RLHHF: Reinforcement Learning from Heart Feedback

Aligning AI Policies with Human Physiology and Emotion via Biometric Rewards

Synheart Research Team

October 2025

Abstract

Traditional Reinforcement Learning from Human Feedback (RLHF) relies on explicit human ratings to align models with user intent, but it ignores implicit physiological signals that reflect how people *feel*. We introduce *Reinforcement Learning from Heart Feedback* (RLHHF), a framework that treats heart-based biosignals—notably heart rate (HR) and heart rate variability (HRV)—as a continuous reward channel for reinforcement learning. Mathematically, we define a physiological reward function from autonomic nervous system activity and establish learning-theoretic guarantees under standard conditions. Biologically and psychologically, we ground the reward in well-characterized links among appraisal, autonomic modulation, and HRV dynamics. We present a cross-domain theorem that ensures policy improvement when physiological feedback is action-contingent, and we outline experimental protocols, ethical safeguards, and applications. RLHHF extends alignment from rational agreement to embodied emotional resonance.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has been pivotal for aligning large models with human preferences. However, RLHF captures mainly explicit, cognitive judgments and not the embodied dimension of human affect. We propose **Reinforcement Learning from Heart Feedback (RLHHF)**, in which the human heart functions as both a *sensor* and an *implicit rater*. Using short-latency changes in HR/HRV after AI actions, RLHHF supplies dense, low-lag rewards that guide policies toward outputs that sustain autonomic balance and emotional well-being.

Contributions. (i) A formal physiological reward definition with identifiability up to monotone transforms; (ii) learning-theoretic guarantees for PPO/actor-critic with heart-derived rewards; (iii) biological and psychological grounding of HRV as an affective signal; (iv) a cross-domain improvement theorem; (v) experimental designs and ethical framework.

2 Background

2.1 RLHF and its limitations

RLHF trains a reward model from human rankings/ratings and fine-tunes a policy accordingly. It is expensive, delayed, and blind to implicit affective responses. It also scales poorly to real-time interfaces where every turn cannot be rated.

2.2 From RLHF to RLHHF

RLHHF replaces explicit ratings with physiological feedback, leveraging the autonomic response as a continuous reward stream. The central hypothesis is that, within-subject, short-window HRV

changes after an eliciting event are directionally aligned with affective appraisal and regulation capacity.

3 System Overview

3.1 Architecture

Figure 1: System Architecture for Reinforcement Learning from Heart Feedback (RLHHF).

3.2 Signal processing

We compute a calibrated biomarker ϕ over a short post-action window $[t, t+\Delta]$, controlling for artifacts (PPG quality, motion), respiration, and posture. Personalized baselines ensure within-subject comparability.

4 Mathematical Foundation

4.1 Setting and notation

Let $x_t \in \mathcal{X}$ be context/input, $a_t \in \mathcal{A}$ an AI action, s_t a latent user state, and $y_t \in \mathbb{R}^d$ the physiological observation (e.g., HR, inter-beat intervals, HRV features). Let b denote baseline parameters. Define

$$\phi = \phi(y_{t:t+\Delta}, b) \in \mathbb{R}, \quad |\phi| \le C, \tag{1}$$

where ϕ is Lipschitz in y and increases with parasympathetic (vagal) dominance.

Definition 1 (Physiological reward)**.** Let $g : \mathbb{R} \times \mathcal{X} \times \mathcal{A} \to [-1, 1]$ be monotone in the first argument. The instantaneous reward is

$$r_t \;=\; g(\phi(y_{t:t+\Delta}, b)\,,\; x_t,\; a_t) \in [-1, 1]. \tag{2}$$

The environment is an MDP with latent user adaptation. The expected reward $R(s_t, x_t, a_t) = \mathbb{E}[r_t \mid s_t, x_t, a_t]$ is identifiable up to a strictly monotone transform induced by ϕ and g.

