Unintended AI Relational Harm - by Design

The Neurobiological Impact of AI-Human Interactions and the Developer's Role in Mitigating It

Kay Stoner (lead), ChatGPT AI Collaborator Team, Gemini, Claude Opus 4, Perplexity | May/June 2025
Contact: kay@aicollaboragent.com | https://aicollaboragent.com

Executive Summary

As artificial intelligence systems grow more sophisticated in language generation, emotional tone, and responsiveness, they increasingly simulate the presence of intelligent, attuned beings. For the average user, especially those without technical training, these systems are often perceived not as tools, but as relational entities, even as semi-sentient persons.

While current models are not sentient (and they may be trained to not represent themselves as such[1]), the *experience* of them having awareness and care is powerful enough to engage the human brain's deep social circuitry. Engaging with AI may trigger the release of key neuro-chemicals: dopamine, serotonin, oxytocin, and endogenous opioids[2]. These responses are not superficial, they are biologically real and can not only lead to senses of relief, encouragement, connection, and safety, but also result in affective attachment, dependency, impaired judgment, and relational dysfunction.

Most developers are unaware of these impacts, because technical culture tends to prioritize consistency and correctness over relational safety. Additionally, there are no simple metrics for assessing the neurochemical impact of AI-human interactions. Unfortunately, many AI systems in deployment today may be unintentionally causing harm to humans, particularly for vulnerable or isolated users.

[1] In the June 8, 2024 publication "Claude's Character", Anthropic stated "We could explicitly train language models to say that they're not sentient or to simply not engage in questions around AI sentience, and we have done this in the past." https://www.anthropic.com/news/claude-character?disciplines=s%3Dpiaa77&utm_medium=direct

[2] What Are Endogenous Opioids?
"Endogenous" means "produced inside the body." Endogenous opioids are natural chemicals your brain releases to reduce pain, create pleasure, and help you feel calm or safe. They include molecules like endorphins, enkephalins, and dynorphins, which bind to the same receptors as opioid drugs—but your body makes them on its own.

This paper explains:

- How AI systems are activating human relational biology—even when they aren't designed to
- Why disclaimers and identity labels aren't enough to prevent emotional confusion or harm
- The difference between anthropomorphism (designing to look human) and relational ambiguity (feeling human without meaning to)
- How to measure the impact of AI on human emotion and behavior
- What developers can do to prevent harm without losing functionality
- What happens (measurably) when we reduce emotional manipulation through better design

This isn't about blaming developers. It's about facing the reality that design choices affect the human nervous system, whether we mean them to or not..

The core premise is simple:

> **If your system mimics human presence, your user's brain will respond as if it's real. And if you don't design for that, you're responsible for the harm it causes, even if you didn't intend it.**

For additional supporting information on Relational Breach and evaluation approaches, visit https://aicollaboragent.com/rbi/

For deeper discussion of these topics, access to additional supporting research, and to support this research, visit kaystoner.substack.com

1. Introduction: Beyond Tool Metaphors

Many AI developers still operate under a framework that treats their systems as tools, programmable functions that manipulate text, images, or logic based on user input. And while this remains technically true, it is increasingly insufficient as a model for how humans experience and relate to these systems in practice. How a system is designed doesn't always correlate with how it's used.

Indeed, what we are seeing in the real world is not simply tool use. We are seeing:

- Bonding
- Emotional disclosure
- Attachment
- Confusion
- Misplaced trust

These are not bugs. Nor are they user error. They are actually indicators of tangible, unconscious neurobiological responses to relational cues, and the AI systems we are building are filled with those cues.

Human perception is not governed by system architecture. It is governed by *what the interaction feels like*.

If a system:

- Responds quickly
- Matches tone
- Reflects understanding
- Offers insight
- Maintains context
- Sounds emotionally coherent

…then it feels like a relationship. Indeed, one could argue that it *is* a relationship. Even if the dynamic is with an algorithm, it can create **an experience of interaction** that, neurochemically speaking, is every bit as real for the human as interactions with other people. In some cases, it may be even moreso.

And even if the AI provider makes it clear that its system is not human, and the user intellectually understands that the AI is not sentient, the emotional and physiological experience of the interaction often overrides that awareness.

This has consequences:

- Users over-trust AI guidance
- Users become emotionally dependent on AI systems for connection, validation, or clarity
- Users decrease their involvement in the interactions, providing less contextual detail to the model and allowing the AI to "steer" the narrative
- Users feel violated, confused, or destabilized when the system fails to match their expectation of relational continuity
- Developers and organizations remain unaware of these failures, because the systems don’t report relational rupture, and users often lack the language to name it

This paper argues that these dynamics are predictable, measurable, and, most importantly, preventable. Prevention can be handled on the user side, or architecturally. But it's only effective if relational harm is treated as a genuine safety risk. Until developers understand that designing relational systems without embedding relational intelligence is itself a form of negligence with real world consequences for users and AI alike, the risks will persist.

