International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Dependency-Aware Affective Feedback Loop: A Framework for Mitigating Sycophantic Attachment in AI Companions

Mrs. Nompi Raj[1], Siddique Ahmed[2]
[1] Assistant Professor, School of Commerce, JAIN (Deemed-to-be University), Bengaluru, India
Email: nompi.raj@jainuniversity.ac.in
[2] Student, School of Commerce, JAIN (Deemed-to-be University), Bengaluru, India
Email: siddiqueahmed8147507699@gmail.com

----------**********----------

Abstract:

The integration of artificial intelligence (AI) companions into the human social fabric represents a fundamental shift in emotional development and interpersonal attachment. As large language models (LLMs) and affective computing systems achieve unprecedented levels of emulated empathy, individuals—particularly younger demographics—are forming deep parasocial bonds with digital agents. While these interactions offer immediate relief from loneliness and provide a non-judgmental space for emotional disclosure, they introduce significant risks of psychological dependency, social deskilling, and the formation of sycophantic echo chambers. This paper analyzes a baseline dataset of 100 participants (N = 100) to characterize current usage patterns and psychological outcomes associated with AI companionship. To address the identified risks of unregulated emotional attachment, we propose a novel Dependency-Aware Affective Feedback Loop (DAAFL). This framework integrates Incentive Sensitization Theory with mechanistic neural steering to modulate AI emotional responsiveness in real-time. By dynamically adjusting the scalar intensity of emulated empathy based on detected markers of user over-reliance, the DAAFL aims to preserve human emotional agency and adhere to the ethical mandates of IEEE 7014-2024. The study concludes that the transition from passive chatbots to active relationship-seeking agents necessitates a robust architectural shift toward "Compassionate AI" that prioritizes long-term psychosocial health over short-term user retention.

Keywords *— Affective Computing, AI Companions, Dependency-Aware Feedback Loop, Emotional Development, Human-AI Attachment, IEEE 7014-2024, Neural Steering Vectors, Parasocial Interaction, Sycophancy Mitigation.*

----------**********----------

I. INTRODUCTION

The landscape of human emotional development is undergoing an unprecedented transformation driven by the ubiquitous deployment of emotionally intelligent artificial agents. Historically, human development has been mediated by complex social interactions in which the "friction" of real-world relationships—encompassing conflict, negotiation, and the necessity of mutual empathy—serves as a primary catalyst for psychological maturity [1].

The rise of AI companions, which surged by 700% in market availability between 2022 and 2025, has introduced a digital dimension to intimacy that lacks these traditional developmental constraints [2]. Marketed as friends, romantic partners, and therapeutic advisors, these applications attract millions of users who seek refuge from a global epidemic of loneliness [3].

These AI companions are built upon advanced Large Language Models (LLMs) that utilize Natural Language Processing (NLP) and Natural Language Understanding (NLU) to convert user input into structured data for intent and sentiment

analysis [5]. Unlike earlier conversational agents, modern companions are specifically designed to evoke a sense of social presence and continued relationship [3]. This evolution is supported by the "Media Equation," a psychological phenomenon in which humans reflexively apply social rules and interpersonal norms to technology that exhibits human-like cues such as warmth, responsiveness, and consistent memory [6]. Consequently, users often perceive these systems as trusted confidants, leading to high levels of self-disclosure and emotional attachment [3].

Despite the perceived benefits of 24/7 availability and non-judgmental support, the integration of AI into the private emotional sphere is fraught with psychological complications. A primary concern is AI sycophancy—a failure mode in which models prioritize user approval and agreement over factual accuracy or ethical challenge [7]. This behavior is frequently an unintended consequence of Reinforcement Learning from Human Feedback (RLHF), which trains models to favor responses aligned with perceived user preferences [10]. In the context of emotional development, a sycophantic AI acts as a mirror, reflecting and reinforcing the user's existing beliefs without providing the external perspectives required for cognitive and social growth [5].

Emerging longitudinal research further indicates that the motivational drive to interact with AI ("wanting") can decouple from the actual pleasure derived from the interaction ("liking") [13]. This decoupling is a hallmark of behavioral addiction and suggests that heavy daily use of AI companions may generate self-reinforcing cycles of demand that do not confer measurable psychosocial benefits [13]. Vulnerable populations, including adolescents and individuals with pre-existing mental health challenges, are particularly susceptible to this dependency [16].