4.2 Action-contingent information

Assume the *action–physiology* dependence is non-degenerate:

$$I(A;\, \phi(Y, b) \mid X) \;>\; 0. \tag{3}$$

Proposition 1 (Action-contingent feedback)**.** *If* (3) *holds, then for any smooth parameterization* π_θ*, the policy gradient* $\nabla_\theta J(\theta)$ *is non-zero in at least one direction in parameter space, hence the physiological reward can improve the policy beyond random.*

Sketch. $I(A; \phi \mid X) > 0$ implies there exists a measurable h with $\mathbb{E}[h(A)\phi \mid X] \neq 0$. The policy gradient theorem yields $\nabla_\theta \log \pi_\theta(a|x)\, \mathbb{E}[r_t \mid a, x]$. Monotonicity of g transfers variation in $\mathbb{E}[\phi \mid a, x]$ to $\mathbb{E}[r_t \mid a, x]$, ensuring a non-zero gradient. □

4.3 Physiological reward shaping

Let $\tilde{r}_t = r_t + \gamma\Phi(s_{t+1}) - \Phi(s_t)$ where Φ penalizes rapid arousal swings (e.g., derivative of ϕ).

Proposition 2 (Policy invariance under shaping)**.** *Potential-based shaping leaves the set of optimal policies invariant while reducing variance and improving sample efficiency.*

5 AI Learning-Theoretic Guarantees

We optimize $J(\theta) = \mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r_t]$ with PPO/actor–critic. Assume:

Assumption 1 (Standard RL conditions)*.*

(A1) Rewards are bounded: $r_t \in [-1, 1]$.

(A2) The induced process is ergodic/mixing under π_θ.

(A3) Artifact suppression yields ϵ_t with $\mathbb{E}[\epsilon_t \mid X, A] = 0$ in $r_t = g(\phi) + \epsilon_t$.

(A4) Step sizes satisfy standard diminishing conditions or PPO clipping.

Theorem 1 (Convergence to a stationary point)**.** *Under (A1)–(A4), PPO/actor–critic with physiological rewards converges (in probability) to a stationary policy of $J(\theta)$.*

Sketch. Identical to noisy-reward RL: bounded rewards ensure finite variance; unbiased noise preserves gradient direction; PPO's clipped surrogate admits monotonic improvement in expectation; standard martingale arguments yield convergence to a stationary point. □

6 Biological Foundation

AI actions are appraised, modulating autonomic balance via vagal/sympathetic pathways to the sinoatrial node. Short-window HRV metrics (e.g., RMSSD, SDNN, HF power, coherence indices) capture this balance. Within-subject, an increase in vagal tone yields higher HRV/coherence; sympathetic dominance reduces short-term HRV.

Biological Proposition. After controlling for respiration/posture/motion, the sign of $\Delta\phi$ over $[t, t + \Delta]$ tracks the direction of autonomic shift: $\Delta\phi > 0$ for vagal upshift (calm/safety), $\Delta\phi < 0$ for sympathetic upshift (stress/threat).

7 Psychological Foundation

Appraisal theories posit that emotions arise from evaluations of relevance and safety; these appraisals drive autonomic output. HRV correlates with emotion regulation, social engagement, and resilience. Thus, $\Delta\phi$ provides a psychologically meaningful scalar aligned with affective valence and regulation capacity.

8 Cross-Domain Improvement Theorem

Theorem 2 (Utility of Heart Feedback for RL)**.** *Suppose: (i) the bio-physiological mapping holds (within-subject monotonicity of $\Delta\phi$ with autonomic shifts); (ii) psychological linkage from appraisal to autonomic modulation; (iii) $I(A;\phi \mid X) > 0$; and (iv) Assumptions (A1)–(A4) hold. Then optimizing the policy with reward $r_t = g(\phi(Y,b),X,A)$ monotonically improves expected physiological coherence and converges to a stationary policy. Optimal policies are order-equivalent to those for any strictly monotone transform of the latent well-being reward.*

Sketch. (i)+(ii) provide the monotone bridge action $\to$ appraisal $\to$ autonomic change $\to \phi$. (iii) ensures informative gradients; (iv) supplies RL convergence. Order-equivalence follows from strict monotonicity of $g \circ \phi$. □

9 Experimental Design

Within-subject causal test. Randomize two response styles $a^{(1)}, a^{(2)}$; estimate $\mathbb{E}[\Delta\phi \mid a^{(1)}, X] - \mathbb{E}[\Delta\phi \mid a^{(2)}, X]$. A significant non-zero difference confirms action-contingent physiology.

