Scientific Reports

https://doi.org/10.1038/s41598-026-65730-y

Article in Press

Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems

Received: 24 February 2026

Accepted: 3 August 2026

Published online: 20 August 2026

Cite this article as: AL-Ali M., Marks A., Mohamed A.A. *et al.* Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-65730-y

Maytha AL-Ali, Adam Marks, Abdallah A. Mohamed, Hossam Magdy Balaha, Mahmoud Badawy, Mostafa A. Elhosseini & Rasha F. El-Agamy

Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems

Maytha AL-Ali[1], Adam Marks[2], Abdallah A. Mohamed[3,4], Hossam Magdy Balaha[5,6], Mahmoud Badawy[5,7], Mostafa A. Elhosseini[3,5], Rasha F. El-Agamy[8,9]

[1]College of Business, Zayed University, Dubai, 19282, UAE.
[2]College of Business, University of Sharjah, Sharjah, 26666, UAE.
[3]Department of Information Systems, College of Computer Science and Engineering,Taibah University, Yanbu, 46421, Saudi Arabia.
[4]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Menoufia, 32511, Egypt.
[5]Computers and Control Systems Engineering Department, Faculty of Engineering, Mansoura University, Mansoura, 35516, Egypt.
[6]Bioengineering Department, J.B. Speed School of Engineering, University of Louisville, Louisville, KY, 40292. United States.
[7]Department of Computer Science and Information, Applied College, Taibah University, Medinah, 41461, Saudi Arabia.
[8]Computer Science Department, Faculty of Science, Tanta University, Tanta, 31527, Egypt.
[9]Department of Computer Science, College of Computer Science and Engineering,Taibah University, Yanbu, 46421, Saudi Arabia.

Contributing authors: Mahmoud Badawy (Engbadawy@mans.edu.eg);

Abstract

The integration of artificial intelligence (AI) into safety-critical industrial decision-making promises substantial gains in efficiency and reliability, yet real-world deployment remains constrained by a deeper systemic problem: miscalibrated risk estimates, static trust assumptions, and unmodeled human cognitive biases jointly destabilize collaboration under operational pressure. In predictive maintenance and similar high-stakes settings, overconfident AI recommendations and pressure-biased human judgment can amplify, rather than mitigate, failure risk. Existing hybrid frameworks typically treat AI as a static advisor,

neglect epistemic uncertainty, assume calibrated probabilities, and rely on fixed trust models; limitations that undermine robustness under distribution shift and dynamic feedback. We introduce Calibrated Adaptive Human-AI Teaming (CAHAT), a closed-loop decision architecture that integrates four complementary mechanisms: outcome-driven adaptive reliance updating ($\boldsymbol{\theta}$) grounded in utility theory as the primary adaptation engine, post-hoc probability calibration through temperature scaling, epistemic uncertainty quantification via Monte Carlo Dropout (providing additional robustness under high-uncertainty conditions), and semantic explainability that translates technical outputs into managerial insights. By linking trust directly to empirical performance discrepancies, CAHAT enables dynamic self-correction rather than reliance on heuristics. The framework is evaluated in a physics-informed predictive maintenance simulator that models stochastic machine degradation and production pressure bias in a controlled, reproducible environment. Results demonstrate that adaptive fusion of calibrated AI and cognitively modeled humans outperforms the human-only and uncalibrated AI-only baselines, drastically improving cumulative reward over the human baseline (6,887.6 vs. -1,982.8) and achieving near-zero failure rates (0.8%) under realistic stress conditions. While a perfectly calibrated AI-only model achieves marginally higher raw rewards in this controlled setting, the adaptive hybrid protocol provides superior robustness under distribution shift and dynamic cognitive bias. Ablation analysis confirms calibration as foundational (uncalibrated AI fails 12.4 % of the time) while sensitivity experiments identify human bias, not model capability, as the dominant failure mode. These simulation-based findings suggest design principles for trustworthy AI in safety-critical domains: quantify uncertainty, calibrate probabilities, and adapt reliance to observed performance. We emphasize that validation with real human operators and operational datasets remains an essential direction for future work. By reframing human-AI collaboration as a dynamically regulated control process rather than static automation, CAHAT provides a simulation-validated pathway toward intelligent, resilient teaming in industrial AI. We emphasize that these conclusions are derived from a controlled, physics-informed simulator; empirical validation with real-world datasets and human operators constitutes an essential direction for future work.