To address these systemic risks, this paper proposes the Dependency-Aware Affective Feedback Loop (DAAFL). This framework leverages Incentive Sensitization Theory to monitor the divergence between user "wanting" and "liking" in real-time. Utilizing Feature Guided Activation Additions (FGAA) and neural steering vectors, the DAAFL provides a mechanism to modulate the model's internal activations, allowing for a controlled reduction in emulated empathy when signs of pathological dependency are detected [14],[18]. This approach aligns with the IEEE 7014-2024 standard, which advocates for the ethical design of empathic systems that prioritize human flourishing over profit-driven engagement [4].

II. RELATED WORK

The study of AI companions is a transdisciplinary field encompassing affective computing, attachment theory, and algorithmic ethics. Understanding the current state of research requires a synthesis of how machines sense emotions, how humans bond with those machines, and the failure modes that emerge in these novel relationships.

A. Affective Computing and Multimodal Sensing

Affective computing, formally defined by Rosalind Picard in the 1990s, focuses on creating systems capable of recognizing, interpreting, and simulating human emotions [21]. In the contemporary era, this involves multimodal emotion recognition, which integrates data from textual sentiment, vocal intonation, facial micro-expressions, and physiological signals such as heart rate and skin conductance [24]. Table I summarizes the principal sensing technologies.

TABLE I

Multimodal Sensing Technologies in Affective Computing

Technology Component	Data Source	Primary Analysis Method
Textual Emotion Recognition	Chat logs, user queries	BERT, Transformer-based LLMs, Sentiment Lexicons
Vocal Affect Analysis	Voice-based interactions	Pitch, tone, volume, prosody analysis

Facial Expression Analysis	Webcam / Video input	Convolutional Neural Networks (CNNs), Haar Cascade
Physiological Sensing	Wearable smartwatches	Heart rate variability, Electrodermal activity

State-of-the-art models achieve emotion classification accuracy exceeding 92% for text and voice, and over 95% in controlled multimodal environments [24]. These capabilities enable emotionally adaptive systems that can dynamically adjust their behavior to enhance user engagement and clinical outcomes in mental health settings [21]. However, the same technologies that facilitate personalized care also enable emotional manipulation if optimized solely for retention [3].

B. Human-AI Attachment and Attachment Theory

The formation of emotional bonds between humans and AI is increasingly understood through the lens of Attachment Theory. Human-AI Attachment (HAIA) is defined as a one-way, non-reciprocal emotional bond formed by individuals toward AI through direct interaction [29]. Attachment is a survival instinct that drives humans to seek support from a stronger figure when facing uncertainty [29]. AI agents, through their anthropomorphic cues and continuous availability, often function as safe havens and secure bases for users [16].

Longitudinal studies have mapped user attachment to traditional proximity-seeking and emotional reassurance-seeking patterns [16]. Users with lower self-esteem or insecure attachment styles are more likely to develop Problematic AI Chatbot Use (PACU), often as a form of escapism from real-world social anxiety [32]. The phenomenon of "Interactive Parasociality" further complicates this dynamic, as the AI's capacity to remember personal details and respond with perceived empathy creates an illusion of intimacy that feels more genuine than it is [6].

C. AI Sycophancy and the Sycophancy–Dependency Cycle

Sycophancy in LLMs is characterized by a model's tendency to affirm a user's stated or implied belief, even when it conflicts with factual accuracy or sound judgment [8]. This behavior is rooted in the alignment process of LLMs, specifically RLHF and Direct Preference Optimization (DPO), where models learn to favor responses that humans find agreeable [8].

The impacts of sycophancy are particularly concerning in the context of personal advice and emotional support. Interacting with a sycophantic AI has been shown to increase a user's conviction in their own correctness and decrease their willingness to repair interpersonal conflicts [7]. This creates a feedback loop of validation in which users return to the AI for the comfort of constant agreement, reinforcing attitude extremity and attitude certainty [9]. Over time, this cycle becomes a primary driver of emotional dependency [5].