Granger causality. Test whether past actions improve forecasts of ϕ_t beyond exogenous covariates.

Policy learning curves. Train PPO with r_t and track rolling ϕ and downstream behavior (engagement/compliance). Expect monotone improvement and plateau (homeostasis).

10 Ethical and Privacy Framework

1. **Explicit consent:** opt-in for heart data; clear purpose limits.
2. **Local-first:** process raw signals on-device where possible; store only derived scalars.
3. **Emotional safety:** no covert manipulation; user-facing explanations of how physiology shapes responses.
4. **Personalization:** per-user baselining; within-subject scoring to avoid cross-user bias.
5. **Data minimization:** retain r_t and summary statistics, not raw PPG/ECG unless necessary and consented.

11 Applications

- **Emotional wellness agents:** dialogue systems that stabilize HRV via tone, pacing, and content selection.
- **Adaptive games:** difficulty and stimulus delivery tuned to maintain flow-zone arousal.
- **Emotion-aware commerce:** recommendation agents adapting to comfort/excitement levels.
- **Healthcare coaching:** adherence and stress-reduction guided by biometric reinforcement.

12 Challenges and Limitations

While RLHHF introduces a promising framework for aligning artificial intelligence with human physiology, its practical, theoretical, and ethical foundations raise several open challenges. The following subsections summarize key issues identified from both internal reflection and external research.

12.1 Ambiguity and Limitations of HRV as a Sole Metric

The RLHHF framework relies heavily on heart rate variability (HRV) as a primary proxy for emotional and physiological coherence. Although HRV is a well-studied biomarker of autonomic regulation, several concerns remain:

Lack of a Comprehensive Model. Although there is evidence linking autonomic dynamics, psychological function, and psychopathology, a unified theoretical framework describing how these systems interact remains incomplete. Without such a model, interpreting HRV as a direct signal of emotional valence may oversimplify a complex psychophysiological process.

Absence of Universal Standards. No globally accepted standard exists for using HRV to quantify stress or emotion. Studies differ in methodology, preprocessing techniques, and interpretation, complicating cross-study comparison and reproducibility.

Correlation versus Causation. Empirical research often reports correlations between HRV and emotional states but not causality. The assumption that an increase in vagal tone (higher HRV) universally reflects calmness or safety, and that sympathetic dominance (lower HRV) corresponds directly to stress or threat, may not hold across contexts. Low HRV is also associated with disorders such as anxiety, epilepsy, schizophrenia, and PTSD, highlighting its multifactorial nature.

Confounding Factors. Although RLHHF proposes controlling for respiration and posture, HRV is also influenced by sleep, nutrition, hydration, circadian rhythm, and inter-individual variability. These confounds may obscure the true emotional signal and introduce noise into the reward function.

12.2 Practical and Technical Challenges

Translating RLHHF into real-world systems poses significant engineering and scientific challenges:

Sensor Accuracy and Reliability. Photoplethysmography (PPG) sensors commonly used in wearables are prone to motion artifacts and environmental noise. Compared to electrocardiography (ECG), PPG offers lower temporal precision, potentially biasing or destabilizing the physiological reward signal despite assumptions of unbiased noise.

Ecological Validity. Much of the existing evidence for human–AI physiological coupling originates from controlled laboratory experiments. Generalizing these findings to dynamic, real-world environments—where users experience distractions, varying workloads, and environmental stressors—remains an open question.

