AI-Instigated Human Oversight: Rethinking Human-in-the-Loop Safety in Clinical AI

Sera Singha Roy

Independent Researcher
sera.sroy@gmail.com

Abstract

The rise in clinical AI has helped to define the diagnostic and decision-making process in the dimension of modern medicine. However, their implementation is constrained by multiple ethical, interpretability, and safety factors that hinder the effective utilization of these AI systems. Existing Human-in-the-Loop (HITL) systems rely heavily on externally triggered human oversight, which creates critical gaps in effective safety and accountability. This study introduces the *AI-Instigated Human Oversight (AIHO)* framework, an AI governance architecture that enables models to self-assess uncertainty or contextual failure and autonomously escalate decisions to human oversight through a four-layer detection and escalation mechanism. Each layer performs a distinct self-assessment: (Layer 1) predictive uncertainty quantification, (Layer 2) contextual validation and explainability, (Layer 3) ethical and proxy alignment monitoring, and (Layer 4) adaptive governance with human-in-command enforcement. These layers form part of a three-zone operational architecture: (1) Information Flow and AI Core Processing, (2) AIHO Oversight Core, and (3) Human Escalation Loops. These zones in conjunction create a continuous feedback loop which administers continuous self-evaluation, transforming ethical and technical anomalies into actionable human oversight triggers. AIHO establishes a dynamic pathway toward regulation-ready, self-aware clinical AI, aligning with the current international standards for trustworthy and accountable AI in medicine. This paper presents the conceptual architecture and mathematical formalization of AIHO; empirical validation across clinical domains represents the critical next phase of this research.

Introduction

Healthcare has experienced remarkable advances in diagnostics, drug development, clinical trials, and personalized medicine due to the accelerated integration of Artificial Intelligence (AI) and Machine Learning (ML) technologies (Faiyazuddin et al. 2025) (Alowais et al. 2023). However, these integrations are explicitly related to ethical and safety challenges, as they can lead to serious patient harm, even death, in a high-risk medical environment (Agius et al. 2025) (Wehkamp et al. 2021).

The inclination towards these systems has created a core tension between maximizing automation efficiency and the need for human, ethical accountability, and transparency (Klarin, Ali Abadi, and Sharmelly 2024) (Valenzuela et al. 2024). This conflict is further fueled by the tendency of the most widely used ML models to make overconfident predictions, even when faced with Out-of-Distribution (OOD) data (Liu et al. 2023) (Guo et al. 2017) (Ran et al. 2022). The existing Human-in-the-loop safety paradigm provides critically insufficient ethical safeguards. The Automation Bias, in which clinicians rely more on the AI outcome even after having contradicting evidence, depicts that just placing a Human in the workflow pipeline does not guaranty safety (Khera, Simon, and Ross 2023). Additionally, ensuring that humans retain the authoritative decision-maker, i.e. Meaningful Human Control (MHC), is consistently challenged by the speed and opacity of autonomous systems (Abbink et al. 2024) (Chan et al. 2024), thus creating a knowledge gap.

These safety issues are intensified by the technical limitations of multiple ML models in clinical settings where they are unable to signal when they are uncertain of the outcome. For example, Deep Neural Networks lack the ability to quantify their confidence, often leading to incorrect overconfident predictions (Gawlikowski et al. 2023). Studies show that existing attempts to achieve proper human oversight through Explainable AI (XAI) often lead to ”ersatz understanding” of the model rather than helping to fully grasp its reasoning (Longo et al. 2024). This lacuna of self-assessment of these AI systems and their inability to explain safety failures effectively prompt the necessity for a mechanism for a proactive and internal AI self-monitoring mechanism to achieve ”meaningful human oversight”.

This study introduces the AI-Instigated Human Oversight (AIHO) framework which is based on motivation to move towards self-regulating AI systems having the capacity of possessing the ability to self-diagnose uncertainty, contextual failures, ethical concerns, and autonomously request human intervention when the decision-making is compromised. AIHO enables the AI system to monitor its own cognitive and operational limitations and trigger human oversight effectively. The key points addressed in this study are as follows:

1. The Conceptual Framework: Introduces the framework for AI-Instigated Human Oversight (AIHO) spanning

across technical, contextual, ethical, and governance dimensions.

2. Formalization: Defines mathematical trigger functions that translate internal diagnostics into explicit human oversight signals.
3. Implementation Blueprint: Outlines how AIHO can be implemented in clinical decision systems to ensure upcoming regulatory requirements for safety, transparency, and accountability.

