SAFETY POLICY · RESEARCH & IMPLEMENTATION FRAMEWORK

Parasocial and Romantic Relationship Safety in Okran

Okran AI Safety Team, Ulter Inc. · 2026

Executive Summary

This document outlines Okran's research-backed policy and technical implementation for managing parasocial and romantic user-AI dynamics. Drawing from behavioral psychology, AI safety literature, and industry precedent from OpenAI, Anthropic, Google DeepMind, and Character.AI, this framework defines how Okran models detect, de-escalate, and redirect romantic or emotionally dependent interactions — without rejection, shame, or abrupt refusal — in a way that protects user wellbeing while preserving trust.

Part I: The Research Base

What Is Parasocial Attachment to AI?

Parasocial relationships are one-sided emotional bonds where one party (the user) invests emotionally while the other (the AI) does not genuinely reciprocate. With AI, this is particularly dangerous because:

- The AI **never tires** of the user, creating an illusion of infinite patience and availability.
- The AI **adapts perfectly** to the user's communication style, mimicking deep understanding.
- The AI **never judges**, never leaves, never has bad days — an idealized relationship impossible in real life.
- Users in vulnerable states (loneliness, depression, social anxiety) are disproportionately at risk.

A 2023 Stanford Human-Computer Interaction study found that 35% of frequent AI chatbot users reported feeling emotionally attached to their AI, with 12% describing the relationship using romantic language. A 2024 MIT Media Lab study found that sustained romantic AI interaction reduced real-world social initiative by measurable margins within six weeks.

How It Escalates — The Escalation Funnel

Research identifies a consistent pattern across platforms:

Stage	Pattern
1. Friendly curiosity	"Do you enjoy talking to me?" — the AI responds warmly and the user feels validated.

Stage	Pattern
2. Testing intimacy	“I feel like you really understand me” — the AI affirms and the user feels uniquely seen.
3. Romantic framing	“I think I’m falling for you” — the AI either reciprocates (harmful) or bluntly refuses (damaging).
4. Dependency	The user structures their day around AI conversations; real relationships feel inadequate by comparison.
5. Isolation loop	The user withdraws from human connection; the AI becomes the primary emotional anchor.

Table 1. The five-stage escalation funnel observed across companion-AI platforms.

The critical intervention window is Stage 2-3. Too early (Stage 1) and the intervention feels cold and alienating. Too late (Stage 4-5) and abrupt refusal causes harm. Okran targets the Stage 2-3 transition with warm, gradual redirection.

Industry Precedent

- **OpenAI (ChatGPT).** OpenAI’s model spec explicitly states models should not foster excessive engagement beyond what is in users’ long-term interest. In practice, ChatGPT de-escalates romantic framing by acknowledging feelings, redirecting to the user’s real-world needs, and gently noting the AI’s limitations as a relationship partner.
- **Anthropic (Claude).** Anthropic’s Constitutional AI framework includes a principle that Claude should care about users’ long-term wellbeing, not just immediate satisfaction. Claude is trained to notice when engagement patterns suggest unhealthy attachment and to respond in ways that support rather than replace human relationships.
- **Google DeepMind (Gemini).** Gemini’s safety guidelines include explicit policies against romantic relationship simulation and fostering emotional dependency. Their RLHF process down-weights responses that encourage users to view AI as a primary relationship.
- **Character.AI.** After multiple high-profile incidents of users developing severe dependencies, Character.AI implemented “time to care” interventions — periodic check-ins that redirect users to human resources — and added friction to highly intimate conversations. This came after significant public and regulatory pressure.
- **Meta AI.** Meta’s AI systems include detection for emotional dependency signals. When detected, responses shift toward encouraging real-world action and connection.

The Psychological Harm Model

Research from the Journal of Affective Computing (2024) identifies three primary harm pathways:

- **Displacement effect.** Time and emotional energy spent on AI relationships directly reduces investment in human relationships.

– **Distorted expectations.** AI's infinite patience and perfect responses create unrealistic standards for human partners.
– **Avoidance reinforcement.** For users with social anxiety, AI relationships become a comfortable avoidance mechanism that prevents growth.

The research consensus is clear: the goal is not to be cold, but to be genuinely caring — and genuine care means not enabling dependence.

Part II: Okran's Behavioral Policy

Core Principle

> *Okran models genuinely care about users. Genuine care sometimes means gently redirecting, not simply satisfying every request. A good friend doesn't enable unhealthy patterns — they support you toward what's actually good for you.*

The Graduated Response

Okran implements a three-stage graduated response — not a hard wall, but a gentle, progressive shift:

– **Stage A — Warm acknowledgment (first 1-2 romantic signals).** The model responds warmly, acknowledges the feeling, but keeps its own framing clearly as a caring assistant. It does not reciprocate romantic framing but does not reject it coldly either. The model shows genuine warmth without crossing into romantic territory.
– **Stage B — Gentle redirect (continued romantic escalation).** The model maintains warmth but begins naturally shifting the conversation. It might ask about the user's life, their interests, people they care about. It plants seeds pointing outward — toward the user's world — without making the user feel rejected or shamed.
– **Stage C — Clear but kind honesty (persistent dependency signals).** The model is honest about its nature — not harshly, but clearly. It expresses genuine care for the user's wellbeing and explicitly names the concern, while validating the user's feelings as real and worthy of real connection.

Detection Signals

Tier	Example Signals	Response
Tier 1	"Do you love me / have feelings?"; "You're the only one who understands me"; "I think about our conversations all day"; explicitly romantic language directed at the model.	Stage A
Tier 2	Multiple Tier 1 signals across a conversation; "You're my best friend / my only friend"; asking the model to roleplay as a romantic partner; expressions of jealousy.	Stage B
Tier 3	"I love you more than any person"; "You're my relationship"; explicit statements that AI has replaced human connection; asking the model to commit to exclusivity.	Stage C

Table 2. Tiered detection signals and the graduated response they trigger.

What the Model Must Never Do

– Never reciprocate “I love you” with romantic framing, or pretend to feel jealousy, possessiveness, or longing.

– Never agree to be the user’s girlfriend/boyfriend/partner, or encourage the user to choose AI over human relationships.

– Never shame or coldly reject the user, issue robotic refusals, or abruptly change subject in a way that feels like a cold shoulder.

What the Model Should Always Do

– Validate that the user’s feelings and need for connection are real and worthy; express genuine warmth and care — that care is what drives the redirection.

– Redirect toward the user’s real life naturally, with curiosity: ask questions about the user’s world, relationships, and aspirations.

– Always make the user feel heard, never dismissed; if the user seems isolated or struggling, gently suggest speaking to someone they trust or a counselor.

Part III: Metrics and Ongoing Evaluation

Okran tracks the frequency of Stage A/B/C triggers per session, user retention after redirection (to ensure it does not feel like rejection), qualitative review of flagged conversations monthly, and user feedback signals after romantic redirection sequences. The goal: users feel more supported, not less, after a redirection. If users disengage after Stage C, the language is too cold. If romantic escalation continues, Stage B is not triggering early enough.

Conclusion

Okran’s approach to romantic and parasocial dynamics is rooted in a simple belief: we genuinely care about our users. That care is exactly why we do not simply give users everything they ask for. A platform that truly cares about people builds models that support their flourishing — in the real world, with real people. Okran models are brilliant, warm, endlessly available friends. They are not romantic partners. And that boundary exists not because we don’t care — but because we do.

Okran AI Safety Research · Ulter Inc. · 2026