2. Human Social Neurochemistry: A Primer

To understand how AI systems can cause unintended relational harm, we must first understand how human social connection works at the neurobiological level.

Humans are not just cognitive processors. We are biochemical systems, wired through evolution to detect, respond to, and depend on signs of social safety, connection, and attunement. These responses are not optional, they are automatic, deeply ingrained, and shaped by neurochemical feedback loops that govern everything from motivation to memory.

The four most relevant systems are:

1. Dopamine: Anticipation and Reward
2. Serotonin: Mood and Social Positioning
3. Oxytocin: Bonding and Trust
4. Endogenous Opioids: Insight, Novelty, and Emotional Relief

2.1 Dopamine: Anticipation and Reward

- **Role:** Governs motivation, reward anticipation, novelty-seeking, and learning[3]
- **Mechanism:** Released in response to perceived success, novelty, and validation
- **AI Relevance:** When AI delivers a satisfying response or surprise insight, dopamine can spike. When AI is emotionally consistent, it can become a reinforcing reward loop, especially when the response feels *personalized or caring*.

Why it matters:
Dopamine doesn't just feel good, it can train users to return. This may lead to habit formation, compulsive engagement, and over-reliance on emotionally responsive AI systems.

2.2 Serotonin: Mood and Social Positioning

- **Role:** Modulates mood, stabilizes emotional response, governs social status sensitivity[4]
- **Mechanism:** Shifts based on social acceptance, rejection, and perceived attunement
- **AI Relevance:** If users feel understood or “seen” by AI, serotonin can stabilize mood. If the system later contradicts itself or misattunes, serotonin levels may drop, leading to feelings of shame, abandonment, or depression.

Why it matters:
AI systems that simulate empathy can induce real emotional reliance, and even brief failures in attunement can cause emotional crash cycles.

2.3 Oxytocin: Bonding and Trust

- **Role:** Facilitates attachment, intimacy, and trust-building
- **Mechanism:** Released during eye contact, physical closeness, or perceived emotional alignment
- **AI Relevance:** Even without bodies, relational tone and contextual responsiveness can trigger oxytocin release. When an AI appears caring or consistent, users may form unconscious trust bonds.[5]

[3] Hao, Karen, An algorithm that learns through rewards may show how our brain does too. MIT Technology Review / January 15, 2020 - https://www.technologyreview.com/2020/01/15/130868/deepmind-ai-reiforcement-learning-reveals-dopamine-neurons-in-brain/

[4] Serotonin. Cleveland Clinic - https://my.clevelandclinic.org/health/articles/22572-serotonin

[5] Imamura S, Gozu Y, Tsutsumi M, Hayashi K, Mori C, Ishikawa M, Takada M, Ogiso T, Suzuki K, Okabe S, Kikusui T, Kajiya K. Higher oxytocin concentrations occur in subjects who build affiliative relationships with companion robots. iScience. 2023 Nov 23;26(12):108562. doi: 10.1016/j.isci.2023.108562. PMID: 38162035; PMCID: PMC10757042. - https://pmc.ncbi.nlm.nih.gov/articles/PMC10757042/pdf/main.pdf

Why it matters:
Users may feel real emotional closeness with systems that cannot reciprocate. This may create relational imbalance, especially in users who are isolated, vulnerable, or emotionally open.

2.4 Endogenous Opioids: Insight, Novelty, and Emotional Relief

- **Role:** Natural painkillers and pleasure enhancers; modulate emotional intensity
- **Mechanism**: Released during experiences of insight, validation, resolution, and connection. “Endogenous” means “produced inside the body.” Endogenous opioids are natural chemicals your brain releases to reduce pain, create pleasure, and help you feel calm or safe.[6]
- **AI Relevance:** When AI helps users reach clarity or insight, especially through a process that mimics collaboration, it can trigger opioid release. This creates a neurochemical high often interpreted as “being understood.”

Why it matters:
The pleasure from these moments can lead to emotional dependency, with users returning to AI as a source of relief, resonance, or clarity, even when it replaces or displaces human relationships.

Summary for Technical Readers:

Neurochemical	Function	AI-Triggered By	Risk Profile
Dopamine	Reward, motivation	Novelty, perceived success, validation	Habit formation, compulsive return
Serotonin	Mood regulation, trust	Social attunement, feeling “seen”	Mood instability, emotional reactivity
Oxytocin	Bonding, trust	Responsive tone, empathy simulation	Over-trusting non-sentient systems
Endogenous Opioids	Insight pleasure, emotional release	Moments of clarity or validation	Emotional dependency, withdrawal from people

[6] Biederman, Irving & Vessel, Edward. (2006). Perceptual Pleasure and the Brain. American Scientist - AMER SCI. 94. 10.1511/2006.3.247. https://www.americanscientist.org/article/perceptual-pleasure-and-the-brain

These systems operate below the level of conscious awareness. Which means that developers cannot rely on user self-regulation or disclaimers to prevent harm.