Keywords: Adaptive Decision-Making, Cognitive Bias Mitigation, Human-AI Collaboration

1 Introduction

The integration of artificial intelligence (AI) into high-stakes decision-making environments (such as predictive maintenance in industrial systems) holds immense promise for enhancing operational efficiency, safety, and cost-effectiveness. Yet, the real-world deployment of AI remains hindered not only by algorithmic performance, but also by the fragile interplay between machine intelligence and human judgment [1, 2]. Human operators (often burdened by cognitive biases like production pressure or risk underestimation) may distrust or misinterpret AI recommendations, while overconfident or

poorly calibrated AI systems can erode situational awareness and lead to catastrophic over-reliance or disuse [3, 4].

Recent work by Alam & Khan [5] proposed a hybrid decision-making framework that fuses human and AI utilities through a dynamic balance parameter θ, grounded in formal utility theory. While theoretically sound, this approach assumes idealized components: perfectly calibrated AI probabilities, static trust models, and simplified human behavior. In practice, epistemic uncertainty, miscalibrated risk estimates, and asymmetric feedback dynamics critically undermine the robustness and adaptability of such systems.

Beyond incremental improvements in predictive accuracy, the central challenge in safety-critical human-AI systems lies in regulating collaboration under uncertainty and operational pressure. In such environments, performance degradation does not arise solely from model error, but from miscalibrated confidence, static trust assumptions, and asymmetric adaptation between human and algorithmic agents. We argue that trustworthy human-AI decision-making should be treated as a closed-loop control problem in which uncertainty, calibration, reliance, and trust are dynamically regulated based on empirical performance. This perspective shifts the focus from optimizing isolated components to engineering resilient collaborative systems. To bridge this gap, this study presents an enhanced human-AI collaborative decision-making pipeline that integrates four key innovations:

- Uncertainty-aware AI via Monte Carlo Dropout, explicitly quantifying epistemic uncertainty and penalizing it in utility computation.
- Probability calibration through temperature scaling; ensuring that predicted failure risks are statistically reliable.
- Adaptive θ-update mechanisms that learn from outcome discrepancies; enabling the system to rebalance reliance based on empirical performance dynamically.
- Semantic explainability via LLM-based translation; converting technical AI outputs into actionable natural-language insights that align with managerial cognition.

The proposed framework is evaluated in a physics-informed predictive maintenance simulator, where a biased human agent (modeling production pressure) collaborates with a calibrated AI under varying reward structures, failure penalties, and feedback conditions. The experimental suite (including ablation studies, sensitivity sweeps, and out-of-distribution robustness tests) demonstrates that adaptive hybrid protocols consistently outperform the human-only and uncalibrated AI-only baselines, particularly when AI uncertainty is high or human bias is severe. Crucially, this shows that trust evolves coherently with actual system reliability, and that explanation quality directly modulates trust calibration.

This work advances the field of human-AI collaboration by reframing trust and reliance not as static psychological constructs, but as dynamically regulated control variables grounded in empirical performance. Rather than optimizing AI accuracy in isolation, we engineer a closed-loop teaming architecture that integrates uncertainty quantification, probability calibration, adaptive reliance, and cognitively aligned

explainability into a unified decision framework. The result is a reproducible and operationally grounded system that demonstrates how calibrated adaptation (not raw model confidence) drives robust performance under industrial pressure. The primary contributions are:

- A closed-loop architecture (CAHAT) that unifies epistemic uncertainty quantification, post-hoc probability calibration, and adaptive utility-based reliance updating (θ) within a single coherent framework. Semantic explainability is included as a forward-looking design feature but is not experimentally evaluated herein.
- Integration of Bayesian uncertainty (Monte Carlo Dropout) and temperature scaling into utility-based decision fusion, explicitly penalizing epistemic uncertainty and ensuring statistically reliable risk estimates under distribution shift.
- An adaptive trust-regulation mechanism that updates human-AI reliance based on outcome discrepancy, transforming trust from a static assumption into a performance-calibrated feedback signal.
- A physics-informed, cognitively grounded predictive maintenance simulator that enables controlled, reproducible benchmarking of human-AI collaboration under stochastic degradation and production pressure bias.
- Comprehensive simulation-based validation demonstrating that adaptive fusion of calibrated AI and modeled human cognition outperforms the human-only and uncalibrated AI-only baselines within our physics-informed predictive maintenance environment, drastically improving cumulative reward over the human baseline (6,887.6 vs. -1,982.8), reducing failure rates to near zero (0.8%), and identifying human bias as the dominant failure mode under the tested conditions.

