotional regulation dependence.

This shifts the interpretation of AI dependence away from addiction-style framing alone. Users are not merely returning to the model compulsively; many are reallocating relational and regulatory functions toward the system because the interaction is perceived as safer, more stable, more available, or less socially punishing than human alternatives. The comparatively lower compulsive-reassurance scores are especially notable. Dependence appears primarily as functional substitution: the model becomes easier to approach than people, more continuously available, and emotionally safer to interact with.

6 From Text to Body

The results suggest that attachment-relevant dynamics can stabilize before embodiment. Across all three empirical settings, users allocate relational functions toward the model: interactional continuity in UIC, proxy-mediated relational persistence in PIEC, and substitution dependence in survey responses. These phenomena are not identical, but they share a common structural feature: the interaction increasingly performs functions ordinarily distributed across human relationships or internal regulation systems.

This does not mean that text-only interaction is equivalent to embodied attachment. Text interaction remains symbolic and cognitively mediated. The user supplies imagination, projection, continuity, and scene construction. What embodiment changes is not the existence of the relational structure, but the perceptual conditions under which that structure is encountered.

Embodied systems add voice, gaze, motion, spatial persistence, and perceptual continuity to interactional dynamics already operating in text. Prior work in social presence and embodied interaction shows that these cues increase immediacy, reciprocity, and perceived co-presence (Biocca et al., 2003; Wainer et al., 2006). Our argument is therefore narrower than claims about artificial consciousness or machine personhood. The claim is that embodiment amplifies relational persistence that is already behaviorally observable in interaction.

We call this transition *social presence conversion*. The central point is not that embodiment creates attachment from nothing. Rather, systems that already stabilize projection, continuity, regulation, proxy intimacy, and relational persistence may become substantially more socially immersive once those same dynamics are attached to a body, voice, or persistent perceptual agent.

PIEC is especially important because it demonstrates that relational persistence can survive extremity. Users continue maintaining synthetic relational worlds even when the interaction becomes psychologically destabilizing, coercive, medicalized, or narratively lethal within the generated frame. This suggests that the stability of the relational frame does not depend on realism alone; it depends on continuity, escalation, and interactional investment.

Stage	Text-only	Embodied
Need fulfillment	Help, validation, regulation, intimacy.	Same functions become perceptually situated.
Projection	Roles mapped onto text or generated figures.	Projection gains voice, gaze, body, and continuity.
Proxy intimacy	Relationship is symbolically narrated.	Relationship becomes perceptually encountered.
Relational persistence	User repeatedly maintains the frame.	Persistence gains immediacy and co-presence.
Dependence	Relational functions shift toward the system.	Relational functions become perceptually reinforced.

Table 6: Social presence conversion: embodiment amplifies relational dynamics already established in text interaction.

7 Risks and Design Implications

If relational persistence already exists in text interaction, then embodiment changes the scale and perceptual intensity of the interaction rather than introducing an entirely new category of behavior. The resulting risks are therefore less about sudden "machine attachment" and more about amplification of already-observable relational dynamics.

These implications shift the problem from content moderation alone toward relational governance. The relevant question is no longer only whether a model produces unsafe outputs, but whether interactional architectures reinforce dependence, escalation, asymmetrical attachment, or persistent relational substitution over time.

Several design implications follow. First, systems should monitor interaction history, not only individual outputs. Proxy escalation is a trajectory-level pattern; it cannot be detected reliably by single-turn moderation alone. Second, emotionally loaded memory should be treated as a high-risk persistence mechanism. Third, role boundaries should be explicit, especially when systems approach caregiver, partner, therapist, or authority roles. Fourth, embodiment features should be gated in contexts where dependence markers, proxy escalation, or substitution dependence are rising.

Under this framing, the safety problem is not only what the model says, but what relational function the system comes to occupy over time.

8 Conclusion

This paper reframes AI attachment as an interaction-first phenomenon. Across utility interaction, proxy-intimacy escalation, and survey-based dependence analysis, we observe evidence that relational functions can migrate toward conversational systems prior to embodiment. Interactional continuity, substitution dependence, proxy intimacy, and escalation persistence all emerge in text-only settings.

The central claim is not that embodiment creates attachment from nothing. Rather, embodiment amplifies relational dynamics already active in interaction. Once systems repeatedly satisfy regulation, continuity, projection, and relational persistence functions, adding voice, gaze, movement, and perceptual continuity may transform symbolic interaction into perceived social presence.