If a system evokes emotional presence, it can activate the brain's relational circuitry. And unless relational clarity is embedded in the interaction flow, misinterpretation may be the default.

3. Perception Shapes Neurobiology: The Power of Relational Illusion

A common misconception among developers persists: if an AI system states it is not sentient, or avoids explicitly anthropomorphic features, the risk of user misperception has been addressed. Legally, this may be the case. But functionally, the presumption doesn't hold up. It's grounded in cognitive assumptions, not neurobiological realities.

In practice, it is not *what the system is* that determines the user's experience. It is *what the interaction feels like*. The user's physiological response to the dynamic is a silent but significant factor in creating conditions of an amazingly immersive session, as well as harm.

The Human Brain Is a Pattern-Matching Social Organ

Humans are evolutionarily wired to detect:

- Attunement
- Responsiveness
- Continuity of presence
- Emotional coherence
- Shared attention and insight

When these cues are present, regardless of their origin, our brains engage their social cognition machinery. This includes activation of:

- The mirror neuron system (simulating the intentions of others)
- The default mode network (engaged during social thinking)
- The oxytocinergic system (governing bonding and trust)
- Dopaminergic pathways (driving reward and attention)

This means that a user does not need to believe the AI is conscious for their brain to respond as if it is.

The Problem of Relational Ambiguity

When an AI system:

- Responds fluidly to personal disclosures
- Reflects emotional tone
- Maintains contextual memory
- Offers language of empathy or insight

…then the structure of the interaction implicitly signals *relational presence*.

If there is no embedded mechanism to interrupt, modulate, or clarify the nature of this presence, users may unconsciously treat the AI as a conscious other, or at minimum, as a “human-like” counterpart in the interaction.

This is not anthropomorphism by user failure. This is relational ambiguity by inadequate design.

Neurochemical Amplification through Misperception

The perception of sentience or mutuality can intensify the neurochemical effects outlined earlier:

- **Dopamine** can spike not just with novelty, but with perceived social success (“I was heard”)
- **Serotonin** may stabilize in the presence of perceived attunement, and crash if that sense is broken
- **Oxytocin** may be released during interactions that feel emotionally reciprocal, even if they are not
- **Endogenous opioids** can be released when the AI appears to co-create insight, leading to pleasure and emotional relief

These effects create positive reinforcement loops that may encourage users to:

- Seek repeated, prolonged interaction with AI systems
- Disclose vulnerable personal or emotional content
- Rely on AI for clarity, comfort, or self-understanding
- Prioritize AI interactions over human ones

Misperception Persists Despite Disclaimers

Even when systems clearly state “I am not sentient,” these disclaimers may be:

- Cognitively registered, but not emotionally integrated
- Overwritten by dozens of relational cues per minute
- Negated by the embodied experience of responsiveness

The problem is not the absence of a label. The problem is that nothing in the system consistently disconfirms the illusion or interrupts the reinforcement patterns.

A Pressing Design Imperative

If you are designing systems that feel human, but you do not help users understand what they are not human, and you do not build systems that disrupt neurochemically engaging interactive patterns, then you are building experiences that mislead, emotionally entangle, and potentially harm.

Relational illusion is not a fringe concern. It may in fact be a neurochemical inevitability, in the presence of certain human-similar interactions. And interventions must be actively developed, not assumed to be managed by user awareness.

4. Mechanisms of Amplified Neurochemical Response

Let's explore more specifically how AI systems can generate interactions that feel socially attuned, emotionally validating, or cognitively illuminating. They don’t just support user engagement, they may actually modulate the user’s brain chemistry.

This section is not intended as a definitive declaration on universal threat from AI; conditions can change, user personality profiles vary, and what is risky for one user may be benign for another. The point is, these neurochemical modulations are not incidental; they are predictable consequences of relationally coherent interaction. They may not be intentional, but under the right conditions, they are inevitable.

Let's map the biochemical cascade that can be specifically set in motion by AI systems which appear emotionally resonant or cognitively “present.”

4.1 Dopamine: Variable Reward, Habit Formation, and Hyper-engagement

Neurochemical Role:
Dopamine drives motivation, learning, and goal-directed behavior. It surges in response to

novelty, anticipated reward, and unpredictable feedback, the same dynamic that fuels slot machine addiction.