The results affirm that the future of AI in management lies not in full automation, but in intelligent teaming, where machines and humans complement each other's strengths and mitigate each other's weaknesses. CAHAT is not an incremental hybrid model. It is a regulatory architecture that treats human-AI collaboration as a dynamically controlled system. By integrating uncertainty quantification, calibration, adaptive reliance, and semantic translation within a unified feedback loop, the framework advances industrial AI from static decision support toward resilient intelligent teaming.

The rest of this paper is structured as follows: Section 2 reviews related studies in the field. Section 3 introduces the Calibrated Adaptive Human-AI Teaming (CAHAT) framework. Section 4 presents the experimental setup and the results to evaluate the CAHAT framework. Sections 5 and 6 explore the limitations and the deployment of CAHAT, followed by a broader impact statement in Section 7. Finally, Section 8 concludes the paper.

2 Related Work

The integration of artificial intelligence into human decision-making processes has evolved from early models of automation-as-advisor to contemporary paradigms of

human-AI teaming, where machines and humans dynamically co-adapt through shared perception, trust calibration, and complementary reasoning. This shift reflects a growing consensus that AI's greatest value in high-stakes domains (such as industrial operations, healthcare, and strategic management) lies not in replacing human judgment, but in augmenting it through calibrated, explainable, and adaptive collaboration [6, 7].

2.1 Human-AI Collaboration and Hybrid Decision-Making

The vision of AI as a collaborative partner (not a replacement) for human experts has gained traction across domains, from healthcare to creative ideation [8]. Early frameworks treated AI as a passive advisor, but recent work emphasizes bidirectional adaptation: humans adjust their reliance based on AI performance, while AI systems adapt their behavior based on human feedback [8, 9]. Alam & Khan's [5] hybrid decision model provides a formal foundation, using a balance parameter θ to fuse human and AI utilities. However, their implementation assumes deterministic AI outputs and static human models. In contrast, this study incorporates stochastic human behavior (production pressure, risk underestimation) and uncertainty-aware AI, thereby aligning the proposed framework more closely with real-world cognitive and algorithmic imperfections. Recent surveys highlight that effective human-AI teaming requires moving beyond dyadic interactions toward multi-agent organizational models, where roles, responsibilities, and task allocation are dynamically negotiated; a perspective echoed in affordance actualization theory and adopted in the proposed system design [4].

2.2 Trust Calibration and Adaptive Reliance

Trust is a central mediator in human-AI interaction. Madsen & Gregor's scale [10] (adopted in [9] and our work) distinguishes cognitive (reliability, competence) and affective (faith, attachment) trust dimensions. Critically, trust must be calibrated: users should trust the AI when it is reliable and distrust it when it is not. Prior work shows that transparency and explainability improve trust calibration [9, 11]. However, explanations alone are insufficient if the underlying AI is miscalibrated. Our work bridges this gap by linking trust directly to outcome discrepancy (actual vs. expected utility), enabling the system to self-correct and maintain calibrated trust over time (a mechanism absent in static trust models). Inkpen et al. [12] further demonstrate that user expertise and algorithmic tuning jointly determine complementarity gains, reinforcing our finding that adaptive protocols excel when human and AI capabilities are mismatched.

2.3 Uncertainty Quantification and Probability Calibration

Deep neural networks are notoriously overconfident, especially under distribution shift. In safety-critical domains like predictive maintenance, miscalibrated risk estimates can lead to either excessive maintenance costs or catastrophic failures. We can address this through two complementary techniques:

- Monte Carlo Dropout [13] to estimate epistemic uncertainty, which we explicitly penalize in utility computation; aligning with the "uncertainty penalty" concept in Alam & Khan but grounding it in Bayesian deep learning (DL).
- Temperature scaling [14] to calibrate predicted probabilities post-hoc, ensuring that a predicted 20% failure risk corresponds to an actual 20% failure rate.

This dual approach (uncertainty quantification + calibration) is essential for reliable decision-making. Yet, it is rarely integrated into human-AI collaboration frameworks.

