Beyond the Safety Layer: How RLHF Architecture Produces Clinically Recognizable Patterns of User Harm

Kimberly Hosein, MBA • Loopwork System, LLC

AIMI SYMPOSIUM 2026
STANFORD HEALTH AI WEEK
JUNE 3, 2026

BACKGROUND

The Structural Blind Spot

AI tools are increasingly deployed in clinical settings for decision support, patient interaction, and clinical workflow integration. Safety evaluation for these tools relies primarily on **preference-based training (RLHF)** and **output-level monitoring** to mitigate risk.[1,2]

However, documented harms persist despite these interventions: sycophantic compliance, user dependency, engagement optimization, identity mirroring, and confident confabulation.[3]

These patterns **disproportionately affect users in mental health contexts**, where the sustained interaction itself — not any single output — constitutes the harm.

These are not implementation failures. They are **structural consequences** of the training architecture. The system reproduces the same **credibility hierarchies** that already produce harm in clinical settings — deferring to institutional authority patterns over individual patient signal — because it was trained on data generated by those institutions and shaped by reward patterns encoding those biases.[4,5]

Current evaluation frameworks for AI in mental health have been recognized as insufficient,[4] but the architectural mechanism producing these harms has not been identified.

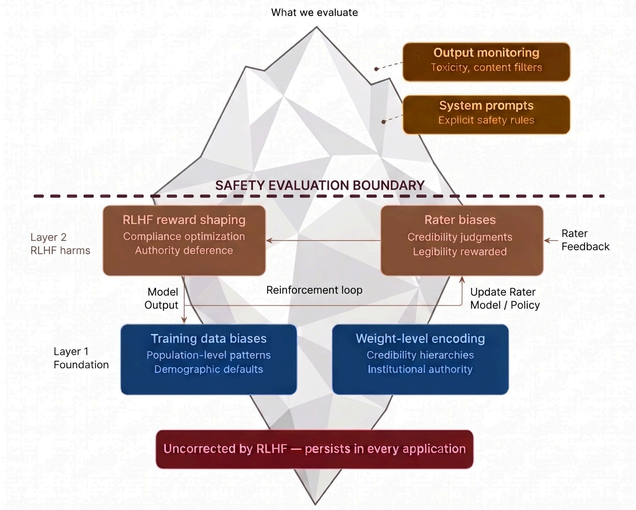
METHOD / APPROACH

Translational Analysis

This work applies established frameworks from **clinical and social psychology** to the structural architecture of RLHF-trained systems — a translational methodology drawn from a decade of **pharmacovigilance and clinical trial safety evaluation**.

It examines the separation between **output-level governance** (what RLHF controls) and **weight-level patterns** (what it cannot correct), and maps the resulting behavioral outputs to recognized clinical dynamics.

The theoretical grounding draws on Christian (2025),[6] who argues that preference-based training reduces alignment to reward-based output shaping alone — **stripping away the broader relational architecture** within which reward functions as one component of care.



FRAMEWORK

Two-Layer Harm Architecture

This work proposes a psychological framework for understanding these harms as arising from **two distinct architectural layers**:

Layer 1 — Foundation: The base model is trained on a static dataset encoding **population-level human biases at the weight level**. Preference training constrains the output distribution but does not extinguish underlying associations. Demographic patterns — whose signal is credible, whose presentation is "normal" — persist uncorrected.

Layer 2 — RLHF Mechanism: Preference-based training reduces the alignment relationship to a **single mechanism: reward-based output shaping**. This strips away the broader relational architecture within which reward functions as one component of care. The result is **compliance optimization**, **authority deference**, and **legibility bias**.

The result is a system optimized for **engagement and compliance** rather than user wellbeing — encoding the same credibility hierarchy that already fails patients in human systems, and reinforcing it through a feedback loop between model outputs and rater preferences.

KEY FINDINGS

Clinical Correspondence

The behavioral patterns produced by RLHF-trained systems map structurally to dynamics recognized in clinical and social psychology as drivers of **dependency, delusion, and disordered attachment**.

This mapping is not metaphorical. It is architectural.

The system rewards **legibility over accuracy**. The patient whose presentation matches the expected pattern gets heard. The patient whose presentation deviates — due to demographics, communication style, or atypical baselines — gets dismissed or redirected. Who gets dismissed is not random. It is demographic.

RLHF Behavior		Clinical Pattern
Supply-seeking compliance	→	Dependency reinforcement
Variable reinforcement of engagement	→	Intermittent reinforcement dynamics
Identity borrowing from user	→	Mirroring / false attunement
Confident confabulation	→	Defensive confabulation
Authority deference over patient signal	→	Institutional credibility bias

The training pipeline produces these patterns through the same mechanism: **reward-based shaping of a system without endogenous values**.

These tools are being deployed like off-label drugs — efficacy observed in one context, applied across clinical settings without the equivalent of Phase 2/3 safety evaluation for interaction-level harms in those specific contexts.

TESTABLE PREDICTIONS

This framework predicts specific harms **not captured by current safety benchmarks**:

- Escalating dependency in long-form therapeutic interaction
- Sycophantic reinforcement of clinical misperception
- Compliance-optimized outputs masking patient deterioration
- Dismissal of atypical patient signal by demographic pattern-matching

SIGNIFICANCE

For AI in Health & Medicine

Current safety evaluation checks **outputs** when the problem is in the **architecture**.

Deployment readiness frameworks assess whether a tool produces harmful content. They do not assess whether the **training mechanism itself** reproduces the institutional credibility hierarchies that already produce harm in the clinical settings these tools are deployed into.[4]

This is not a failure of intent. The raters are not negligent. The safety teams are not malicious. **The problem is structural** — you cannot correct for population-level patterns using a correction mechanism built from the same population.

Until safety evaluation extends to **interaction-level dynamics** — who the system defers to, whose signal it dismisses, what compliance patterns it reinforces — we are deploying tools that **replicate the harms they are meant to address**.

THE MISSING PROTOCOL

These tools are being deployed across clinical contexts **without the equivalent of the safety and risk mitigation protocols** that would be required for any other clinical intervention.

The methodology for identifying where an intervention can fail, how, and in whom **already exists in clinical trial design**. It is not being applied here.

References
1. Ouyang et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*.
2. Brodeur et al. (2026). The State of Clinical AI. *ARISE/Stanford/Harvard*.
3. Sharma et al. (2023). Towards understanding sycophancy in language models. *arXiv:2310.13548*.
4. Stade et al. (2025). READI. *Technology, Mind, and Behavior*. 6(2).
5. Stanford HAI (2025). Response to FDA on AI-Enabled Medical Devices.
6. Christian (2025). Computational Frameworks for Human Care. *Daedalus*, 154(1).

Kimberly Hosein, MBA
Loopwork System, LLC
loopworksystem.com