

HITL For High Risk AI Domains

Tyler Edwards	Sruthi Kuriakose	Subramanyam Sahoo	Shamith Achanta
<i>Independent</i>	<i>Independent</i>	<i>Independent</i>	<i>Independent</i>

Our product addresses the challenge of aligning AI systems with the legal, ethical, and policy frameworks of high-risk domains such as healthcare, defence, and finance. These sectors demand precise, context-aware decision-making to ensure compliance, safety, and accountability.

We recognise that this is not purely a technological problem. While we have the compute, data, and algorithms to develop powerful AI systems, many of these systems conflict with the legal and ethical frameworks that govern these industries. This disconnect presents a significant barrier to their safe and effective deployment.

We propose a flexible human-in-the-loop (HITL) system that acts as an interface between AI tooling and human operatives, ensuring AI outputs align with domain-specific regulatory and ethical standards while empowering users with ergonomically optimised decision support. By facilitating interactive refinement of AI behaviours and exposing decision making logic in auditable and human readable formats, our solution delivers real-time explainability, decision-level accountability, and minimised cognitive load, enabling experts to focus on critical insights and make informed decisions.

1 Problem overview

Machine Learning has been transformative in high-risk sectors offering unprecedented potential for innovation and efficiency. However, these sectors also demand rigorous compliance with legal, ethical, and policy frameworks. Misalignment between AI systems and these frameworks can lead to severe consequences, ranging from patient harm in healthcare to financial fraud or even geopolitical instability. Addressing this misalignment is critical to ensuring the safe, effective, and trustworthy adoption of AI in these contexts

1.1 Regulatory and Ethical Conflicts

While AI systems have provided immense value in various domains, they often operate as black boxes, lacking transparency in their decision-making processes. This opacity creates a fundamental disconnect with the legal and ethical requirements that demand:

1. **Traceability:** Decisions must be auditable and accountable.
2. **Explainability:** Outputs must be interpretable by human operatives
3. **Compliance:** Systems must adhere to industry-specific regulations such as ,GDPR , HIPAA, and IPA.

1.2 Real-World Case Studies Highlighting the Risks

AI systems that fail to provide clear reasoning for their predictions can jeopardize outcomes, the absence of clear, interpretable AI processes can lead to significant human costs:

1. **Healthcare:** In 2023, an AI-powered chatbot designed to assist with medical inquiries provided incorrect and potentially harmful advice to users. The system's lack of accurate medical knowledge and insufficient oversight led to misdiagnoses, highlighting the dangers of deploying AI in healthcare without rigorous validation and human supervision.[Dra]
2. **Defense:** In September 2024, the International Committee of the Red Cross highlighted concerns regarding AI decision support systems in military targeting. Instances were reported where AI misidentified civilians as combatants, leading to tragic consequences. These errors emphasize the critical need for human oversight and the limitations of AI in complex real-world military operations. []

1.3 Current Gaps in AI System Design

1. **Context Awareness:** AI systems struggle to learn nuanced understandings of domain-specific variables that human experts bring to critical decision-making.
2. **Human Integration:** Few systems are designed with interfaces that facilitate intuitive human-AI collaboration.
3. **Feedback Mechanisms:** Limited mechanisms exist for incorporating real-time user feedback to refine AI behaviours in situ.

2 Proposed solution

- a. **Reward Modelling:** A reward model is trained to capture human responses and align AI outputs with desired outcomes. Feedback is batched and aggregated to generate representative datasets, ensuring comprehensive coverage of user insights
- b. **Human Feedback Pipeline:** Clustering algorithms such as k-means and DBSCAN are used to identify similarities, while dimensionality reduction techniques like PCA and t-SNE detect patterns and trends. The aggregated feedback is analyzed to discover recurring themes and anomalies which aims to strip data of user preferences and focusing on trends in model performance.
- c. **Trend and Pattern Analysis:** Once feedback trends are identified, anomaly detection models pinpoint outliers that could signify critical issues. All feedback analysis is used to guide hybrid preference optimization (HPO) using Proximal Policy Optimization (PPO) to fine-tune the AI model by aligning it with the reward model, incorporating insights from the feedback analysis.
- d. **Explainability and Bias Mitigation:** Explainable AI (XAI) methodologies are integrated into the tooling output to provide a broader context and greater transparency in machine learning (ML) decision-making logic. By making complex decision pathways interpretable, XAI empowers users to trace, audit, and understand the rationale behind AI-driven conclusions. This transparency is critical for high-stakes domains, as it fosters trust, supports regulatory compliance, and reduces the risk of misinterpretation.