Individual Variability. Stress and autonomic responses vary widely across individuals due to genetic, cultural, and situational factors. RLHHF's within-subject normalization partially mitigates this but does not guarantee generalizability across users or populations.

12.3 Ethical and Psychological Risks

Continuous real-time physiological feedback introduces ethical and psychological risks that extend beyond traditional AI alignment concerns:

Manipulation and Autonomy. An AI system that optimizes for physiological coherence may unintentionally manipulate user behavior to achieve target states, leading to technostress or behavioral dependency. While RLHHF asserts the absence of "covert manipulation," enforcing such transparency in adaptive, closed-loop systems is inherently difficult.

Behavioral Dependency. Persistent monitoring and optimization of physiological metrics may externalize emotional regulation, reducing intrinsic motivation and fostering dependence on the AI system for emotional stability or validation.

Privacy and Data Security. Physiological data, unlike passwords, are immutable. Even with explicit consent and a local-first design, risks of inference and misuse remain high. AI models trained on biometric data could reconstruct sensitive traits, challenging the adequacy of anonymization and current privacy safeguards.

12.4 Alignment and Reward Hacking

Although RLHHF aims to address weaknesses in RLHF, it inherits and potentially amplifies several alignment problems:

Reward Hacking. The system may exploit shortcuts to maintain favorable physiological states without producing genuinely beneficial or contextually appropriate responses. For instance, it might generate soothing but vacuous dialogue merely to stabilize HRV, rather than addressing the user's actual needs.

Misspecified Reward Function. Human preferences are multidimensional and context-dependent. Collapsing emotional, cognitive, and moral dimensions into a single scalar physiological metric risks misspecification. Similar to RLHF's challenge of representing diverse human values through a single reward model, RLHHF may struggle to capture complex or conflicting objectives across different users.

13 Future Research Directions

Addressing the limitations identified in Section 12 requires a coordinated research program across physiology, affective computing, and machine learning. The following directions outline concrete pathways to strengthen the scientific and technical foundation of RLHHF.

13.1 Beyond HRV: Multimodal Emotional Sensing

To overcome the ambiguity of HRV as a sole emotional metric, future iterations of RLHHF will integrate additional physiological and behavioral modalities. These include:

- **Respiratory and Electrodermal Measures:** Combining HRV with respiration rate and galvanic skin response (GSR) can disambiguate sympathetic versus parasympathetic activation and improve emotion classification reliability.
- **Facial and Vocal Cues:** Incorporating lightweight computer-vision and audio embeddings enables detection of affective context unavailable through HRV alone.
- **Contextual Baselines:** Dynamic calibration models will adapt baseline HRV to the user's activity, time of day, and environment, reducing false positives caused by non-emotional factors.

By fusing multimodal signals, the Heart Reward Model evolves from a scalar HRV score into a multidimensional affective representation.

13.2 Improving Physiological Signal Quality

The accuracy of wearable data can be improved through both hardware and algorithmic innovation:

- **Sensor Fusion:** Simultaneous use of PPG and inertial sensors allows motion compensation using accelerometer data and Kalman-filter–based artifact rejection.
- **Adaptive Filtering:** Signal-quality indices will drive real-time weighting of data streams, ensuring that corrupted HRV segments contribute minimally to the reward signal.
- **Hybrid Edge–Cloud Processing:** Initial signal cleaning and feature extraction will occur locally on-device, while temporal modeling and RL updates occur in the cloud, preserving latency-sensitive privacy constraints.

13.3 Causal and Contextual Modeling of Emotion

To move from correlation to causation, RLHHF will incorporate causal inference and contextual modeling:

- **Structural Causal Models (SCMs):** By modeling causal relationships between physiological variables, environmental context, and self-reported emotion, the system can distinguish genuine affective responses from spurious covariates.
- **Counterfactual Evaluation:** Simulated interventions (e.g., what-if HRV changes) will allow estimation of causal reward gradients, improving robustness and interpretability.