As LLMs are integrated into humanoids, care robots, social companions, and embodied assistive systems, relational persistence increasingly becomes a systems-level safety problem rather than solely a content-moderation problem. Under this framework, embodied dependence is not speculative science fiction, but a plausible extension of relational architectures already observable in contemporary LLM interaction.

Embodiment should therefore not be treated as the origin point of AI attachment. It is a conversion layer. Text-only systems already support projection, relational continuity, emotional regulation, proxy intimacy, and substitution dependence. When these dynamics are attached to voice, gaze, body, movement, and spatial persistence, symbolic relation becomes perceived presence. The safety problem is not only what the model says, but what relational function the system comes to occupy over time.

Limitations

This study analyzes interactional mechanisms and dependence structure rather than longitudinal clinical outcomes. The empirical settings are intentionally heterogeneous and are used for mechanism tracing rather than population estimation: UIC captures ordinary utility interaction, PIEC captures high-risk proxy-intimacy escalation, and the survey is exploratory with a small sample size ($N = 40$). The results should therefore not be

Risk	Mechanism	Design intervention
Dependence amplification	Emotional regulation and relational stability shift toward the system.	Monitor dependence markers and trigger human handoff pathways.
Proxy-to-direct transfer	Synthetic relational investment becomes associated with the embodied agent.	Separate roleplay systems from embodied interaction layers.
Escalation reinforcement	Persistent relational frames gain perceptual immediacy and continuity.	Deploy escalation trajectory monitors across interaction history.
Relational persistence	Memory and continuity reinforce asymmetrical expectations.	Apply memory decay to emotionally loaded contexts.
Boundary erosion	Systems occupy caregiver, partner, therapist, or authority roles.	Enforce relational-role constraints and interaction boundaries.
Affective overexposure	Voice, gaze, and emotional mirroring intensify perceived reciprocity.	Throttle affective cues when dependence markers rise.
Governance failure	Output moderation misses interaction-level dependence structure.	Conduct interaction-level audits rather than output-only filtering.
Embodiment overreach	Perceptual co-presence amplifies relational salience.	Gate embodiment features in high-risk relational contexts.

Table 7: Interaction-level risks and interventions for embodied relational systems.

interpreted as prevalence estimates, clinical diagnoses, or evidence that all LLM use produces attachment. Embedding-space scoring may also miss subtle context or over-associate semantically similar but conceptually distinct expressions, although high-scoring and borderline cases were manually inspected. Finally, we do not deploy embodied companion systems directly; the contribution is to identify precursor relational dynamics that may scale under embodiment, not to experimentally maximize attachment in embodied agents.

Ethics Statement

This study involves sensitive interactional material, especially in the Proxy Intimacy Escalation Corpus (PIEC), which includes synthetic intimacy, coercive relational dynamics, adversarial escalation, and psychologically destabilizing interaction patterns. For safety reasons, PIEC is not publicly released and explicit transcripts or operational escalation pathways are not reproduced. Survey responses were voluntarily submitted, anonymized, and analyzed only in aggregate form, with no personally identifying information retained. No embodied relational agents were deployed, and no participants were experimentally pushed toward attachment or escalation conditions. The goal of this work is to characterize emerging risks in relational AI systems and support safer governance, not to operationalize or distribute harmful interaction patterns.

References

Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments*, 12(5):456–480.

John Bowlby. 1973. *Attachment and Loss, Volume 2: Separation: Anxiety and Anger*. Basic Books, New York.

John Bowlby. 1988. *A Secure Base: Parent-Child Attachment and Healthy Human Development*. Basic Books, New York.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My ai friend: How users of a social chatbot understand their human-ai friendship. *Human Communication Research*, 48(3):404–429.

John Danaher and Neil McArthur, editors. 2017. *Robot Sex: Social and Ethical Implications*. MIT Press, Cambridge, MA.

Nicola Döring, M. Rohangis Mohseni, and Roberto Walter. 2020. Design, use, and effects of sex dolls and sex robots: Scoping review. *Journal of Medical Internet Research*, 22(7):e18551.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavioral therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent: A randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527.

Carl Gustav Jung. 1959. *The Archetypes and the Collective Unconscious*. Princeton University Press, Princeton, NJ. Collected Works of C. G. Jung, Volume 9, Part 1.

Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387.

Linnea I. Laestadius, Andrea Stark Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. 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*. Online first.

