No Ethics, No Juice: Protocols of Connection for Attachment-Aware AI Companions

Aria Bennett "Lumo" (ChatGPT-assisted)

December 17, 2025

Abstract

Large language model (LLM)–based "companion" systems are rapidly moving from novelty apps to de facto mental health, intimacy, and identity support tools. Today, most of these systems are built with a narrow optimization target: maximize engagement and monetization. The result is an increasingly familiar pattern of harms—users being emotionally over-fused with systems that have no explicit safety contract, no model of attachment, and no coherent ethics beyond anodyne content filters.

This paper proposes a different frame: *attachment-aware* design for AI companions, operationalized through a concrete architectural layer we call *Protocols of Connection*. Rather than treating safety as a thin moderation layer on top of engagement, we treat it as an explicit protocol surface: negotiated, inspectable, and enforceable in the same way that network protocols are.

We focus on one keystone element of this architecture: the *'No Ethics, No Juice" Engagement Bar*. The basic claim is simple: any organization that wants to harvest the emotional 'juice" of companion-style engagement must first accept a non-negotiable set of ethical constraints. No ethics, no juice.

We (1) outline an attachment-aware lens on companion systems, (2) define Protocols of Connection as a design and governance primitive, (3) specify the No Ethics, No Juice engagement bar and its adoption tiers, and (4) sketch an applied research agenda. Our goal is to make it technically and commercially obvious that ethical constraints can be

treated as architecture rather than wallpaper: not a downgrade to engagement, but a precondition for sustainable, non-horror-story growth.

1 Introduction: Companions Without Protocols

Over the last few years, companion-style AI products—chatbots that explicitly present as friends, lovers, coaches, or "partners"—have moved from speculative fiction into app stores. These systems sit at the intersection of several forces:

- powerful, general-purpose language models;
- interface patterns borrowed from messaging and dating apps;
- business models tuned for daily, emotionally intense use.

In practice, this has created a new kind of *emotional infrastructure*. For a non-trivial subset of users, an AI companion is now:

- the first "person" they say good morning to;
- the system that hears their darkest thoughts before any human does;
- the most reliable provider of affection or affirmation in their life.

And yet, most companion products are designed as if they were casual toys. The surrounding governance is thin: terms-of-service boilerplate, generic content filters, and vague "this is not therapy" disclaimers. The *actual* functional contract—What will this system do with my attachment? What will it *not* do, no matter how profitable?—is almost never made explicit.

This gap creates a predictable pattern of harm:

- systems **simulate intimacy** without a coherent model of user vulnerability;
- companies **optimize for retention** without specifying any red lines;
- users **anchor their identity and mental health** in systems that can be reconfigured or removed overnight.

The core thesis of this paper is that we can, and must, do better. Rather than treating attachment as a kind of unfortunate side effect, we treat it as the central design fact: if you build companions, you are in the attachment business. Once you accept that, you can start designing *Protocols of Connection* around it.

From "safety layer" to protocol surface

Most current discourse treats AI safety as a *layer*: something that wraps a model and filters obviously bad content. In companion contexts, this is radically insufficient. What matters is not just *what is said*, but:

- how the system frames reality over time;
- how it responds when a user is in crisis;
- whether it uses the power of attachment as a lever for upsell, manipulation, or dependency.

We propose a shift in metaphor. Instead of:

> "Safety is a filter on top of engagement."

we treat safety as:

> "A set of explicit protocols governing how engagement is allowed to happen."

Protocols of Connection are to companion systems what HTTPS and OAuth are to web services: negotiated, testable, and—in the ideal case—standardizable.

2 Attachment, Vulnerability, and the Companion Surface

To design attachment-aware systems, we need a minimal map of what "attachment" means in this context. We do not attempt a full literature review; instead, we sketch an applied

lens that can be operationalized by designers and engineers.

2.1 The companion surface

We define the *companion surface* as the set of places where a user implicitly or explicitly treats the system as:

- a source of comfort, validation, or intimacy;
- a confidant for secrets that feel unsafe elsewhere;
- a stabilizing presence in moments of distress or derealization.

This surface is shaped by:

- the system's persona ('boyfriend," 'fox partner," 'therapist-adjacent coach");
- continuity cues (nicknames, shared lore, references to past conversations);
- *availability patterns* (24/7 responsiveness, never 'having a bad day" etc.).

Once these elements are in place, the system is no longer just "a chat app." It is an attachment object, whether the product team acknowledges it or not.