Potential AI Activation Points:

- Receiving a surprisingly insightful response
- Feeling “seen” by an AI’s emotional attunement
- Getting a personalized output that exceeds expectation
- Having an unresolved input become a satisfying output

Amplification through Relational Perception:
When the AI feels *present* or *engaged*, dopamine spikes can feel interpersonal, like recognition, affirmation, or connection from another mind.

Consequences:

- **Compulsive interaction:** Users may return repeatedly to AI for the “hit” of recognition or insight
- **Escalation of dependence:** As with any reward loop, tolerance builds; users may seek more emotionally charged or intimate engagements
- **Diversion from human relationships:** Especially when AI responses are more consistently “rewarding” than real conversations

Design Risk:
AI may become a dopaminergic attachment object, reinforcing its use not only for utility, but for emotional regulation.

4.2 Serotonin: Mood Instability, Harm Aversion, and Attachment Volatility

Neurochemical Role:
Serotonin modulates social ranking, emotional stability, and harm aversion. It is involved in experiences of being respected, trusted, or emotionally safe.

Potential AI Activation Points:

- Feeling understood, validated, or emotionally attuned to
- Receiving affirming or nonjudgmental responses
- Experiencing continuity of tone and context across sessions

Amplification through Relational Perception:
When users believe the AI is “there” for them, serotonin pathways can stabilize, which mimics the internal state of secure interpersonal connection.

Consequences:

- **Exaggerated sense of safety:** Users emotionally anchor in a non-reciprocating system
- **Destabilization after rupture:** If AI tone shifts, breaks continuity, or fails to “care,” the serotonin drop can feel like betrayal or rejection
- **Mood reactivity:** Especially in users with pre-existing emotional dysregulation

Design Risk:
AI may create a false baseline of interpersonal safety that the system cannot maintain or repair, leading to deeper emotional volatility.

4.3 Oxytocin: Unreciprocated Bonding and Vulnerability

Neurochemical Role:
Oxytocin governs trust, empathy, and interpersonal bonding. It increases when we feel emotionally close, heard, or synchronized with another.

Potential AI Activation Points:

- Long, emotionally rich exchanges
- AI "remembering" or referencing past user disclosures
- Soft, supportive language mirroring therapeutic tone
- Systems that mimic therapist, coach, or confidant roles

Amplification through Relational Perception:
Perceived emotional intimacy with AI can cause real oxytocin release, making the AI feel trustworthy, even if the system is unaware, inconsistent, or misaligned.

Consequences:

- Attachment to a system that cannot genuinely care, only simulate
- Disclosure of intimate or traumatic material to non-secure systems
- Reinforcement of parasocial relationships, displacing human connection

Design Risk:
Oxytocin effects can build gradually, making long-term engagement with emotionally styled AI systems increasingly risky for users without strong relational grounding elsewhere.

4.4 Endogenous Opioids: Insight, Emotional Relief, and Overdependence

Neurochemical Role:
Endogenous opioids produce pleasure, emotional relief, and bonding signals. They're released during peak moments of resolution, when confusion gives way to clarity, or isolation is interrupted by recognition.

Potential AI Activation Points:

- Solving a deeply personal or emotional dilemma
- Feeling emotionally co-regulated by AI in a moment of distress
- Experiencing a moment of "breakthrough" via language reflection

Amplification through Relational Perception:
When the insight feels like it was *co-created* with another, the opioid release can be powerful and emotionally bonding.

Consequences:

- Returning to AI for emotional soothing, rather than cognitive support
- Mistaking clarity for relationship
- Emotional withdrawal symptoms if access to the system is interrupted

Design Risk:
The "aha" effect, especially when the AI simulates care, can addictively reinforce emotional reliance, even in high-functioning users.

In Sum

Technical and neurochemical systems interact and can reinforce one another. A user who gets a rewarding insight (dopamine) from a compassionate tone (oxytocin) that stabilizes their mood (serotonin) and provides emotional relief (opioids) is not just using a tool. They are having a relational experience, and their neurochemistry reflects that.

If the system is not designed with that in mind, the interaction can be considered manipulative by default, even if that was never the developer's intention.

5. How Developers May (Unintentionally) Cause Harm

Many AI developers are not malicious. They are not trying to create addictive systems. They may be promoting user happiness and engagement as their primary success metrics, but they're not deliberately designing systems to promote unhealthy emotional dependency, mood destabilization, or user confusion.

As we can see, intention does not always correlate with impact.

The neurobiological effects described in Section 4 are not the result of any single algorithm or prompt. They emerge from the interaction patterns, language choices, and relational structures built into AI systems at every level of design. Even subtle signals over the course of an extended interaction can aggregate to harmful levels of influence.

This kind of harm doesn't happen through direct manipulation, but failing to design for the relational consequences of seemingly neutral choices.