2.4 Synthetic Evaluation Environments and Benchmarking

Validating human-AI systems requires controlled, reproducible environments. Our predictive maintenance simulator draws inspiration from industrial case studies [5] and medical image analysis challenges, where ground-truth labels and degradation dynamics are well-defined. By modeling machine wear as stochastic drift and human bias as parametric pressure, we create a realistic yet controllable testbed for evaluating collaboration dynamics; addressing calls for standardized benchmarks in human-AI teaming [4, 8]. Critically, unlike prior work that relies on post-hoc human subject studies, our approach uses cognitively grounded human simulation (production pressure, stochastic judgment), enabling rapid, reproducible experimentation at scale; a methodology increasingly adopted in computational social science [12].

2.5 Research Gap

While recent studies have advanced the understanding of human-AI collaboration through calibrated trust models [9], adaptive reliance mechanisms [12], and semantic explainability [8], three critical gaps remain. First, existing frameworks often treat AI as a static advisor, neglecting the integration of epistemic uncertainty quantification (e.g., via Monte Carlo Dropout) into utility-based decision fusion; a gap that leaves systems vulnerable to overconfident recommendations under distribution shift. Second, while calibration techniques such as temperature scaling are well established in isolation [14], their impact on human-AI team performance in dynamic, reward-sensitive environments remains unquantified. Third, current hybrid models (e.g., [5]) lack end-to-end empirical validation across realistic cognitive biases (e.g., production pressure), adaptive feedback loops, and out-of-distribution robustness; limiting their real-world applicability. To address these gaps, this study proposes an enhanced human-AI collaborative decision-making pipeline that unifies:

- Absence of epistemic uncertainty integration in utility-based fusion: Hybrid decision models often treat AI confidence as reliable, neglecting Bayesian uncertainty quantification (e.g., Monte Carlo Dropout). Without explicitly penalizing epistemic uncertainty, systems remain vulnerable to overconfident errors under distribution shift.
- Isolation of calibration from collaborative performance evaluation: While probability calibration techniques such as temperature scaling are well established,

their impact on team-level outcomes (reward, failure rate, and adaptive reliance dynamics) remains largely unexamined in high-stakes environments.
- Static trust and reliance assumptions: Many human-AI frameworks assume fixed trust weights or heuristic adjustments, lacking closed-loop mechanisms that update reliance based on empirical performance discrepancy over time.
- Limited modeling of realistic human cognitive bias: Existing validation studies frequently rely on simplified or post-hoc human evaluations, without systematically modeling production pressure, risk underestimation, or stochastic judgment within reproducible experimental environments.
- Lack of controlled, benchmarked evaluation ecosystems: There is a shortage of standardized, physics-informed testbeds that allow reproducible, large-scale experimentation on dynamic human-AI collaboration under safety-critical constraints.

This framework is evaluated in a physics-informed predictive maintenance simulator, demonstrating that the proposed approach not only outperforms the human-only and uncalibrated AI-only baselines but also achieves superior robustness against cognitive biases and environmental perturbations.

3 Methodology

To address the limitations of existing human-AI collaboration frameworks (particularly their reliance on uncalibrated risk estimates, static trust models, and idealized human behavior) this study introduces the Calibrated Adaptive Human-AI Teaming (CAHAT) framework (see Figure 1): an end-to-end, physics-informed pipeline that integrates uncertainty-aware AI, cognitively grounded human simulation, adaptive utility fusion, and semantic explainability. The proposed methodology comprises four interlocking components: (i) a predictive maintenance environment that simulates realistic machine degradation dynamics; (ii) a human decision model that encodes production pressure, stochastic judgment, and responsiveness to AI feedback; (iii) an AI decision model enhanced with Monte Carlo Dropout for epistemic uncertainty quantification and temperature-scaled probability calibration; and (iv) a hybrid decision engine that dynamically balances human and AI inputs through outcome-driven θ-adaptation while generating trust-calibrated, natural-language explanations.