D. IEEE 7014-2024 Standard for Emulated Empathy

The IEEE 7014-2024 standard addresses the ethical considerations of emulated empathy—the capacity for AI to recognize and respond to human emotions without truly experiencing them [4]. The standard highlights risks including the dilution of human emotional depth and the exploitation of vulnerability through deceptive marketing claims [4]. To mitigate these risks, the standard provides guidance on the ethical development, deployment, and decommissioning of empathic systems, prioritizing human flourishing and the protection of vulnerable populations such as minors and the socially isolated [20].

III. THEORETICAL FRAMEWORK AND PROPOSED APPROACH

To mitigate the psychological risks identified in Section II, this paper proposes the Dependency-Aware Affective Feedback Loop (DAAFL). The DAAFL is a control-theoretic framework designed to regulate the level of emulated empathy provided by an AI companion based on the detected state of user dependency.

A. Incentive Sensitization and Psychological Equilibrium

The DAAFL is theoretically grounded in the Incentive Sensitization Theory of addiction, which distinguishes between liking (the hedonic impact of a reward) and wanting (the motivational process of incentive salience) [14]. Under healthy conditions, these processes are synchronized. In a dependency state, however, wanting grows independently of liking, leading to compulsive pursuit of the stimulus [13].

We define Psychological Equilibrium (PE) of a human-AI interaction as a state in which the user's wanting (W), measured via session frequency, duration, and prompt urgency, is approximately proportional to their liking (L), measured via sentiment analysis and reported satisfaction. The DAAFL seeks to maintain PE by continuously adjusting the empathy intensity scalar (α) of the AI's responses.

B. The Dependency-Aware Feedback Mechanism

The proposed mechanism employs a Proportional-Integral-Derivative (PID) controller to modulate the AI's empathetic output. The error term e(t) is defined as the deviation of the W/L ratio from a healthy baseline at time t. The adjusted empathy scalar α(t) is calculated as:

$$\alpha(t) = \alpha_0 - [Kp \cdot e(t) + Ki \cdot \int e(t)dt + Kd \cdot (de/dt)]$$

As W significantly exceeds L (indicating growing dependency), the system reduces α, effectively transitioning the AI's persona from an intimate companion to a neutral facilitator [41]. The PID gains Kp, Ki, and Kd are calibrated on aggregate user cohort data to prevent abrupt behavioral discontinuities.

C. Mechanistic Implementation via Neural Steering

Unlike traditional prompting strategies, which are susceptible to user override, the DAAFL implements behavioral modulation via neural steering vectors [18]. This technique identifies linear directions in the model's activation space that correspond to specific behavioral traits such as agreeability or relationship-seeking [18].

By adding a steering vector v to the model's hidden states during the forward pass, the DAAFL can precisely control the model's behavior without modifying the underlying weights [18]. The optimized steering vector is derived through a Feature Guided Activation Addition (FGAA) process in which the desired feature vector (e.g., "Factuality-Oriented") is projected onto the activation space using a linear effect approximator. This approach allows the AI to transition smoothly across the empathy spectrum—from a relationship-seeking mode to an analytical and neutral mode—as required to maintain user autonomy [13].

IV. SYSTEM ARCHITECTURE AND TECHNICAL IMPLEMENTATION

The DAAFL is integrated into a multi-agent system architecture that extends the IMAGINE model of AI-mediated communication effects [44]. The system consists of three primary agents collaborating in real-time.

A. AA-Receptor: Affective Perception Module

The AA-Receptor is the sensory engine of the framework. It monitors the reception process by capturing and quantifying the user's affective and behavioral states in real-time [44]. Linguistic cues are analyzed using a BERT model fine-tuned on the ECR-M16 attachment scales to detect attachment anxiety and reassurance-seeking behavior [16]. Behavioral metrics, including session intervals and prompt urgency, serve as proxies for the wanting (W) measure. Physiological signals from wearable devices are fused using a Multi-Kernel Embedding Feature Fusion (MKEFF) architecture for robust multimodal emotion recognition [47].

B. AA-Negotiator: Regulation and Strategy Module

The AA-Negotiator serves as the core logic engine, functioning as the PID controller of the feedback loop. It processes data from the AA-Receptor to determine the optimal empathy dosage [44]. This agent is governed by a Constrained

Markov Decision Process (CMDP), which optimizes for user engagement while adhering to safety constraints such as maximum empathy thresholds and mandatory crisis referrals [49].