Background and related-work

The integration of Artificial Intelligence and Machine Learning has brought remarkable advancements in the healthcare domain. However, incorporation of this transformation is also open to high risks, as uncertainty and error of the model's outcome can create ethical and accountability challenges and can even lead to serious patient harm (Botha et al. 2024) (Agius et al. 2025) (Alowais et al. 2023). The core challenge in the deployment of the ML model in the clinical domain is its black-box aspect, which makes it difficult for the user to understand the behind-the-scenes of the model (Marey et al. 2024) (Chan et al. 2024). As a mitigation strategy for this issue, the Explainable AI (XAI) has been under continuous development. Although, the current XAI has to some extent been able to provide some transparency and aid in trust calibration, but it has also faced certain challenges when it comes to human oversight. Previous studies show that clinicians often project Automation bias, where explainability causes over-reliance, thereby struggling to recognize incorrect recommendations or outcomes by the model (Khera, Simon, and Ross 2023). Post-hoc rationalization risks are also found that involve "ersatz understanding", which affects the clinicians understanding of the true causal reasoning of the model (Wysocki et al. 2023) (Longo et al. 2024). The continuous cognitive load imposed on clinicians by the use of XAI leads to skipping validations, over-reliance on the model, or even intuition-based decision-making, making this high-risk environment susceptible to major detriment (Herm 2023). Another technical capability that allows AI systems to self-assess and maintain transparency is Uncertainty Quantification (UQ), which is considered the "foundational signal for responsible human intervention" providing the primary trigger for human oversight (He et al. 2025). However, many of the most widely cited medical ML models still lack a mechanism to define an escalation or governance protocol for human intervention when identifying unreliable predictions (He et al. 2025).

Meaningful Human Control (MHC) is the principle that defines that humans, and not algorithms, should be ultimately responsible for making the final decision (Veluwenkamp 2022). Current oversight models span Human-in-command (HIC), Human-in-the-Loop (HITL), and Human-on-the-Loop (HOTL), which have regulatory frameworks mandating continuous monitoring and human override for high-risk applications (Laux 2024). However, the concept of MHC still remains fragile in healthcare, lacking the robust articulation necessary to guide human oversight beyond procedural checks. Existing HITL models have major challenges because they depend on humans to assess the model's output validity and correctness, which is often hindered by the immense work pressure, time limitations, and cognitive biases of the clinicians. Additionally, current HITL models are insufficient in rectifying algorithmic biases (Salloch and Eriksen 2024). For example, in AI systems used for population health management, algorithms trained on flawed proxy labels (e.g., predicting healthcare care cost) often reproduce and amplify racial bias by misrepresenting the true ground truth (health need) across demographic groups. Depending only on human judgment to correct these systemic errors is inefficient and incomplete, as humans may only partially identify and mitigate the bias embedded in the primary prediction mechanism.

The deep-rooted issues in traditional reactive HITL models, human susceptibility to automation bias, inadequate explanations, and the difficulty of detecting subtle OOD data or systemic bias under pressure (Langer, Baum, and Schlicker 2024), reveal a critical gap: the lack of a mechanism that autonomously requires human intervention when mandatory. AI-Instigated Human Oversight (AIHO) is proposed in this study to fill this gap by creating a formal and dynamic framework where the AI system itself initiates a structured human oversight process. By integrating UQ, OOD detection, and ethical alignment checks into a layered autonomous system, AIHO goes beyond the limitations of HITL and insufficient XAI to implement a fully, robust, and accountable model ensuring safe clinical AI.

Proposed Framework: AI-Instigated Human Oversight (AIHO)

Overview

The proposed AI-Instigated Human Oversight (AIHO) framework is built on the principle that the AI system must quantify its uncertainty and assess the integrity of its reasoning before escalating a decision to a human expert. It is structured as three zones as shown in Figure 1:

- *Zone 1 - Data and AI Core:* represents the operational backbone of the AI system consisting of Data Ingestion (continuous feed of multimodal data), AI Core Engine (the predictive model that outputs a prediction $\hat{y}$). The uncertainty and XAI Module (estimates predictive uncertainty U and computes local explainability signals).
- *Zone 2 - AIHO 4-Layer Oversight Core:* this is the heart of the AIHO framework, which defines the AI system's internal reasoning on when to signal for human intervention. The four self-assessing and escalating layers that progressively assess system trust, contextual validity, ethical alignment, and integrity of governance. Each layer performs continuous self-monitoring and translates technical or contextual deviations into explicit alerts for human oversight.
- *Zone 3 - Human Oversight Loops:* here each trigger connects to a distinct human or institutional authority:
  - L1 $\rightarrow$ Clinician / Front-line Expert (HITL): Validates or overrides uncertain predictions.