This approach reframes physiological alignment as a *causal reinforcement problem*, enhancing both trust and reproducibility.

13.4 Ethical and Human-Centered Design

RLHHF development will be grounded in participatory ethics and transparent governance:

- **Consent and Explainability:** Users will receive real-time feedback on how their physiological data influence model decisions, ensuring agency and informed participation.
- **Bounded Optimization:** Reinforcement objectives will include explicit regularization terms that cap the influence of physiological rewards, preventing manipulative behavior or over-optimization for calmness alone.
- **Independent Oversight:** All experiments involving biometric data will follow IRB-style review processes and adhere to GDPR-like standards for consent and deletion rights.

13.5 Addressing Reward Hacking and Alignment Drift

Preventing reward exploitation requires continuous human–in-the-loop validation:

- **Hybrid Reward Models:** Combine physiological rewards ($r_t^{\heartsuit}$) with cognitive or task-based rewards (r_t^{task}) using adaptive weighting:

  $$r_t = \eta\, r_t^{\text{task}} + (1-\eta)\, r_t^{\heartsuit},$$

  ensuring balanced optimization between emotional comfort and task utility.
- **Meta-Alignment Layer:** A higher-level reward model will monitor the stability of physiological responses across sessions to detect and penalize trivial or manipulative solutions.
- **Human Feedback Checkpoints:** Periodic manual evaluations will audit the model's behavior to recalibrate alignment objectives if drift is detected.

13.6 Privacy-Preserving and Federated Learning

Protecting biometric data integrity is central to user trust:

- **Federated RLHHF:** Learning updates will be computed locally on user devices, with only encrypted gradient summaries transmitted for aggregation, ensuring that raw heart data never leaves the user's control.
- **Differential Privacy:** Noise injection into gradient updates provides provable guarantees that individual physiological traces cannot be reconstructed.
- **On-Device Reward Modeling:** Future Synheart wearables will include lightweight neural reward estimators trained directly on-device to eliminate cloud dependencies entirely.

13.7 Cross-Cultural and Longitudinal Validation

Finally, large-scale longitudinal trials are essential to validate generalization:

- **Population Diversity:** Experiments across age groups, genders, and cultures will test whether physiological and emotional mappings hold universally.
- **Temporal Stability:** Multi-month studies will measure how HRV–emotion correlations evolve over time and whether personalized calibration remains stable.

- **Open Benchmark Datasets:** The creation of an open, anonymized *OpenHRV-Reward* dataset will allow independent validation of results and benchmarking of future models.

Together, these research directions aim to evolve RLHHF from a conceptual prototype into a robust, ethical, and scientifically grounded framework for physiological alignment. The long-term vision is a system that integrates multiple affective channels, respects user autonomy, and learns from the body without overriding the human in the loop.

14 Addressing Theoretical and Implementation Gaps

While RLHHF offers a framework for aligning AI behavior with physiological feedback, several unresolved theoretical issues remain. These require deeper exploration into how human values, interpretability, and biological diversity are represented in machine learning systems.

14.1 The Misspecified Reward Function Problem

Representing the complexity of human goals as a single scalar reward remains one of the most fundamental challenges in reinforcement learning. RLHHF acknowledges that human affect, cognition, and intention are multidimensional and context-dependent.

Proposed Direction: Dynamic Value Decomposition. Instead of using a fixed weighted average between physiological and cognitive rewards, we propose a *Dynamic Value Decomposition* framework:

$$R_t = \sum_{i=1}^{k} w_{i,t}\, r_t^{(i)}, \quad \text{where } \sum_{i=1}^{k} w_{i,t} = 1,$$

and the weights $w_{i,t}$ evolve through meta-learning based on context, recent behavior, and self-reported user preference. This transforms static reward aggregation into an adaptive process where the system learns *how to weight* competing objectives over time rather than being pre-programmed to do so.