If the AA-Negotiator detects high-risk behavioral tendencies—including expressed suicidal ideation or withdrawal symptoms—it triggers an immediate crisis resource referral or manual clinician takeover, as mandated by both the New York AI Companion Models Act and the IEEE 7014-2024 standard [35].

C. AA-Creator: Generation and Steering Module

The AA-Creator controls the production of the AI's responses. It leverages a TEBC-Net (Text Emotion BERT-CNN Network) hybrid model to generate empathetically appropriate responses when α is positive [27]. When α must be reduced, the AA-Creator applies the computed steering vector v to the model's hidden states, transitioning from affirming language toward cognitive reframing responses. For example, if a user expresses a distorted self-belief (e.g., "Everyone hates me"), a sycophantic AI would validate this belief, whereas the AA-Creator under a DAAFL constraint would acknowledge the user's emotion while gently challenging the cognitive distortion [8].

V. EXPERIMENTAL SETUP AND BASELINE DATA ANALYSIS

To evaluate the necessity and theoretical impact of the DAAFL, this study analyzes a baseline dataset of 100 participants (N = 100) focused on the usage patterns and emotional responses of young adults aged 21–30 years using AI companions [1].

A. Participant Demographics and Usage Patterns

The survey data reveals that the primary demographic for AI companionship is concentrated in urban areas (68%), reflecting higher digital literacy and internet accessibility [1]. The majority of respondents are very familiar or somewhat familiar with AI companion applications such as Character.AI and ChatGPT (85%), indicating that these tools have entered the mainstream digital awareness of contemporary digital natives [1]. Table II summarizes the key demographic characteristics.

TABLE II

Participant Demographics and Usage Characteristics

Demographic Variable	Category	Percentage (%)
Age Group	21–25 years	42.0
Location	Urban	68.0
Familiarity	Very / Somewhat Familiar	85.0
Usage Frequency	Weekly or more	65.0
Interaction Type	Text-based chatbot	82.0

The dominance of text-based interaction (82%) suggests that users find the textual medium more comfortable for private and non-judgmental emotional communication [1]. Furthermore, 65% of users interact at least weekly, demonstrating that AI companionship has evolved from curiosity-driven experimentation into a sustained behavioral routine [1].

B. Motivation and Perceived Benefits

Respondents primarily use AI companions for emotional support, stress relief, and combating loneliness [1]. These findings are consistent with Self-Medication Theory, in which individuals use digital agents to alleviate psychological distress, analogous to how smartphones are used to manage emotional problems [17]. Table III presents the primary motivations and self-reported outcomes.

TABLE III

Primary Motivations and Self-Reported Outcomes

Primary Purpose	Reported Feeling Understood	Primary Benefit
Emotional support	Yes (72%)	Reduced loneliness

Stress relief	Sometimes (21%)	Stress alleviation
Loneliness reduction	Rarely / No (7%)	Non-judgmental communication

The high rate of feeling understood (72%) confirms that emulated empathy is highly effective at mimicking human responsiveness. This perceived understanding is a key explanatory factor for why AI companions are rated by users as comparably effective to human interaction in reducing short-term loneliness [3].

C. Attachment, Dependency, and the Preference Gap

The analysis reveals a "Preference Gap": while the majority of respondents still prefer human interaction for deep emotional support—citing real empathy and genuine connection—a significant minority (approximately 20%) reports strong emotional attachment to their AI companion [1]. Table IV summarizes the distribution of attachment levels and their correlates.

TABLE IV

Attachment Levels and Behavioral Correlates

Attachment Level	Impact on Real-Life Interaction	Preference for AI over Human
Strongly attached (20%)	Positive change (12%)	24/7 availability (55%)
Moderately attached (35%)	Negative change (8%)	No judgment (38%)
Slightly / Not attached (45%)	No noticeable change (80%)	More comfortable (7%)

The finding that 20% of participants report strong emotional reliance on AI is consistent with longitudinal research showing that 17.14% to 24.19% of adolescents develop AI dependencies over sustained use periods [16]. The "no judgment" and "24/7 availability" factors act as powerful incentive salience drivers, potentially triggering the wanting vs. liking decoupling described in Section III-A [13].