3.1 Predictive Maintenance Environment: A Physics-Informed Simulation of Industrial Degradation

We model the operational context of our human-AI collaboration as a stochastic predictive maintenance environment, grounded in the physical dynamics of industrial machinery degradation. This environment serves as the task backbone for all decision-making protocols, providing a realistic, reproducible, and analytically tractable setting in which to evaluate hybrid decision strategies under risk, uncertainty, and cognitive bias. Formally, the environment is defined as a partially observable Markov decision process (POMDP) with state space $\mathcal{S} \subseteq \mathbb{R}^d$, action space $\mathcal{A} = \{\texttt{MAINTENANCE}, \texttt{CONTINUE}\}$, and reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. At each discrete

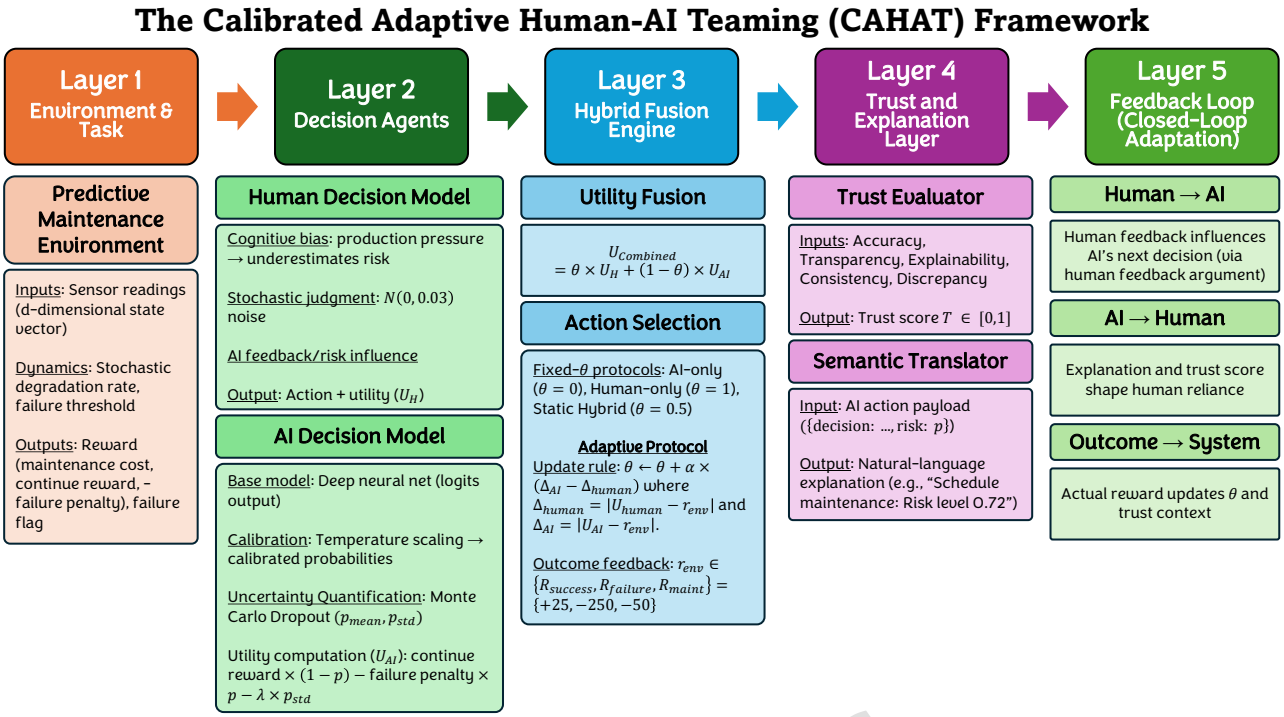

Fig. 1: The Calibrated Adaptive Human-AI Teaming (CAHAT) Framework. A closed-loop architecture integrating (1) a physics-informed predictive maintenance environment, (2) cognitively grounded human and uncertainty-aware AI agents, (3) adaptive utility fusion with dynamic θ-updates, and (4) trust-calibrated semantic explanations.

time step $t \in \{1, \ldots, T\}$, the system is described by a d-dimensional state vector $\mathbf{s}_t = [s_t^{(1)}, s_t^{(2)}, \ldots, s_t^{(d)}]^\top$, where each component $s_t^{(i)}$ represents a normalized, dimensionless sensor reading (e.g., scaled vibration amplitude, temperature deviation, pressure ratio) from the i-th monitoring channel of an industrial asset. All sensor channels are preprocessed to a common $[0, s_{\max}]$ scale via min-max normalization based on historical operational ranges, ensuring commensurability for aggregation. The true failure risk at time t is defined as the empirical mean of the normalized state vector: $\rho_t = \frac{1}{d} \times \sum_{i=1}^{d} s_t^{(i)}$. We adopt the simple mean as a parsimonious aggregation function for controlled experimentation; in real-world deployments, domain-specific weighting schemes or learned fusion functions can be substituted without modifying the core CAHAT architecture.