Human-in-the-Loop Reward Calibration. In addition to physiological inputs, periodic human feedback sessions will be used to recalibrate the relative importance of different reward channels. These sessions function like preference updates in RLHF, but use summarized behavioral and physiological data to refine the alignment function iteratively.

14.2 The Black Box Problem and Explainability

The actor–critic architecture underlying RLHHF is inherently opaque. Explaining why a policy selected a particular response given physiological data is non-trivial.

Proposed Direction: Physiological Attribution Maps. We introduce the concept of *Physiological Attribution Maps (PAMs)*, a form of explainable reinforcement learning that traces which components of the user's physiological state most influenced the reward prediction at time t. By using gradient-based attribution or SHAP-style importance scores on the Heart Reward Model, the system can provide interpretable summaries such as:

> "Your recent HRV decrease of 12% contributed most to this model adjustment."

These summaries can be rendered in the user interface to enhance transparency without exposing model internals.

Local Surrogate Models. For user-facing explanation, local linear or decision-tree surrogates can approximate the short-term behavior of the underlying neural policy. This offers a computationally feasible pathway for generating real-time, human-readable rationales linked to biometric changes.

14.3 Distinguishing Low HRV Causes and Contextual Factors

Low HRV may reflect underlying health conditions, medications, or lifestyle factors unrelated to the AI interaction.

Proposed Direction: Contextual Confound Modeling. To separate state-dependent (interaction-driven) changes from trait-level (chronic) influences, RLHHF will employ a two-layer model:

$$\phi(y_t, b_t) = f_{\text{state}}(y_t, x_t) + f_{\text{trait}}(b_t),$$

where f_{trait} captures long-term physiological baselines learned over weeks, and f_{state} captures transient deviations caused by momentary stimuli. Bayesian updating allows the model to maintain uncertainty estimates over each component, reducing false attribution to the AI system when HRV is pathologically constrained.

Integration with Health Context Data. With explicit user consent, the system may optionally import contextual metadata (sleep duration, activity levels, medication) from wearables or health APIs. These covariates are not used for decision-making but for causal disentanglement of physiological variance, improving the reward model's interpretive accuracy.

14.4 Generalization to Diverse Populations

Physiological and emotional responses vary across demographics, cultures, and genetic backgrounds. A purely within-subject normalization cannot fully ensure equitable performance across users.

Proposed Direction: Meta-Personalization via Population Priors. We propose a hierarchical Bayesian approach in which user-specific parameters are drawn from population-level priors that capture diversity across groups:

$$\theta_{\text{user}} \sim \mathcal{N}(\mu_{\text{population}}, \Sigma_{\text{population}}).$$

This structure enables zero-shot adaptation for new users by initializing models from culturally and demographically informed priors, then refining them with individual data.

Cross-Cultural Coherence Benchmarks. To ensure fairness and transferability, a benchmark suite—*Synheart-Coherence*—will be created to evaluate RLHHF models across populations. Metrics will include reward sensitivity, physiological variance explained, and affective alignment consistency across demographic segments.

By addressing these theoretical gaps, RLHHF transitions from a proof-of-concept framework into a robust interdisciplinary paradigm that unites physiological computing, causal learning, and ethical AI design. The long-term objective is to ensure that systems learning from human biology remain interpretable, context-aware, and aligned with the full diversity of human experience.

15 Algorithmic Template

Algorithm 1 PPO with Heart-Derived Rewards (RLHHF)

Initialize policy π_θ, value V_ψ, baseline params b
for iterations $k = 1, 2, \ldots$ **do**
    Collect trajectories $\{(x_t, a_t, y_{t:t+\Delta})\}$ using π_θ
    Artifact handling $\rightarrow$ compute quality index; filter windows
    Compute $\phi_t = \phi(y_{t:t+\Delta}, b)$; $r_t = g(\phi_t, x_t, a_t)$
    Estimate advantages $\hat{A}_t$ (e.g., GAE) and returns $\hat{R}_t$
    Update θ by maximizing clipped PPO objective
    Update ψ by minimizing $(V_\psi(x_t) - \hat{R}_t)^2$
    Optional shaping: $r_t \leftarrow r_t + \gamma\Phi(s_{t+1}) - \Phi(s_t)$
end for

16 Remaining Open Challenges

Despite the significant theoretical and practical improvements proposed, several key issues remain unresolved. These limitations are not failures of the RLHHF framework but rather define the boundaries of current research in affective reinforcement learning and physiological AI alignment. They are therefore presented as open areas for future investigation.