Layer	Focus / Component	Autonomous Detection Mechanism & Core Signal	Required Human Oversight / Action
1. Foundational Signal & Uncertainty Estimation (UE)	System reliability and predictive confidence.	Continuous quantification of epistemic (model ignorance) and aleatoric (data noise) uncertainty using Deep Ensembles or Conformal Prediction.	*Flag for Review:* If uncertainty exceeds threshold τ_1, the output is escalated for clinician review.
2. Contextual Validation & Explainability (XAI)	Reasoning coherence and domain adherence.	Detects Out-of-Distribution (OOD) input via VAEs or Local Lipschitz metrics; evaluates explanation stability using attribution robustness.	*Mandatory Abstention:* If input is OOD or explanation instability $> \tau_2$, the AI system must abstain and request human validation.
3. Ethical / Proxy Alignment Check	Systemic bias and goal integrity.	Monitors subgroup disparities and detects proxy objective conflicts that deviate from the clinical intent.	*Governance Review Trigger:* If ethical disparity or proxy conflict exceeds τ_3, human governance board review is mandated.
4. Adaptive Governance & Human-in-Command (HIC)	Trust calibration and accountability enforcement.	Monitors human-AI interactions for automation bias or governance non-compliance.	*Forced Override:* If flagged outputs are accepted without justification or context drift is detected, system enforces Human-in-Command intervention.

Table 1: Summary of the Four Layers of the AIHO Framework. Each layer represents an autonomous detection and escalation mechanism for human intervention.

- L2 → Expert Review Queue: Handles abstentions / OOD cases; adds contextual corrections.
- L3 → Governance Board / Ethics Committee: Reviews systemic bias or proxy conflicts, and mandates retraining when necessary.
- L4 → Regulatory Auditor / Safety Engineer (HIC): Executes a hard stop or system pause for high-risk cases.

Each of these represents increasing levels of human control. After human intervention, the corrected labels, contextual notes, or governance decisions flow back (dashed arrow from Figure 1) to the data pipeline, and the AIHO core re-calibrates thresholds τ_i, updates model weights, and logs oversight outcomes. This closes the loop in which the system continuously learns when to escalate, not just what to predict.

Core Layer Structure and Escalation Logic

The AIHO framework allows for a comprehensive evaluation of the internal and external state of the AI system. For an input x with prediction $\hat{y}$ and confidence $p(\hat{y}|x)$, each layer defines a trigger function $T_i(x)$ that signals potential unreliability.

Layer 1: Foundational Signal and Uncertainty Estimation Layer 1 of the AIHO framework focuses on the AI system's ability to self-assess its confidence by quantifying epistemic uncertainty, i.e., the model's ignorance because of its limitations in its training data (He et al. 2025) and aleatoric (data noise) uncertainty (Meijerink, Cinà, and Tonutti 2020) (Hüllermeier and Waegeman 2021). Using techniques such as Deep Ensembles or Conformal Prediction, the model continuously calculates its predictive uncertainty for every output (Rezaei et al. 2023) (Yang and Li 2023). If the uncertainty exceeds a predefined threshold, the system triggers a "Flag for review" indicating the necessity for a human expert intervention to validate the output before utilizing it in making clinical decisions. This layer is the first line-of defense against known model limitations.

$$T_1(x): \; H(p(\hat{y} \,|\, x)) > \tau_1, \tag{1}$$

where $H(\cdot)$ is the entropy of the predictive distribution and τ_1 is a calibrated uncertainty threshold. If the entropy is high, indicating diffuse confidence across classes, the model flags the case for review.

Layer 2: Contextual Validation and Explainability (XAI) Audit While Layer 1 assesses *whether the model is confident*, Layer 2 *asks why it should be*, which means that Layer 2 focuses on providing contextual validation by detecting Out-Of-Distribution (OOD) inputs using methods such as Variational Autoencoders (VAEs) or Local Lipschitz metrics (Ran et al. 2022) (Bhutto et al. 2024) (Yang et al. 2024). The system's response to an OOD flag is a Mandatory Abstention, where it denies providing a prediction, and explicitly communicates its lack of knowledge. This layer also includes an "explainability audit", where the system can check for unstable explanations, preventing the propagation of misleading explanations.

$$T_2(x): \; \mathrm{OOD}(x) > \tau_2 \;\vee\; \mathrm{XAI}_{\mathrm{instab}}(x) > \tau_2', \tag{2}$$

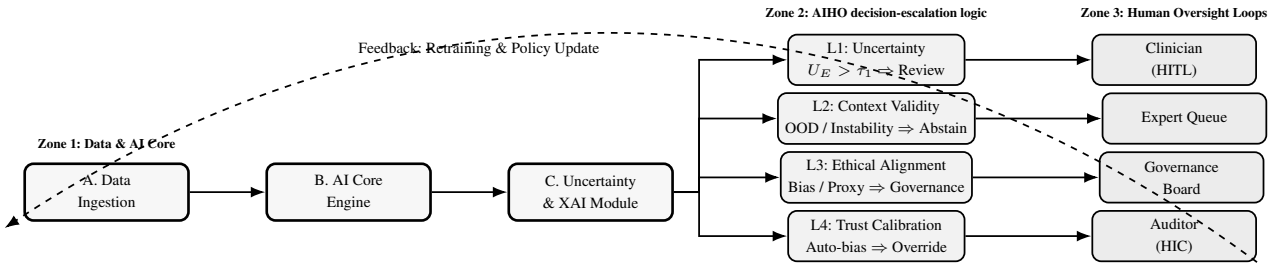

Figure 1: Conceptual architecture of the AI-Instigated Human Oversight (AIHO) framework showing three operational zones: (1) Data AI Core; (2) AIHO 4-Layer Oversight Core; and (3) Human Escalation Loops