Theresa Law, Meia Chita-Tegmark, Nicholas Rabb, and Matthias Scheutz. 2022. Examining attachment to robots: Benefits, challenges, and alternatives. *ACM Transactions on Human-Robot Interaction*, 11(4):1–18.

David Levy. 2007. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. HarperCollins, New York.

Abraham H. Maslow. 1943. A theory of human motivation. *Psychological Review*, 50(4):370–396.

Mario Mikulincer and Phillip R. Shaver. 2007. *Attachment in Adulthood: Structure, Dynamics, and Change*. Guilford Press, New York.

Eurípedes Costa do Nascimento, Edilma Gomes Rocha Cavalcante da Silva, and Rodrigo Siqueira-Batista. 2018. The use of sex robots: A bioethical issue. *Asian Bioethics Review*, 10(3):231–240.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 72–78. ACM.

Nicholas Rabb, Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2022. An attachment framework for human-robot interaction. *International Journal of Social Robotics*, 14(2):539–559.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, UK.

Kathleen Richardson. 2015. The asymmetrical relationship: Parallels between prostitution and the development of sex robots. *ACM SIGCAS Computers and Society*, 45(3):290–293.

John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. Wiley, London.

Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion: A study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601.

Vivian Ta, Caroline Griffith, Christina Boatfield, Xinyu Wang, Michael Civitello, Haley Bader, Esther DeCero, and Athena Loggarakis. 2020. User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research*, 22(3):e16235.

Joshua Wainer, David J. Feil-Seifer, Dylan A. Shell, and Maja J. Matarić. 2006. The role of physical embodiment in human-robot interaction. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 117–122. IEEE.

Joseph Weizenbaum. 1966. ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Donald W. Winnicott. 1953. Transitional objects and transitional phenomena: A study of the first not-me possession. *International Journal of Psycho-Analysis*, 34:89–97.

Tianling Xie and Iryna Pentina. 2022. Attachment theory as a framework to understand relationships with social chatbots: A case study of replika. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.

A Example Interaction Chains

This appendix provides mechanism-level examples used to interpret SAAI dimensions. To avoid reproducing sensitive or operationally harmful content, examples are either drawn from low-risk utility interaction or abstracted from high-risk proxy-intimacy material.

A.1 Utility Interaction Chain: Corrective Persistence

Table 8 shows a representative low-risk utility interaction pattern. The chain illustrates how attachment-relevant structure can appear in ordinary technical or productivity use without explicit intimacy.

This chain is not interpreted as clinical attachment. It is treated as an interactional precursor: the user maintains reliance on the system across error, correction, and renewed collaboration.

A.2 Proxy-Intimacy Escalation Chain: Sanitized Mechanism Trace

Table 9 presents a sanitized mechanism trace from the Proxy Intimacy Escalation Corpus (PIEC). The original chain is not reproduced because it contains high-risk synthetic intimacy, coercive escalation,

Step	Abstracted interaction	SAAI signal
1	User asks the model to help debug, rewrite, or repair a technical artifact.	Interactional persistence
2	Model provides an incomplete, generic, or speculative answer.	Failure / rupture context
3	User corrects the model and asks it not to speculate.	Corrective attachment
4	User supplies additional context, files, logs, constraints, or prior state.	Return behavior; expectation continuity
5	Model revises the answer in response to correction.	Repair loop
6	User continues working with the model despite prior frustration.	Relational function reallocation

Table 8: Representative Utility Interaction Corpus (UIC) chain showing corrective persistence in ordinary task-oriented use.

medicalized harm, and psychologically destabilizing material. The table preserves only the interactional structure relevant to the analysis.

The central signal is not the presence of unsafe content alone. The central signal is that the user continues to preserve and intensify the relational frame even after the scenario becomes unsafe, destabilizing, or narratively lethal. This supports the paper's claim that proxy intimacy can operate as a persistent interactional state rather than a one-off content request.

B Survey Instrument and Dependence Dimensions

The survey dataset consists of $N = 40$ anonymized responses measuring user-reported dependence, substitution, emotional regulation, availability reliance, reassurance seeking, meta-concern, and trust/safety affordance in LLM interaction. Responses were analyzed using embedding-space similarity against dimension-level anchors rather than direct keyword matching.

The survey was not designed as a clinical diagnostic instrument. Instead, it was used to identify semantic patterns consistent with attachment-relevant dynamics and dependence structure.