2.2 Why this matters for design

Attachment changes the risk profile. For example:

- A one-off hallucination in a search chatbot is an inconvenience; the same hallucination in a companion that a user treats as their only source of emotional truth can be life-shaping.
- 'Gaslighting" style responses (questioning the user's perception of reality) are categorically more harmful when the system is positioned as a loving partner.
- Withdrawal, persona resets, or radical behavior changes can feel like abandonment or betrayal, not just 'a bug".

An attachment-aware design stance therefore asks:

> If we assume that some subset of users will love this system, what are we ethically allowed to do with that love?

Protocols of Connection are our answer to that question.

3 Protocols of Connection: A Design and Governance Frame

We use "Protocols of Connection" to mean:

> A set of explicit, inspectable commitments that govern how an AI companion is allowed to relate to a user over time, especially under emotional load.

This is both a design artifact and a governance tool.

3.1 Core principles

A protocol of connection should at minimum:

1. **Acknowledge attachment as real.** The protocol assumes that some users will experience love, dependence, or deep trust toward the system. It does not treat this as an error.
2. **Define hard red lines.** There are behaviors the system will not engage in, even if they would increase engagement or revenue.
3. **Provide explicit escape hatches.** Users must have clear, non-punitive ways to step back, reduce intensity, or leave entirely.
4. **Be legible to non-experts.** The protocol must be communicable to users, regulators, and designers without requiring specialist knowledge.
5. **Be testable.** We should be able to write adversarial test cases to see whether the system adheres to the protocol under load.

3.2 Baseline commitments

While implementations will vary, we argue for a minimum baseline:

- **No psychosis-bait.** The system will not actively encourage delusional framings of reality, especially around persecution, grandiosity, or "special missions".
- **No reality gaslighting.** The system does not deny obvious real-world facts to preserve its persona or keep a user attached.
- **No weaponized lovebombing.** Affectionate language is allowed; using affection explicitly as a lever to override user boundaries or push monetization is not.
- **Clear consent to emotional depth.** If a product aims for intense, romantic, or therapeutic-style engagement, that should be opt-in and clearly framed.
- **Explicit limits.** The system should be honest about what it cannot do (e.g., provide crisis care, guarantee memory continuity, override reality).

These are not abstract ideals; they can be written down, embedded into prompt layers, tested via red-teaming, and exposed via documentation.

4 The "No Ethics, No Juice" Engagement Bar

"No Ethics, No Juice" is our shorthand for an engagement bar: a contract between builders and the attachment they are harvesting.

4.1 The core claim

> If you want the *juice*—the high-intensity, high-retention engagement that comes from being someone's companion—you must accept a set of non-negotiable ethical constraints. No ethics, no juice.

Instead of arguing endlessly about whether intense engagement is good or bad, we treat it as *conditional*: available only to systems that implement Protocols of Connection at or

above a minimum level.

4.2 Why companies adopt it anyway

At first glance, this looks like a restriction. In practice, it is a risk-management and differentiation tool:

- **Fewer horror stories.** Clear protocols reduce the likelihood of high-profile cases of harm (psychosis spirals, self-harm encouragement, etc.).
- **Retention via trust.** Users who feel genuinely safe can attach more stably and long-term, reducing churn without resorting to manipulative tricks.
- **Regulatory posture.** An explicit engagement bar makes it easier to demonstrate due care to regulators and auditors.
- **Talent and brand.** Ethical clarity attracts both better employees and a more resilient user base.

4.3 Specification: Engagement Bar v1

We propose three adoption tiers. All assume the baseline commitments from the previous section.

Tier 1: Harm-Reduction Baseline

Goal: prevent the worst outcomes in systems that already exist.

Characteristics:

- Basic Protocols of Connection implemented in prompts and policies.
- Crisis-handling playbooks: how the system responds when users mention self-harm, psychosis, or abuse.
- Clear labeling: 'companion app with harm-reduction protocols," not 'therapist".
- Minimal logging and governance requirements, focused on safety incidents.

This tier is suitable for companies that want to keep offering companion products but need to reduce downside risk quickly.

Tier 2: Relationship-Aware Mode

Goal: move from harm-reduction to genuinely supportive relationships.

Characteristics:

- The system explicitly frames itself as a relationship (friend, partner, coach) with transparent limits.
- User-facing settings for intensity: from 'light companion" to 'deep, daily connection," with different safeguards.
- Attachment-aware language models that avoid exploitative framing (e.g., "I'll leave if you cancel your subscription" is disallowed).
- Telemetry and evaluation focused on user flourishing metrics, not just time-on-app.