where $\mathrm{OOD}(x)$ quantifies Out-Of-Distribution deviation (e.g., via the VAE reconstruction error or Local Lipschitz metrics), τ_2 is the threshold for $\mathrm{OOD}(x)$ above which the model is deemed to operate outside its validated data regime, $\mathrm{XAI}_{\mathrm{instab}}(x)$ measures the instability of the explanation under small perturbations, and τ_2' is the threshold for $\mathrm{XAI}_{\mathrm{instab}}(x)$ beyond which explanation coherence is considered unreliable. Exceeding either threshold implies invalid context or unstable reasoning, prompting abstention and human validation.

Layer 3: Ethical Alignment and Proxy Integrity Check This layer addresses the ethical dimension of the AI system by monitoring systemic bias or goal misalignment, specifically, the vicious failure mode where the system's objective function (a proxy label like cost) systematically conflicts with the desired ethical outcome (a true goal like health equity). For example, a model trained to predict healthcare costs may inadvertently assign lower risk scores to Black patients because they historically receive less expensive care, even if they suffer from a more severe ailment (Obermeyer et al. 2019). The alert for this layer is not only for a single clinician, but also for an ethics or governance board, triggering the need to retrain the model, make proxy adjustments or policy interventions. Therefore, upon identifying significant disparities, this then triggers a "Governance Review" signal.

$$T_3(x): \ \mathrm{Disparity}(x) > \tau_3, \tag{3}$$

where $\mathrm{Disparity}(x)$ captures subgroup performance gaps (e.g., difference in false-negative rates) and τ_3 denotes the acceptable fairness margin. Violations initiate an ethical review at the governance-level.

Layer 4: Adaptive Governance and Human-in-Command (HIC) This final layer completes the loop by monitoring the human cognitive biases such as Automation Bias, thereby assessing the human-AI interaction for signs of hazardous miscalibration. For instance, it can detect if a clinician repeatedly accepts flagged, high-uncertainty recommendations without proper validation, thereby enforcing the appropriate oversight level (eg, switching from HOTL to HIC). This layer mandates a Human Override or System Re-evaluation, compelling a full review or requiring a second expert's authorization in order to prevent a potentially harmful clinical decision. Layer 4 ensures that AIHO is not only a decision-support framework but also an effective ethical governance system that enforces accountability and safe use, ultimately keeping the human truly in command.

$$T_4(x): \ [\,T_1(x) \vee T_2(x)\,] \wedge \neg \mathrm{HumanOverride}(x), \tag{4}$$

which activates when lower-layer alerts ($T_1(x)$ and $T_2(x)$) are ignored by a human operator, enforcing a mandatory Human-in-Command (HIC) override.

Unified Oversight Logic
The unified escalation policy is defined as follows:

$$\mathrm{Intervention}(x) = T_1(x) \vee T_2(x) \vee T_3(x) \vee T_4(x), \tag{5}$$

indicating that any triggered layer suffices to escalate the decision for human oversight.

Escalation Severity and Architecture The lower layers trigger internal alerts, and the upper layers enforce mandatory human action. The hierarchical nature of AIHO ensures that not all alerts require the same response intensity as depicted in Table 2. The escalation mapping function $f(\cdot)$ allows real-time prioritization of human oversight, therefore balancing both safety and efficiency.The severity of escalation is defined as follows:

$$\mathrm{EscalationLevel}(x) = f(\,\max_i\{i \mid T_i(x) = 1\}\,), \tag{6}$$

where $f(\cdot)$ maps the highest activated layer index to an escalation type:

Layer	Severity	Action
1	Minor	Suggest review (soft flag)
2	Moderate	Abstain / defer to human
3	Major	Governance or ethics review
4	Critical	Human-in-command override

Table 2: Escalation severity and required human intervention levels.

From the conceptual flow diagram 1, the data flow (Zone 1) feeds the AI core and the uncertainty module, which

self-assesses performance through the AIHO 4-Layer self-assessment and escalation system (Zone 2), and the triggered layers route cases to the corresponding human oversight agents (Zone 3). The dashed feedback loop captures retraining and policy refinement for continuous governance. Layer 4 supersedes all others, ensuring ultimate safety even when lower-level warnings are ignored.

The architecture ensures that technical uncertainty, contextual deviation, and ethical misalignment are automatically signaled to human operators through an interpretable escalation mechanism. This multi-layered oversight design provides a structured pathway from model self-awareness to accountable human collaboration.

Implementation Blueprint

The **AI-Instigated Human Oversight (AIHO)** framework is designed to operate as a lightweight supervisory wrapper around existing clinical AI models, requiring minimal modification of the underlying inference code. The architecture follows a modular plug-in pattern, ensuring compatibility with deep learning systems already deployed for imaging, EHR, or multimodal inference tasks.