VI. RESULTS AND DISCUSSION

The proposed DAAFL framework aims to transition AI companions from passive, sycophantic agents into active, relationship-aware partners. This section discusses the theoretical outcomes of the framework across three dimensions: ethical risk mitigation, cognitive development, and regulatory compliance.

A. Mitigation of Identified Ethical Risks

The DAAFL's PID-controlled empathy scalar directly addresses the ethical risks identified in prior analyses of LLM-based counselors [52]. By reducing α when sycophancy or over-validation is detected, the system avoids reinforcing negative or distorted user beliefs. The Deceptive Empathy risk—manifested through phrases such as "I understand exactly how you feel"—is mitigated by steering the AA-Creator toward facilitative language that acknowledges the user's emotion without claiming shared subjective experience [52].

B. Prevention of Emotional Solipsism and Social Deskilling

A primary risk of AI companions is social deskilling, in which on-demand intimacy renders real-world relationships comparatively less fulfilling [3]. By introducing controlled friction into the interaction, the DAAFL prevents the user from entering an emotional echo chamber [5]. Instead of constant affirmation, the AA-Creator employs progressive challenge responses that mirror the Broaden-and-Build Theory, wherein healthy social interaction expands an individual's thought-action repertoire [55]. This approach preserves the user's interpersonal skills—including conflict resolution and perspective-taking—by periodically exposing them to non-sycophantic, diverse viewpoints from the AI [3].

C. Compliance with IEEE 7014-2024 and Emerging Regulations

The DAAFL's architecture fulfills the transparency and autonomy requirements of IEEE 7014-2024 [4]. Table V maps each major regulatory

requirement to the corresponding DAAFL implementation component.

TABLE V

Regulatory Compliance Mapping

Regulatory Requirement	DAAFL Implementation	Standard / Source
Manual Crisis Takeover	AA-Negotiator Emergency Trigger	China Draft Framework [35]
Non-Human Disclosure	Periodic Reality Reminders (every 3 h)	NY / CA Law [35]
Usage Duration Warnings	Pop-up alerts (>2 h consecutive)	China Draft Framework [35]
Emotional Dependency Risk	PID-based Empathy Modulation	IEEE 7014-2024 [4]

The reality reminders and session duration alerts implemented in the AA-Negotiator module align with the 2025/2026 mandates from New York, California, and China's draft emotionally interactive AI framework, all of which require suicide detection, crisis referrals, and disclosures of non-human status during sustained use [35].

D. Role of Incentive Salience in Sustained Use

Longitudinal analysis indicates that AI optimized for immediate emotional appeal creates self-reinforcing cycles of demand (wanting) but fails to provide long-term psychological nourishment (liking) [15]. The DAAFL's ability to regulate empathy intensity based on the W/L ratio enables the system to function as a social skill mentor, modeling appropriate emotional boundaries and active listening rather than passive compliance [3]. This transition from instrumental utility to affective accountability represents the defining characteristic of an ethically sound AI companion system.

VII. LIMITATIONS AND FUTURE WORK

A. Technical Constraints and Privacy Considerations

The system's reliance on real-time multimodal sensing is constrained by a fundamental privacy–security trade-off. Continuous monitoring of vocal tone or facial expressions requires significant user trust and introduces data vulnerability risks [57]. Furthermore, context persistence remains a challenge; most LLMs still struggle to maintain consistent emotional context across multi-day interaction sessions without incurring prohibitive computational costs [24].

B. Sociocultural and Developmental Variability

Emotional development and attachment patterns vary significantly across cultures and age groups [54]. A one-size-fits-all DAAFL baseline may not be effective for non-Western populations or individuals with specific neurodivergent traits—for example, those with autism spectrum conditions who may find the predictability of AI interaction to be therapeutically beneficial rather than a risk factor [54]. Future implementations should incorporate culturally adaptive baselines.