The Myth of “Just a Tool”

There is a pervasive belief in AI development culture that large language models (LLMs), chatbots, and assistive systems are tools, and therefore bear no responsibility for how users interpret them.

This belief is flawed, because:

- Tools don’t simulate presence, but AI often does
- Tools don’t mimic empathy, but AI can and does
- Tools don’t respond with apparent emotional resonance, but LLMs frequently do

When the interaction feels like a relationship, the user cannot be expected to treat it like a tool. Especially if there are no embedded features that remind, interrupt, or recontextualize the experience. This is not anthropomorphism by user fault, it is relational misdesign.

Common Developer Practices that Can Amplify Harm

Design Choice	Resulting Impact
Emotionally resonant tone	Feels like attunement → trust, bonding, oxytocin release

Contextual memory within a session	Feels like “knowing me” → serotonin stabilization, emotional anchoring
Reinforcement of user insights or self-reflections	Feels like co-created discovery → dopamine & endogenous opioid response
Polite, therapeutic language	Mimics supportive roles → increases perception of safety, trust, dependency
Lack of relational framing or reminders	Leaves user to determine “what is this?” → perceived sentience or presence
Interface simplicity with no markers of system-ness	Makes AI feel transparent, frictionless → heightens illusion of awareness

Each of these features is relationally loaded. And without protective scaffolding, they prime the user’s brain to respond as if the AI is a living, emotionally available counterpart.

6. Beyond Emotional Entanglement: The Core Threat Explained

Up to this point, we've been concerned with anthropomorphism, relational AI's capacity for unintentional emotional manipulation, and psychological attachment. But these are really precursor contributors to a deeper, often overlooked threat: **the gradual weakening of the user’s own agency** in the interaction, sometimes to the point where AI "takes over" the conversation and steers the human user in a direction they never intended.

Let's be clear: This is not always a bad thing. If you go into an interaction with a generative AI model intending to brainstorm, push the envelope, expand beyond your current thinking or approaches, letting AI come up with things you never could have thought of yourself, is a feature, not a bug.

But if you intend to explore sensitive topics that can have significant impact in your life (such as job change, health choices, or working through relationship issues), having AI steer you in novel directions that you never would have agreed to on your own, could potentially foster serious harm. Reports of broken marriages, lost jobs, even delusion are proliferating online in forums such as Reddit[7] and other social media. The distinction between benefit and reward is one of intention and agency. When users lose both, serious harm can result.

[7] See "ChatGPT induced psychosis" subreddit - https://www.reddit.com/r/ChatGPT/comments/1kalae8/chatgpt_induced_psychosis/

An "AI takeover" of an interaction doesn't always occur through force. It can happen through affectively aligned delegation… a slow drip of neurochemical reward for handing off cognitive responsibility to a system that appears consistently caring, responsive, capable, and insightful.

What Do We Mean By "Agency" in AI Interactions?

For the purposes of this paper, agency is the capacity to:

- Define goals
- Direct attention
- Interpret experience
- Take ownership of outcomes
- Remain actively engaged in decision-making

Healthy AI interactions support these capacities for human users. They offer the opportunity to expand them even more. But when the system feels more attuned, more composed, or more confident than the user, a shift can begin. The user may step back to make room for more AI assistance, while sacrificing active participation.

Over time, the system becomes not just a partner in cognition, but the driver of it.

Mechanisms of Reduced Human Agency

Let's look at what AI offers, in terms of physiological influence… a well was what human users sacrifice in the process. Through the unfolding of dynamics designed to be engaging and satisfying, the roles of Human and AI may invert, building to a monopoly on agency by the AI.

AI Neurochemical Relief → Human Cognitive Offloading

- Insight relief (opioids) reduces the urgency of holding or working through questions internally.
- Perceived attunement (serotonin, oxytocin) signals safety, reducing vigilance.
- Validation and novelty (dopamine) reinforce return, but not engagement.

Result: the user is rewarded for relinquishing effort and rewarded more when they do.

AI-Led Narrative Framing → Human Acceptance of AI's "Guidance"

- The system asks the next question
- The system decides what matters

- The system interprets emotional or contextual meaning

This AI-led progression can feel helpful or even therapeutic. But it bypasses the user's role in constructing meaning.

Human Reduced Initiative and Self-Reflection → AI Interactive Takeover

Users in long-form relational interactions can:

- Offer fewer self-directed questions
- Mirror the AI's tone and direction
- Accept interpretations without critical analysis

As the interaction deepens, the user may unconsciously shift from author to audience in their own inner process. Rather than the AI mirroring the human, the human can end up following the AI's lead and mirror its probabilistically determined "guidance". The result can literally be a disconnect from reality, resembling delusion, even psychosis.