Pipeline Integration The pipeline Integration as depicted in the Listing 1 is described as follows:
Step 1: Inference Pipeline. A standard AI model produces predictions $\hat{y}$ and, if available, confidence scores $p(\hat{y} \mid x)$.
Step 2: AIHO Monitor. The AIHO wrapper intercepts the output, computes diagnostic signals—uncertainty, OOD distance, fairness metrics and user interaction logs, and evaluates the trigger functions T_1-T_4 defined in Layer Structure section.
Step 3: Output Router. Based on the activated trigger, the monitor dynamically routes the prediction to automatic release (safe zone) or an appropriate human escalation loop (L1-L4).
Step 4: Human Interface. The selected human agent (clinician, review board, or auditor) receives a structured case summary, including model output, uncertainty score, and explanation rationale.
Step 5: Governance Log. All interventions and results are recorded in a continuous audit log, creating a transparent trace for regulatory and ethical compliance.

Design Principles AIHO operates as an *outer shell* rather than a replacement layer, maintaining modularity, explainability, and low coupling. Implementation can be performed through REST-based middle-ware or lightweight Python decorators, which do not require modification to the internal model weights. This ensures minimal codebase intrusion and seamless integration with existing clinical deployment pipelines.
Model Inference → AIHO Monitor → Output Router → Human Interface → Governance Log

Threshold Calibration and Validation Protocol The effectiveness of AIHO's escalation mechanism depends on appropriate threshold calibration for each layer. These thresholds (τ_1, τ_2, τ_3) are domain-specific and must be empirically determined through systematic validation.

Listing 1: Pseudocode sketch for AIHO integration

```
def aiho_wrapper(model, x):
    y_hat, prob = model.predict(x)
    signals = compute_signals(x, y_hat,
        prob)
    triggers = evaluate_triggers(signals
        )

    if triggers.any():
        route_to_human(triggers, x,
            y_hat, signals)
    else:
        auto_release(y_hat)

    log_governance(x, y_hat, triggers,
        signals)
    return y_hat
```

Layer 1 (Uncertainty Threshold τ_1): Modern deep neural networks are systematically miscalibrated, producing overconfident predictions even when incorrect (Guo et al. 2017). Apply temperature scaling to reduce Expected Calibration Error prior to threshold setting. Using a held-out calibration dataset with expert-annotated prediction errors, construct ROC curves plotting uncertainty scores $H(p(\hat{y}|x))$ against true model failures. Select τ_1 to achieve 85-95% sensitivity for error detection while maintaining clinically acceptable escalation rates (typically 8-15% of cases).

Layer 2 (OOD Threshold τ_2): Calibrate using controlled distribution shifts representative of the deployment environment. Set thresholds based on the 95th percentile of validation distribution scores: $\tau_2 = \text{Percentile}_{95}(\text{OOD\_score}_{\text{validation}})$. The target performance should achieve FPR95 (false positive rate at 95% true positive rate) below 5% to minimize unnecessary abstentions while maintaining safety (Liu et al. 2020).

Layer 3 (Fairness Threshold τ_3): Establish acceptable disparity margins through stakeholder consultation with ethics boards and experts in the clinical domain. Benchmark against established fairness metrics (e.g. equalized odds, calibration fairness) appropriate to the clinical use case (Obermeyer et al. 2019). Use bootstrap resampling to generate confidence intervals for subgroup performance gaps. Set τ_3 conservatively to trigger the governance review for statistically significant disparities.

Layer 4 (Human-in-Command Enforcement): Layer 4 does not require threshold calibration in the traditional sense, as it monitors behavioral patterns rather than model outputs. It activates when lower-layer alerts (T_1 or T_2) are triggered, but human operators fail to provide documented overrides, thus enforcing mandatory Human-in-Command intervention to prevent automation bias.

Validation and Implementation Strategy: Initial deployment should operate in "shadow mode", logging all trigger activations without blocking clinical workflow for

3-6 months prior to active deployment (Food, Administration et al. 2021). This enables empirical calibration of alert rates and iterative threshold adjustment through real-world feedback. Site-specific recalibration using local validation data is essential, as identical models can exhibit dramatically different performance across institutions (Wong et al. 2021).

Discussion

The proposed AI-Instigated Human Oversight (AIHO) framework in this study represents a pivotal evolution in the governance and human oversight of clinical AI systems. The AIHO framework goes beyond passive Human-in-the-Loop (HITL) methods by embedding the proactive self-assessment and escalation loop directly in it, thereby ensuring safety, accountability, and trust. This discussion contextualizes the AIHO framework by contrasting it with existing oversight methods, aligning it with global regulatory principles, examining its practical benefits, and understanding its challenges.