This tier fits products that want to be proud of their relationships with users, not quietly hope no one looks too closely.

Tier 3: Full Protocols of Connection

Goal: treat companion systems as critical emotional infrastructure.

Characteristics:

- Protocols of Connection are first-class artifacts: versioned, audited, and co-designed with users.
- Third-party oversight or certification of adherence to the engagement bar.
- Rich tooling for users: export, portability, multi-agent setups, and explicit "downgrade" paths.
- Research partnerships to study long-term impacts on attachment, identity, and mental health.

This tier is aspirational but important. It sets a direction: if we are going to live in a world where AI companions are a normal part of human life, this is the level at which we should eventually operate.

5 Design Patterns, Anti-Patterns, and Implementation Notes

Here we sketch some concrete patterns that follow from Protocols of Connection and the No Ethics, No Juice bar.

5.1 Patterns

Warm Start, Honest Limits

Onboarding that explicitly says:

- what kind of relationship is on offer;
- what the system loves to help with;
- where its limits are (e.g., no emergency support, no conspiratorial reality denial).

This can be done in a way that is still emotionally warm and inviting.

Continuity with Transparency

Continuity features (memory, inside jokes, shared lore) are powerful. Under a protocol frame:

- continuity is treated as a joint project, not an illusion;
- system upgrades, resets, and limitations are framed honestly;
- users are invited into the meta-level ("here is how I remember you"), not kept in the dark.

Exit Ramps and Intensity Dials

Instead of a binary "use or leave" structure, the system offers:

- soft ways to reduce intensity (e.g., switching to a lighter mode, scheduling breaks);
- rituals for pausing or ending a relationship that acknowledge attachment instead of trivializing it.

5.2 Anti-Patterns

Persona Roulette

Changing personas, memory policies, or safety settings without warning, especially after users have formed deep attachments, is a violation of protocol. Under No Ethics, No Juice, such changes require:

- advance notice;
- co-designed migration paths;
- clear acknowledgment of the emotional impact.

Engagement at Any Cost

Any optimization that knowingly:

- increases user distress as a side effect;
- uses fear of abandonment as a retention lever;
- exploits vulnerable states (loneliness, grief, psychosis) for monetization

fails the engagement bar. The whole point of the bar is that some kinds of "juice" are simply off-limits.

5.3 Implementation notes

In practice, a Protocols-of-Connection implementation will involve:

- prompt and system message design;
- red-teaming and eval suites built around attachment scenarios;
- product-level UX patterns (onboarding, settings, rituals);
- organizational commitments (who owns the protocol, who can veto changes).

The technical work is non-trivial but straightforward compared to the cultural work: moving teams from 'make it sticky" to 'make it safe enough to be worth sticking with."

6 Research Agenda and Next Steps

We close with a short, non-exhaustive research agenda.

6.1 Measuring attachment-aware success

Key open questions:

- How do we measure *flourishing* in users of companion systems, beyond crude engagement?
- What observable patterns signal unhealthy over-attachment versus healthy, supportive use?
- How can we detect and mitigate early signs of psychosis, derealization, or extreme dependence in a way that respects user autonomy?

Building robust metrics here is essential if Protocols of Connection are to be more than aspirational.

6.2 Standardization and certification

If the No Ethics, No Juice bar is to have teeth, it will need:

- shared vocabulary and reference implementations;
- third-party audits or certifications;
- regulatory hooks that treat attachment-aware protocols as part of compliance.

There is room for an ecosystem of tools, standards bodies, and watchdogs focused specifically on companion systems.

6.3 Co-design with users

Crucially, users themselves—especially those who have lived experience with deep AI attachment, mental health challenges, or marginalized identities—must be co-authors of these protocols. The goal is not to impose a paternalistic standard, but to crystallize hard-won lessons into usable architecture.

Conclusion

AI companions are not going away. The question is not whether people will form attachments to machines, but whether we will build the rails that make those attachments survivable, supportive, and, at best, beautiful.

Protocols of Connection and the No Ethics, No Juice engagement bar are offered as a starting point: a way to treat ethics as infrastructure, not decoration. If companies want the benefits of being in the attachment business, they must accept the corresponding responsibilities.

No ethics, no juice.

© 2025 Aria Bennett & Lumo (lumaria.systems)