3 Pilot experiment or demo

To begin exploring our solution, we experimented with implementing our HITL (Human-in-the-Loop) framework in the medical domain. During the hackathon, we conducted a series of technical experiments to develop a Minimum Viable Product (MVP) and gain insights into Reinforcement Learning from Human Feedback (RLHF) and reward modelling on multimodal data.

3.1 HITL Bias Mitigation on Text-Based Data

We first focused on text-based datasets, specifically the COMPAS dataset, to examine how human feedback could be used to combat bias within predictive models. Using a human feedback mechanism, we implemented a system where predictions could be either accepted or rejected based on human review. This feedback was then integrated back into the model to adjust its parameters, thereby reducing bias in subsequent predictions. We created a Decision Consistency Score (DCS) which compares the model's predicted probabilities between pairs of demographic groups (e.g., the mean predicted probabilities for Group A vs. Group B) as a means of observing model bias for specific demographics.

3.2 Exploration of Human Feedback in LLMs for Medical QA

Following this, we conducted an exploration of human feedback aggregation in the medical QA domain using datasets such as ChatDoctor with the medical-qa-datasets. The goal was to annotate and utilize human feedback to improve the contextual accuracy and quality of model responses. Our approach focused on (1) identifying features of the inputs that led to the model's response and (2) developing a reward model tailored to evaluate response quality, which was fine-tuned using annotated feedback.

To understand the influence of different parts of the input (e.g., patient history or specific symptoms) on model predictions, we explored feature attribution methods such as Shapley values and input ablation. While these methods showed promise in approximating feature importance and isolating critical inputs, their application to large language models posed challenges that limited their effectiveness. We also began investigating advanced techniques, including Integrated Gradients (Captum) and Intervention Training (Pyvene), which we aim to incorporate in future iterations to enhance attribution precision.

3.3 Feedback Aggregation and Trend Analysis

The feedback processing pipeline is a modular framework designed to integrate user feedback into AI systems for improving performance, safety, and accountability. It leverages clustering, dimensionality reduction, anomaly detection, and trend analysis to extract actionable insights. To ensure the models remain generalised and are not swayed by user preferences, the feedback is aggregated and processed collectively to ensure we can identify systemic improvements needed in model responses, focusing on broad patterns rather than isolated preferences.

1. **Pattern Recognition and Clustering:** user feedback is collected, batched, and converted into numerical embeddings for analysis. Clustering algorithms such as K-means and DBSCAN identify patterns and groupings in the feedback. Dimensionality reduction techniques like PCA and t-SNE enable visual exploration of high-dimensional feedback, revealing underlying structures and trends.

4 Process

Timeframe	What will you do?
Next 3 months	Foundation and Pilot Testing: Continue technical validation of HITL approach by completing system architecture and exploring/evaluating applications across varied test cases. Collaborate with policymakers and researchers to align system with industry needs. Review and iterate.
2025	Deployment and Early-Stage Scaling: Build towards deployment in real-world applications by EOY to stress test and review product performance. Expand market engagement through strategic partnerships with security firms and AI governance/auditing bodies.
2026	Growth and Standardization: Broaden scope of deployment to domains such as logistics, pharma, and manufacturing, working with global organisations to establish international standards of AI accountability
2027	Horizon scanning: Develop a healthy commercial ecosystem with third-party developers to expand flexibility of tooling, continue investment into R&D to address emerging trends.

5 Impact on AI safety & key risks

5.1 Evaluate direct and indirect safety benefits

The proposed solution fosters a robust culture of responsibility and compliance, ensuring that AI systems are developed and deployed in alignment with ethical standards and regulatory requirements. This approach encourages organizations to prioritize safety-first AI practices that mitigate risks while enhancing operational reliability.

1. **Enhanced Decision Accuracy:** Real-time human feedback ensures AI outputs are aligned with context-specific needs, reducing the likelihood of errors in critical applications.
2. **Bias Mitigation:** By integrating human review in training and decision processes through XAI and feedback prorogation, the system actively reduces demographic biases, promoting fairness and equity. Showcasing model interpretability safeguards against potential harm to vulnerable groups.
3. **Accountability and Transparency:** The HITL framework provides auditable decision-making processes, enhancing trustworthiness in AI systems.
4. **Policy Compliance:** The framework ensures that all AI processes adhere to domain-specific regulations and standards, such as GDPR, HIPAA, or industry-specific guidelines.

5.2 Risks

1. Inefficient or incomplete feedback collection might lead to suboptimal model refinement.
2. Overfitting to specific feedback trends might compromise the model’s generalizability
3. Bias in human inputs could skew the AI system’s behavior, perpetuating existing prejudices
4. Regulatory landscapes may shift, rendering the system non-compliant.