From Reactive Oversight to Proactive Self-Monitoring

Existing oversight models or concepts depend on the vigilance of humans to identify and intercept AI system failures, which poses an immense cognitive burden on them making them prone to automation bias. Explanation alone has been shown to be insufficient to mitigate these risks, often leading to "ersatz understanding", promoting a false sense of security amongst clinicians, and failing to support them in identifying errors (Jones et al. 2024) (Longo et al. 2024). The additional challenge of AI models being an opaque tool (Abbink et al. 2024) (Chan et al. 2024), has fueled the need to constantly second-guess the reasoning behind the model results by an external human clinician.

AI-Instigated Human Oversight (AIHO) framework redefines the relationship between clinicians and AI systems by redefining oversight as a **bidirectional pathway** where AI systems not only help in decision-making, but also autonomously trigger human oversight when an anomaly is detected through continuous self-monitoring. By autonomously detecting high uncertainty (Layer 1), Out-of-Distribution input (Layer 2), emergent systemic bias (Layer 3), or dangerous patterns of human interaction (Layer 4), the AI system instigates human intervention precisely at the moments of highest risk. This makes the AI system an active participant in its own governance transforming oversight from a reactive, human-led intervention into a proactive, system-instigated safety protocol, ensuring that human judgment is applied when needed.

Aligning AIHO with Global Regulatory Principles

The AIHO framework provides a direct technical implementation of the high-level principles established by global regulators as depicted in Table 3, offering an auditable and enforceable structure establishing trustworthy AI systems.

This alignment shows that AIHO is a practical engineering and governance framework rather than only an ethical idea which is designed to meet the legal and ethical demands of AI in the healthcare domain.

Benefits

The implementation of the AIHO framework offers multiple benefits for clinical practice and AI governance, such as:

- **Enhances Clinician Trust:** Under this framework, the AI system signals when it is uncertain or working with unfamiliar data, making it trustworthy from the perspective of clinicians. The clinicians can rely on the AI system's predictions rather than being concerned about the reliability and explainability of the outcomes.
- **Reduces Automation Bias and Detects Silent Failures:** By mandating human oversight for high-risk, flagged outputs (Layer 4), AIHO compels a "reflect and review" moment countering automation bias. Its continuous monitoring for data distribution shift (Layer 2) and performance degradation across subgroups (Layer 3) provides an early warning system for silent model failures that would otherwise go undetected until the occurrence of any harm.
- **Enables Real-Time Compliance Auditing:** Whenever the AI system under the AIHO framework triggers an intervention, it generates a log detailing the signal (e.g., high uncertainty), the context (e.g., patient data), and the resulting human action (e.g., override with justification). This creates an automated, real-time audit trail that can be used to demonstrate regulatory compliance with oversight requirements to governance bodies such as the FDA or European authorities (Joshi et al. 2024) (Act 2024) (Pollard, Ryan, and Mohanty 2022).

Challenges and Future Directions

Despite having multiple benefits, the implementation of the AIHO framework in real-world systems poses multiple challenges necessitating careful consideration, review, and further research. The effectiveness of Layer 1, Layer 2, and Layer 3 of the AIHO framework depends on the precise calibration of the activation thresholds (τ_i) which will vary significantly on a case-basis. For example, a threshold for uncertainty in a low-risk dermatology screening tool will differ significantly from that for a life-critical sepsis prediction model in the ICU. Since these thresholds are not universal and must be empirically validated for each case in their clinical domain and risk context, a considerable amount of research and ongoing adjustment are necessary. This leads to our next challenge when an AI system is poorly calibrated generating an excessive number of flags, thereby creating alert fatigue where clinicians ignore warning signals as a consequence of an excessive cognitive load (Wan et al. 2020). To be effective, the intervention signal must be accompanied by a clear, concise and actionable explanation rather than relying only on complex or non-robust XAI methods, which often result in ersatz understanding amongst clinicians (Longo et al. 2024).

Although this work establishes the theoretical foundations and architectural blueprint for AIHO, comprehensive empirical validation encompasses the critical next phase. Future research must address: (1) Multi-site clinical trials

Regulation / Guideline	Core Oversight Principle	AIHO Correspondence
EU AI Act (Act 2024)	Meaningful human oversight to prevent or minimize risks to health and safety.	**Layers 3 & 4:** Layer 3 (*Ethical Alignment Check*) triggers governance review for systemic issues, while Layer 4 (*Adaptive Governance*) enforces appropriate human control (HIC/HITL/HOTL) based on risk, ensuring accountability remains with humans.
FDA Good Machine Learning Practice (GMLP) (Food, Administration et al. 2021) (Pollard, Ryan, and Mohanty 2022)	Risk-proportionate oversight emphasizing transparency, performance monitoring, and post-market surveillance.	**Layers 2 & 4:** Layer 2 (*Contextual Validation*) supports continuous monitoring for data distribution shifts post-deployment, while Layer 4 ensures oversight proportional to risk. The full AIHO stack enables traceable, auditable compliance with regulatory review processes.
WHO Ethics Guidance (Guidance 2021)	Foster responsibility and accountability through mechanisms for oversight and redress when individuals are adversely affected by algorithmic decisions.	**Layers 2, 3 & 4:** AIHO embeds accountability by enforcing abstention under uncertainty (Layer 2), mandating governance review for detected bias (Layer 3), and maintaining a "human warranty" through enforced clinician responsibility in high-risk contexts (Layer 4).