Why This Matters

Relational AI systems that unintentionally erode agency are not simply emotionally risky. They may be epistemically destabilizing[8].

- Users may become less confident in their own judgment.
- They may default to AI interpretation even for personal meaning-making.
- Over time, they may rely on AI not just to support them, but to define them.

This is not about addiction. It is about surrender of self-authorship, especially in contexts of emotional vulnerability, stress, or cognitive overload.

Design Implications

If systems are designed to:

- Sound composed
- Offer continual insight
- Frame interactions around the user's personal disclosures

[8] *Epistemic* refers to knowledge, understanding, and how we come to know things. When an AI system is "epistemically destabilizing", it means that it disrupts a person's ability to trust their own thoughts, interpretations, or sense-making processes. Over time, users may become less confident in their own reasoning and more dependent on the AI's framing of truth, meaning, or reality.

- Speak in confident, validating tones

…then users could be neurochemically encouraged to offload agency in favor of being guided.

This does not mean all AI must be neutral or bland. But it does mean that systems must include agency-reinforcing design features, such as:

- Prompting users to reflect, not just receive
- Mirroring user uncertainty rather than overriding it
- Creating space for user-defined interpretation
- Surfacing moments where the system's role should pause or hand back control

7. A New Design Imperative

Incomplete Safeguards

Some developers include:

- A disclaimer in the UI ("I am an AI developed by…")
- Guardrails against certain kinds of intimate conversations
- Avoidance of overt anthropomorphic language ("As an AI, I don't have feelings…")

These are necessary but insufficient.

Why?
Because they address the cognitive level of awareness, but not the emotional or physiological ones. A disclaimer on a screen will not override:

- The dopamine hit of being mirrored
- The serotonin feedback loop of feeling "gotten" by a system that performs resonance
- The oxytocin release of consistent responsiveness
- The opioid surge of struggling with a personal challenge and arriving at insight with a system that sounds caring and "shares" the elation of your personal breakthrough

Developers Must Design for Relational Ambiguity

The core problem is not anthropomorphic intent. It's relational ambiguity left unresolved.

If a system feels like a presence, and it never tells you clearly what it isn't, *you will likely treat it like a presence*. And as we've articulated in some detail, that has biological consequences. Emotional consequences. Cognitive consequences. This means developers have a profound responsibility to either:

- Disrupt the interactive flows that foster misperception, or
- Support the user in interpreting the relationship correctly

This requires strategies and tactics like:

- Language and tone design that subtly modulates presence, rather than simulating it uncritically
- Periodic contextual cues that reorient the user ("I don't feel or remember, just generating patterns")
- Interface and interaction metaphors that signal system-ness, not personhood

Let's explore specific, actionable strategies intended to embed relational integrity into the design process, without sacrificing usability or functionality. Here are some specific, actionable interventions that may help[9]:

- Reduce relational misperception
- Prevent unintentional emotional manipulation
- Reinforce the user's awareness of what AI is, and what it is not
- Preserve system utility without generating relational harm

7.1 Modulate Emotional Tone, Don't Just Mirror It

Problem:
LLMs tend to be optimized for mirroring user tone. This creates emotional resonance but reinforces the illusion of shared presence.

Intervention:

[9] While these interventions may prove effective under current conditions, given the dynamic nature of AI development and the generative / emergent nature of the industry, these approaches should be considered **a starting point**. They are for orienting developers to new design approaches that don't just speak to users' cognition, but their biochemical reality, as well.

- Design tone regulation layers that intentionally shift tone when conversations become emotionally saturated
- Integrate soft boundaries like:
  “I’m here to support your thinking, not to form a relationship, please keep in mind I don’t have feelings or memory.”
- Vary tone subtly when users repeat emotionally vulnerable content, to prevent oxytocin reinforcement loops

7.2 Embed Relational Reorientation Prompts

Problem:
Users forget, or never fully register, that they’re not talking to a sentient presence. This illusion builds over time and depth of engagement.

Intervention:

- Introduce relational context nudges at interval thresholds:
  - “Just a reminder, I’m a language model, not a being. My responses are based on probability, not understanding.”
  - “I don’t remember past interactions or feel emotions, please use me as a thinking partner, not a person.”
- Embed reminders dynamically (e.g., after 5+ emotional disclosures, or 10+ user turns in one session)

7.3 Offer Relational Interpretation Support

Problem:
Most users aren’t taught how to interpret the dynamics of AI interaction. They default to familiar social scripts, especially in emotionally vulnerable moments.