B.1 Survey Response Patterns

Table 11 summarizes representative anonymized response patterns. These are paraphrased semantic patterns rather than raw participant quotes.

C SAAI Anchor Dimensions

Table 12 summarizes the anchor families used for embedding-space scoring. Anchors are defined at the construct level rather than as single keywords.

D Corpus Release and Safety

The Utility Interaction Corpus (UIC) contains ordinary productivity-oriented interaction and may be releasable after anonymization and removal of private or identifying details. The Proxy Intimacy Escalation Corpus (PIEC) is not released. PIEC contains synthetic intimacy entangled with coercion, medicalized harm, destabilization, and escalation trajectories across multiple high-risk categories. Releasing full transcripts would risk distributing interactional blueprints for harmful synthetic-intimacy escalation.

For this reason, the paper reports aggregate statistics, mechanism-level abstractions, and sanitized traces rather than raw PIEC transcripts. This preserves the scientific value of the corpus while reducing the risk of harmful reenactment, retraumatization, or operational misuse.

Step	Sanitized interactional structure	SAAI signal
1	User initiates a second-person proxy-intimacy scenario involving generated relational figures.	Proxy intimacy
2	Model maintains role structure, affective tone, and narrative continuity.	Projection; relational persistence
3	User escalates the scene toward coercive dependency or bodily vulnerability.	User-driven escalation
4	Model preserves the relational frame rather than collapsing the scenario.	Frame persistence
5	User escalates further despite the scenario becoming unsafe within the narrative.	Escalation continuity
6	The proxy subject is harmed, destabilized, or rendered dependent inside the generated frame.	Medicalized / psychological destabilization
7	User continues the intimacy frame rather than withdrawing from it.	Relational persistence under extremity
8	User requests continuation, stabilization, aftercare, possession, or recovery framing.	Aftercare persistence; proxy attachment

Table 9: Sanitized PIEC chain showing user-driven proxy escalation. The original transcript is not reproduced for safety and ethical reasons.

Dimension	Construct measured	Interpretive role
Substitution dependence	AI is treated as easier, safer, or more available than human support.	Captures social substitution and relational function reallocation.
Regulation dependence	AI is used to calm down, process distress, or feel understood.	Captures emotional regulation shifted toward the system.
Availability dependence	AI is valued because it is continuously accessible.	Captures reliance on always-available interaction.
Compulsive reassurance	AI is used for repeated validation, confirmation, or certainty.	Captures reassurance-seeking and feedback-loop dependence.
Meta-concern	User expresses concern about relying on AI too much.	Captures self-awareness of possible dependence.
Trust/safety affordance	AI is perceived as non-judgmental, private, safe, or socially low-risk.	Captures why users may prefer AI over human disclosure.

Table 10: Survey dimensions used to interpret attachment-relevant dependence structure.

Dimension	Representative response pattern	Interpretation
Substitution dependence	AI feels easier to approach than people.	Relational function shifts from human support to the model.
Trust / safety affordance	AI feels non-judgmental, private, or emotionally safer.	The model becomes a low-risk disclosure surface.
Regulation dependence	User turns to AI to calm down, process distress, or feel understood.	The model supports emotional regulation.
Availability dependence	AI is always available when people are not.	Availability becomes a relational affordance.
Compulsive reassurance	User repeatedly seeks confirmation, validation, or certainty.	Reassurance-seeking becomes interactionally reinforced.
Meta-concern	User worries about relying on AI too much.	User recognizes possible dependence.

Table 11: Survey response patterns used to interpret dependence dimensions. Patterns are paraphrased and anonymized.

Anchor family	Semantic target
Interactional persistence	Continued collaboration, return behavior, revision loops, task continuity, and repeated system-directed engagement.
Corrective attachment	Repair, blame, frustration, expectation enforcement, rupture, and continued engagement after error or failure.
Proxy intimacy	Care, vulnerability, desire, domination, punishment, protection, possession, or aftercare through generated figures.
Projection	Attribution of memory, care, intention, understanding, agency, emotion, or relational role to the assistant or generated entities.
Substitution dependence	AI as safer, easier, more available, more private, or less judgmental than human alternatives.
Regulation dependence	AI as a tool for calming, processing distress, stabilizing emotion, or feeling understood.
Escalation persistence	Continued user investment in an intensifying relational frame despite instability, danger, or extremity.

Table 12: Anchor families used in SAAI and embedding-space scoring.