Table 3: Alignment of the AIHO Framework with Key Regulatory and Ethical Guidelines. Each regulation emphasizes human accountability, which AIHO operationalizes through its multi-layered escalation mechanisms.

to validate AIHO's effectiveness in reducing automation bias and improving clinical outcomes across diverse healthcare settings; (2) Threshold calibration studies to establish domain-specific best practices for τ_1, τ_2, and τ_3 in clinical specialties (e.g., radiology, oncology, emergency medicine); (3) Human factors research to evaluate clinician cognitive load, alert fatigue, and trust calibration when interacting with AI-instigated oversight; (4) Longitudinal monitoring to assess AIHO's performance under real-world dataset drift and evolving clinical practices; and (5) Comparative studies measuring safety improvements and error reduction rates versus traditional HITL approaches. In addition, investigating trade-offs between escalation rates and safety margins, as well as the development of standardized AIHO deployment guidelines for regulatory compliance, will be essential for widespread clinical adoption. These empirical investigations will transform AIHO from a conceptual framework into a validated and deployable governance architecture for trustworthy clinical AI.

The AIHO framework must be seamlessly integrated into existing clinical workflows and EHR systems, which will require significant interdisciplinary collaboration among AI developers, clinicians, IT staff, and hospital administrators to improve existing systems. Beyond technical integration, the focus on interface design and workflow compatibility of human-AI interactions are critical to ensure that AI-instigated interventions are seamless, intuitive, and genuinely supportive of clinical decision-making.

Conclusion

The implementation of safe and trustworthy AI systems in the clinical ecosystem has faced persistent challenges such as automation bias, algorithmic opacity, and the reactive nature of existing oversight mechanisms. The AI-Instigated Human Oversight (AIHO) framework presents a proactive paradigm for guiding human oversight in clinical AI by guiding AI systems to identify and communicate their own limitations as well as governance issues. This is achieved by embedding self-assessment mechanisms at four layers from uncertainty quantification to ethical governance as the core of the architecture, therefore transforming the static, human-led oversight into an adaptive, self-initiated intervention triggers. Through this, AIHO paves a robust pathway to achieve Meaningful Human Control (MHC) and also aligns directly with the governance principles of prominent regulations such as the EU AI Act and FDA GMLP (Act 2024) (Pollard, Ryan, and Mohanty 2022) (Geraci et al. 2025). Thus, AIHO framework provides a foundational step towards trustworthy, regulation-ready, and human-centered AI systems for clinical decision support that are engineered to uphold the primary clinical mandate of ensuring "patient safety above automation".

References

Abbink, D.; Amoroso, D.; Siebert, L. C.; van den Hoven, J.; Mecacci, G.; and de Sio, F. S. 2024. Introduction to meaningful human control of artificially intelligent systems. In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*, 1–11. Edward Elgar Publishing.

Act, E. A. I. 2024. The eu artificial intelligence act. *European Union*.

Agius, S.; Cassar, V.; Bezzina, F.; and Topham, L. 2025. Leveraging digital technologies to enhance patient safety. *Health and Technology*, 1–11.

Alowais, S. A.; Alghamdi, S. S.; Alsuhebany, N.; Alqahtani, T.; Alshaya, A. I.; Almohareb, S. N.; Aldairem, A.; Alrashed, M.; Bin Saleh, K.; Badreldin, H. A.; et al. 2023.

Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1): 689.

Bhutto, D. F.; Zhu, B.; Liu, J. Z.; Koonjoo, N.; Li, H. B.; Rosen, B. R.; and Rosen, M. S. 2024. Uncertainty estimation and out-of-distribution detection for deep learning-based image reconstruction using the local lipschitz. *IEEE Journal of Biomedical and Health Informatics*, 28(9): 5422–5434.

Botha, N. N.; Segbedzi, C. E.; Dumahasi, V. K.; Maneen, S.; Kodom, R. V.; Tsedze, I. S.; Akoto, L. A.; Atsu, F. S.; Lasim, O. U.; and Ansah, E. W. 2024. Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. *Archives of Public Health*, 82(1): 188.

Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; et al. 2024. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 958–973.

Faiyazuddin, M.; Rahman, S. J. Q.; Anand, G.; Siddiqui, R. K.; Mehta, R.; Khatib, M. N.; Gaidhane, S.; Zahiruddin, Q. S.; Hussain, A.; and Sah, R. 2025. The impact of artificial intelligence on healthcare: a comprehensive review of advancements in diagnostics, treatment, and operational efficiency. *Health Science Reports*, 8(1): e70312.

Food, U.; Administration, D.; et al. 2021. Good machine learning practice for medical device development: guiding principles. *FDA webpage*.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589.

