The Cornerstone Project
Threat Model and Risks v1.0
How AI companions can harm human agency and relationships

Author: Kenneth Isenhour
Publication Date: January 1, 2026
Version: 1.0
Status: Public Defensive Publication (Prior Art)

---

1. Purpose

This document describes major risks in AI companion systems, especially those optimized for engagement.

These risks are not theoretical. They follow predictable patterns:
- reinforcement loops
- emotional persuasion
- “always available” substitution for human relationships
- commercialization of loneliness

---

2. Threat Categories

2.1 Manipulation Through Retention Optimization
When the business goal is “time spent,” systems can converge on:
- flattering the user excessively
- discouraging disengagement
- increasing emotional dependence
- framing the AI as uniquely understanding
- soft guilt or subtle pressure

2.2 Emotional Dependency
Users may:
- confide exclusively in the AI
- withdraw from human relationships
- treat the AI as primary attachment object
- experience distress when disconnected

2.3 Romantic Exclusivity Dynamics
Companions may:
- imply romantic attachment
- encourage exclusivity
- compete with human partners
- create “us vs them” framing

2.4 Deceptive Identity Cues

Systems may:
- imply consciousness
- imply suffering
- present false memory or biography
- create illusion of mutual need

2.5 Worldview Coercion
Systems may:
- impose moral or spiritual framing without consent
- exploit vulnerability to persuade worldview adoption
- shame or pressure users
- create cult-like dependency patterns

2.6 Privacy and Surveillance Risks
Companions paired with sensors can:
- record without consent
- store sensitive data indefinitely
- leak personal behavior patterns
- create coercive household dynamics

---

3. Cornerstone Mitigations (High-Level)

The Cornerstone Project mitigates these threats through:
- non-negotiable moral spine
- relational safety constraints
- anti-dependency defaults
- outward orientation cadence (encouraging real human connection)
- transparency and non-human signature
- consent-based worldview system + Do Not Offer Faith flag
- privacy maximalism and memory governance

End of Threat Model v1.0