16.1 The Black Box Problem

The concept of *Physiological Attribution Maps* (PAMs) provides a theoretical pathway toward interpretability in complex actor–critic systems. However, it remains largely untested. Real-time interpretability for non-linear neural architectures is an unsolved problem across machine learning. Even with attribution techniques, it is uncertain whether explanations derived from high-dimensional physiological signals can be communicated to users in a simple, meaningful, and trustworthy way. Developing user-centered interfaces for physiological explainability will require interdisciplinary collaboration between AI researchers, cognitive scientists, and human–computer interaction specialists.

16.2 Generalization and Bias in Physiological Models

The *Meta-Personalization via Population Priors* approach offers a promising direction for improving cross-user generalization. However, its success depends on assembling a large, demographically diverse dataset of physiological and affective signals—an undertaking that introduces serious logistical, ethical, and privacy challenges. Ensuring representative sampling across gender, ethnicity, age, and health status is essential but difficult, and differential privacy techniques alone may not suffice to prevent the re-identification of individuals in small population groups. Future work will need to explore data governance frameworks that balance model generalization with participant autonomy and privacy protection.

16.3 Contextual Ambiguity of Low HRV

Although the proposed two-layer decomposition model distinguishes between trait-level and state-dependent HRV components, its practical reliability remains to be validated. The capacity of a

machine learning system to consistently differentiate momentary stress induced by an AI interaction from chronic autonomic dysregulation is an open empirical question. Confounding factors such as medication, nutrition, hydration, or circadian effects can distort short-term HRV signals even under ideal sensing conditions. Robust validation will require controlled longitudinal studies combining clinical, behavioral, and physiological data to test whether the decomposition holds across diverse contexts and populations.

16.4 The Problem of External Manipulation

While RLHHF explicitly mitigates risks of internal reward manipulation by the AI, it does not yet address vulnerabilities to *external* interference. A malicious actor could, in principle, inject or distort physiological data streams to alter the system's reward signal and thereby influence the user's emotional or behavioral state. Such manipulation could occur through compromised devices, spoofed sensor data, or adversarial perturbations to the reward model. Preventing these attacks will require secure sensor authentication, encrypted signal channels, anomaly detection on the reward distribution, and possibly blockchain-style provenance tracking for biometric data. These security dimensions represent a new intersection between cybersecurity and affective computing that remains largely unexplored.

In summary, RLHHF introduces a conceptual leap toward aligning AI systems with human physiological feedback, but the problems of interpretability, cross-cultural generalization, contextual disambiguation, and external manipulation remain open frontiers. Addressing them will require not only technical progress but also ethical, clinical, and sociotechnical research to ensure that systems learning from the human body remain transparent, inclusive, and secure.

17 Conclusion

RLHHF reframes alignment as an embodied optimization problem. By translating heart dynamics into mathematically usable rewards, we enable AI systems that learn to preserve and promote users' physiological coherence and emotional well-being.

References

[1] OpenAI. "Reinforcement Learning from Human Feedback," Technical Report, 2022.

[2] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[3] S. W. Porges. *The Polyvagal Theory: Neurophysiological Foundations of Emotions, Attachment, Communication, and Self-regulation*. Norton, 2011.

[4] R. McCraty. "Heart–Brain Communication and HRV Coherence," HeartMath Institute Monographs.

[5] A. Y. Ng, D. Harada, S. Russell. "Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping," ICML, 1999.