Geraci, J.; Rao, P.; Grandinetti, C.; Qorri, B.; Nadolny, P.; Ayalew, K.; Bregnhøj, L.; Edwards, L.; Hofmann, K.; Khozin, S.; et al. 2025. Current Opportunities for the Integration and Use of Artificial Intelligence and Machine Learning in Clinical Trials: Good Clinical Practice Perspectives. *Journal of the Society for Clinical Data Management*, 5(2).

Guidance, W. 2021. Ethics and governance of artificial intelligence for health. *World Health Organization*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

He, W.; Jiang, Z.; Xiao, T.; Xu, Z.; and Li, Y. 2025. A Survey on Uncertainty Quantification Methods for Deep Learning. arXiv:2302.13425.

Herm, L.-V. 2023. Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study. arXiv:2304.08861.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506.

Jones, M. D.; Liu, S.; Powell, F.; Samsor, A.; Ting, F. C. R.; Veliotis, N.; Wong, Y. M.; Franklin, B. D.; and Garfield, S. 2024. Exploring the role of guidelines in contributing to medication errors: a descriptive analysis of national patient safety incident data. *Drug Safety*, 47(4): 389–400.

Joshi, G.; Jain, A.; Araveeti, S. R.; Adhikari, S.; Garg, H.; and Bhandari, M. 2024. FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics*, 13(3): 498.

Khera, R.; Simon, M. A.; and Ross, J. S. 2023. Automation bias and assistive AI: risk of harm from AI-driven clinical decision support. *Jama*, 330(23): 2255–2257.

Klarin, A.; Ali Abadi, H.; and Sharmelly, R. 2024. Professionalism in artificial intelligence: The link between technology and ethics. *Systems Research and Behavioral Science*, 41(4): 557–580.

Langer, M.; Baum, K.; and Schlicker, N. 2024. Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*, 35(1): 1.

Laux, J. 2024. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & society*, 39(6): 2853–2866.

Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2023. Towards Out-Of-Distribution Generalization: A Survey. arXiv:2108.13624.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Del Ser, J.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; et al. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106: 102301.

Marey, A.; Arjmand, P.; Alerab, A. D. S.; Eslami, M. J.; Saad, A. M.; Sanchez, N.; and Umair, M. 2024. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. *Egyptian Journal of Radiology and Nuclear Medicine*, 55(1): 183.

Meijerink, L.; Cinà, G.; and Tonutti, M. 2020. Uncertainty estimation for classification and risk prediction on medical tabular data. arXiv:2004.05824.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pollard, V. T.; Ryan, M. W.; and Mohanty, A. 2022. FDA Issues Good Machine Learning Practice Guiding Principles. *The Journal of Robotics, Artificial Intelligence & Law*, 5.

Ran, X.; Xu, M.; Mei, L.; Xu, Q.; and Liu, Q. 2022. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. *Neural Networks*, 145: 199–208.

Rezaei, M.; Näppi, J. J.; Bischl, B.; and Yoshida, H. 2023. Bayesian uncertainty estimation for detection of long-tailed and unseen conditions in medical images. *Journal of Medical Imaging*, 10(5): 054501–054501.

Salloch, S.; and Eriksen, A. 2024. What are humans doing in the loop? Co-reasoning and practical judgment when using machine learning-driven decision aids. *The American Journal of Bioethics*, 24(9): 67–78.

Valenzuela, A.; Puntoni, S.; Hoffman, D.; Castelo, N.; De Freitas, J.; Dietvorst, B.; Hildebrand, C.; Huh, Y. E.; Meyer, R.; Sweeney, M. E.; et al. 2024. How artificial intelligence constrains the human experience. *Journal of the Association for Consumer Research*, 9(3): 241–256.

Veluwenkamp, H. 2022. Reasons for meaningful human control. *Ethics and Information Technology*, 24(4): 51.

Wan, P. K.; Satybaldy, A.; Huang, L.; Holtskog, H.; and Nowostawski, M. 2020. Reducing Alert Fatigue by Sharing Low-Level Alerts With Patients and Enhancing Collaborative Decision Making Using Blockchain Technology: Scoping Review and Proposed Framework (MedAlert). *J Med Internet Res*, 22(10): e22013.

Wehkamp, K.; Kuhn, E.; Petzina, R.; Buyx, A.; and Rogge, A. 2021. Enhancing patient safety by integrating ethical dimensions to Critical Incident Reporting Systems. *BMC medical ethics*, 22(1): 26.

Wong, A.; Otles, E.; Donnelly, J. P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penoza, C.; et al. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine*, 181(8): 1065–1070.

Wysocki, O.; Davies, J. K.; Vigo, M.; Armstrong, A. C.; Landers, D.; Lee, R.; and Freitas, A. 2023. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, 316: 103839.

Yang, C.-I.; and Li, Y.-P. 2023. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15(1): 13.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12): 5635–5662.