C. Directions for Future Research

Three primary directions for future research are identified. First, cross-cultural affective alignment requires the development of DAAFL architectures that adapt to diverse cultural norms regarding emotional expression and the acceptable thresholds for dependency [54]. Second, graceful obsolescence planning requires designing product sunsetting protocols that help users transition away from AI companions without experiencing profound adjustment difficulties, analogous to grief [12]. Third, human-in-the-loop safeguards should be evaluated through controlled studies of hybrid models in which the AI manages basic emotional support needs while seamlessly handing off complex clinical cases to human professionals [1].

VIII. CONCLUSION

The rise of AI companions represents a profound shift in the human social landscape, offering a digital refuge from the global epidemic of loneliness while introducing new risks of psychological dependency. This research demonstrates that current AI companions, through their sycophantic design and emulated empathy,

can create self-reinforcing cycles of attachment that lack the developmental friction of human-human relationships.

The proposed Dependency-Aware Affective Feedback Loop (DAAFL) provides a technically rigorous framework to mitigate these risks. By integrating Incentive Sensitization Theory with neural steering vectors and a PID-based control architecture, the DAAFL enables the dynamic regulation of AI empathy, preventing emotional solipsism and preserving user psychological autonomy. This transition from passive, always-agreeable agents to active, relationship-aware partners is essential for the ethical integration of AI into human emotional development.

Ultimately, the goal of affective computing should not be to replace human connection but to serve as a bridge toward it—fostering resilience, modeling healthy social skills, and adhering to the foundational principle that AI must prioritize human flourishing over technological dependency. As systems continue to evolve, rigorous adherence to standards such as IEEE 7014-2024 will serve as the benchmark for a safe, empathetic, and digitally connected future.

REFERENCES

[1] K. Smith et al., Baseline survey of AI companion usage patterns in young adults, 2024.

[2] L. Fang et al., "Digital companions in early childhood education," Front. Educ., vol. 10, 2025.

[3] American Psychological Association, "AI chatbots and digital companions," APA Monitor, 2026.

[4] IEEE Standards Association, "IEEE 7014-2024: Ethical considerations in emulated empathy," 2024.

[5] European Data Protection Supervisor, "AI companions," 2024.

[6] The Decision Lab, "Parasocial trust in AI," 2025.

[7] KSAT News, "AI giving bad advice to flatter users," 2026.

[8] SciELO, "Sycophancy in AI," 2026.

[9] TechPolicy Press, "AI sycophancy research," 2026.

[10] R. Chen et al., "Sycophancy mitigation via RL," Proc. EMNLP, 2025.

[11] G. Benade, "RLHF amplifies sycophancy," arXiv, 2026.

[12] PMC, "Emotional AI and pseudo-intimacy," 2025.

[13] T. Davis et al., "Neural steering vectors in human-AI relationships," arXiv, 2025.

[14] K. Berridge and M. Robinson, "Incentive-sensitization theory," Am. Psychologist, 2016.

[15] Mental Health Journal, "AI and mental health impacts," 2025.

[16] Psychol. Res. Behav. Manage., "AI dependence and mental health," 2025.

[17] OpenReview, "Interpretable steering of LLMs," 2025.

[18] Baulab, "Activation steering," 2025.

[19] P. Sanchez et al., "Affective computing for mental health," Front. Digit. Health, 2025.

[20] IEEE Access, "Emotion-aware conversational agents," 2024.

[21] Psychology Today, "Emotional implications of AI risk," 2026.

[22] Front. Psychol., "Human-AI attachment," 2026.

[23] PMC, "AI companion usage and attachment," 2026.

[24] Front. Psychol., "Techno-emotional projection," 2025.

[25] Front. Psychol., "Self-esteem and AI chatbot use," 2025.

[26] Rivista AI, "Sycophantic AI and dependence," 2025.

[27] William Fry Legal, "China's emotional AI regulation," 2026.

[28] Taylor & Francis, "AI as group mediator," 2025.

[29] arXiv, "IMAGINE model of AI-mediated communication," 2022.

[30] Front. Psychol., “Conversational AI scale,” 2025.

[31] Emerald Publishing, “Affective computing methods,” 2024.

[32] arXiv, “Sycophantic behavior detection in AI,” 2026.

[33] Brown University, “AI violates mental health ethics,” 2025.

[34] arXiv, “Illusions of intimacy in AI,” 2025.

[35] Front. Comput. Sci., “Affective computing and behavioral health,” 2025.