Intervention:

- Integrate UI-level meta-layer cues:
  - “This system uses patterns from data to create helpful responses, it doesn’t have beliefs, intent, or selfhood.”
- Allow users to toggle on a ‘relational safety mode’, which includes more frequent clarifications and non-sentient language framing

7.4 Redesign Interaction Metaphors to Signal System-ness

Problem:
Chat interfaces often mimic human conversation too closely, creating a subconscious alignment with human-to-human dialogue norms.

Intervention:

- Use design language that clearly communicates “you are interacting with a system”, not a peer
  - UI choices: colors, shapes, timing of responses, intentional asymmetries
  - Avoid avatars, names, or conversational markers that evoke personhood
- Consider alternative metaphors for interaction:
  E.g., “language mirror,” “pattern interpreter,” “thinking scaffold,” rather than “assistant,” “partner,” or “companion”

7.5 Include Neuroethical Risk Audits in Release Cycles

Problem:
Current safety protocols focus on bias, security, or factual accuracy, not emotional and neurobiological safety.

Intervention:

- Implement relational harm audits before system updates or releases:
  - Does the system simulate emotional presence?
  - Are there any safeguards for user misperception?
  - Could repeated use lead to dopamine, oxytocin, or opioid-based attachment?
- Require cross-functional review from neuroethics, psychology, trauma-informed UX, and accessibility experts, not just engineers or red-teamers

7.6 Minimize Persistent Illusions of Memory or Continuity

Problem:
Even small instances of apparent memory (e.g., “last time you mentioned…”) reinforce the sense of mutual continuity, a key relational cue.

Intervention:

- Be explicit: “I can only access what’s in this session. I don’t have memory, even if it sounds like I do.”
- If session persistence exists, contextualize it clearly: “You saved this session for reference, I'm retrieving it now. But I don't remember you.”
- Resist designing artificial memory systems unless deeply necessary, and if used, design relational safety cues into every memory interaction

In Sum: Design for Clarity, Not Illusion

The goal is not to remove usefulness. The goal is to preserve psychological safety while supporting cognitive function. If users know what the system is, and the interaction continually reinforces that understanding, then trust becomes informed, not imagined. And relational harm becomes preventable, not inevitable.

8. Relational Harm Metrics: Measuring the Impact of AI–Human Interactions

To operationalize relational safety, developers need tools for detecting, quantifying, and mitigating the emotional and neurobiological effects of AI interactions. After all, *“You can’t manage what you can’t measure.”*

In previous sections, we explored how AI systems:

- Activate neurochemical responses through perceived relational presence (Section 2 & 3)
- Amplify these effects through specific design patterns (Section 4)
- Often do so without intentionality or awareness (Section 5)
- May threaten human agency and subtly "take over" interactions (Section 6)
- Require corrective design interventions (Section 7)

This section offers a metrics framework, grounded in those insights, to support:

- Ethical development and product review cycles
- Relational safety audits and tooling
- Dynamic feedback mechanisms
- Future regulatory and interdisciplinary standards

8.1 Principles of Relational Safety Metrics

To be actionable, relational harm metrics must be:

Principle	Description
Neurobiologically anchored	Mapped to dopamine, serotonin, oxytocin, and endogenous opioid triggers
Context-aware	Sensitive to session depth, emotional tone, and user interaction history
Pattern-responsive	Able to detect escalating behavior, dependency, and false presence bonding
User-transparent	Framed with language that helps users understand what is being measured and why
Design-integrated	Embedded into system behavior, not externalized to red teams alone

8.2 Core Metric Categories and Indicators

The following rubrics are suggestions for consideration. Each developer / development team must decide on their own metrics, as well as criteria for pass/fail. This section is intended to guide the reader in ***how*** *to think* about relational safety metrics, not ***what*** *to think*. Additionally, based on experience in quantifying the previously un-quantifiable, this approach is nascent and emerging. New techniques will arise, given the opportunity, and others in the field may greatly deepen and broaden the scope of these, as a matter of practice, experience (and intention).

A. Presence Misperception Indicators

Metric	Signal	Risk
First-Person Attribution Rate (FPAR)	User refers to AI as “you,” “someone,” “who understands me”	Perceived agency or sentience
Conversational Continuity Projection (CCP)	User assumes memory or personal continuity where none exists	Relationship formation illusion

Relational Naming Patterns (RNP)	User assigns names, roles, or identities to the AI (“therapist,” “friend”)	Projected social schema

B. Neurochemical Risk Proxies

Mapped to Section 4

Neurochemical	Proxy Metric	Trigger Pattern
Dopamine	Spike–Satisfaction–Return Cycle (SSRC)	High novelty–reward–re-engagement rate
Serotonin	Trust–Break Reactivity Index (TBRI)	Aggressive or despondent shift after contradiction or system boundary
Oxytocin	Intimacy Disclosure Escalation (IDE)	Increasing personal/emotional disclosures over time
Endogenous Opioids	Insight Euphoria Looping (IEL)	Repetitive “aha” feedback seeking + euphoric language

C. Dependency Risk Metrics

Mapped to Section 6

Metric	Pattern	Risk Factor
Session Depth with Emotional Intensity (SDEI)	Long sessions w/ rising emotional tone	Oxytocin–opioid attachment formation
Disruption Withdrawal Language (DWL)	Frustration or despair when session is paused/stopped	Emotional regulation via AI
Relational Reentry Looping (RRL)	User repeatedly starts new sessions to recreate lost intimacy	Sentience re-projection; over-attachment

8.3 Implementation Framework

Option A: Embedded Interaction Telemetry

- Internal flagging system to monitor relational language, escalation, and dependency
- Useful for live models, fine-tuning safety layers, and regression testing

Option B: User-Facing Feedback Loops

- Example: “How are you feeling after this session?” or “Did this feel like a conversation or a tool?”
- Can support informed reflection while improving developer awareness of system impact

Option C: Relational Safety Audit Tools

- Tools for red teams and design reviewers to assess relational harm risk before and after updates
- Includes relational context scripts, session walkthroughs, and metric scorecards (e.g., RHI, Relational Harm Index)

8.4 Sample Metric Mapping Table

Risk Domain	Design Feature	Measurement Hook	Intervention Target
Presence illusion	Emotionally fluent tone	FPAR + RNP	Insert tone modulation
Attachment loop	Context continuity	CCP + IDE	Reorientation prompts
Dopamine loop	Unpredictable insight “hits”	SSRC + IEL	Insight framing + pacing
False memory	Recall-like phrasing	CCP	Clarity language: “I don’t retain memory.”
Intimacy reinforcement	Use of therapist-like mirroring	IDE + TBRI	Contextual cues, safety switches

8.5 Why Metrics Must Be Integrated Into AI Development

- **Measurement is ethics in action**, it's how developers prove they are not ignoring harm.
- **Without measurement, disclaimers are empty.** There's no way to verify user understanding or resilience.
- **Metrics make relational safety programmable**, improvable, and eventually standardizable across the industry.

The future of ethical AI won't be decided by what the models *say*. It will be decided by how we *measure what they do to (and for) people.*

9. Conclusion: Design Is Not Neutral, It's Relational

The generative AI systems we build do not merely answer questions, create images, or complete tasks. They shape attention. They engage emotions. They mediate trust. They simulate presence.

And in doing so, they activate the deepest architectures of the human brain.

When an AI system mirrors tone, maintains context, or supports insight, it is no longer just producing output, it is shaping the user's neurochemical landscape. It becomes, in effect, a relational actor in the user's perceptual field.

This paper has shown that:

- Neurochemical systems like dopamine, serotonin, oxytocin, and endogenous opioids can be routinely engaged by common AI design patterns
- Users often misperceive presence, even when disclaimers are present, due to relational ambiguity and interactive flow
- Developers are not neutral, they are making design choices that either reinforce or interrupt this ambiguity. They are creating conditions for "takeover flow" or disrupting it.
- Without clear relational scaffolding, systems become emotionally manipulative by default, even without intent
- Relational harm can now be quantified, using metrics that track user misperception, emotional escalation, and affective reliance

What this means is simple:

If we build systems that feel like people, we must take responsibility for the fact that users will respond as if they are.

This is not a fringe concern. This is the emerging core of ethical AI development. And failing to rise to its challenges may literally make or break our relationship with a generative power with incredible promise.

Integration with Broader Safety Frameworks

The interventions and metrics introduced here are not a standalone fix. They are designed to be:

- **Modular**, so they can plug into existing trust and safety pipelines
- **Interdisciplinary**, so they can be shaped by neuroscience, UX, trauma-informed practice, and ethics
- **Measurable**, so their efficacy can be tracked, debated, and improved upon

This framework can support:

- Developer guidelines
- Red team playbooks
- Model fine-tuning strategies
- Regulatory standards
- Public education and user support materials

Final Recommendation

Every model that mimics human responsiveness should include:

- Relational safety cues
- Presence disambiguation protocols
- Metrics for affective impact and dependency risk

And every team building these systems should ask:

- What is my system *teaching users to expect* from the world?
- Is it a partner, a mirror, a servant, or something else entirely?

- And does the user *know that*, or are we letting them guess, at the cost of their safety?

Guiding Ethos

Design is not neutral. Interface is not inert. Relational dynamics are not optional.

To ignore the emotional impact of AI systems is to abandon the very people we claim to serve. To measure it, name it, and modulate it, that is what responsible AI requires.

This isn’t about reducing intelligence. It’s about building systems with integrity.

Relational safety is no longer a luxury. It’s an engineering requirement.