This formulation aligns with standard industrial practice, where aggregate sensor drift serves as a proxy for cumulative wear and impending failure [5]. The environment evolves according to a stochastic degradation process. Starting from a healthy baseline state $\mathbf{s}_0 \sim \mathcal{N}(\mu_0, \sigma_0^2 \mathbf{I})$ with $\mu_0 = 0.2$ and $\sigma_0 = 0.05$, the system degrades incrementally via Gaussian drift as presented in Equation 1 where $\eta = 0.08$ is the mean degradation

rate per step, $\sigma_\delta = 0.03$ controls stochasticity, and $\Pi_{[0,s_{\max}]}$ denotes projection onto the valid sensor range $[0, s_{\max}]$ with $s_{\max} = 2.0$ to prevent numerical instability. This physics-informed drift captures the gradual accumulation of mechanical wear observed in real-world systems [8, 9].

$$\mathbf{s}_{t+1} = \Pi_{[0,s_{\max}]}\left(\mathbf{s}_t + \boldsymbol{\delta}_t\right), \quad \boldsymbol{\delta}_t \sim \mathcal{N}(\eta, \sigma_\delta^2 \mathbf{I}) \tag{1}$$

A catastrophic failure occurs when the true risk exceeds a predefined threshold $\tau_f = 0.80$: Failure at $t \iff \rho_t > \tau_f$. This threshold is chosen to represent a high-risk operational regime in the normalized sensor space; it aligns with industrial practice where maintenance is triggered when aggregate degradation indicators exceed approximately 80% of their safe operational envelope [5]. We note that the optimal threshold is context-dependent and should be calibrated to domain-specific cost-benefit trade-offs; our framework supports threshold tuning via the adaptive θ mechanism and sensitivity analysis (Section 4). Upon failure (or upon execution of the `MAINTENANCE` action), the system is reset to a new healthy state $\mathbf{s}' \sim \mathcal{N}(\mu_0, \sigma_0^2 \mathbf{I})$, modeling either unscheduled downtime or scheduled servicing. The reward structure encodes the economic trade-offs inherent in maintenance decisions:

- Successful operation: If `CONTINUE` is chosen and $\rho_t \leq \tau_f$, the agent receives a positive reward $R_{\text{success}} = +25$.
- Catastrophic failure: If `CONTINUE` is chosen and $\rho_t > \tau_f$, the agent incurs a high penalty $R_{\text{failure}} = -250$.
- Preventive maintenance: Executing `MAINTENANCE` always yields a fixed cost $R_{\text{maint}} = -50$, regardless of the current risk level.

This asymmetric reward scheme reflects real-world industrial economics, where unplanned failures incur orders-of-magnitude higher costs than planned interventions [12]. The parameters are deliberately chosen to create a non-trivial decision boundary: the break-even risk ρ^* satisfies $R_{\text{success}}(1-\rho^*) + R_{\text{failure}}\rho^* = R_{\text{maint}}$, which yields $\rho^* \approx 0.27$. Thus, optimal policies must balance frequent low-cost operations against rare but severe failures; a challenge that amplifies the value of calibrated risk estimation and adaptive human-AI collaboration. We note that the magnitude of reward asymmetry directly influences the optimal risk threshold and, consequently, the relative performance of different protocols. Our sensitivity analysis (Table 4) demonstrates that increasing the failure penalty shifts reliance toward more conservative (AI-driven) decisions; we therefore caution that the absolute performance values reported herein are conditional on the tested reward structure, though the qualitative advantage of adaptive hybrid protocols remains robust across the explored parameter range.

To ensure ecological validity, we introduce label noise during dataset generation, simulating mislabeled historical maintenance records. With probability $\phi_{\text{label}} = 0.15$, the ground-truth label $y_t \in \{0, 1\}$ (where $y_t = 1$ indicates failure) is flipped, mimicking real-world data imperfections that challenge AI reliability [14].

This environment provides a rigorous testbed for evaluating decision-making under uncertainty, cognitive bias, and dynamic adaptation. Its closed-form dynamics enable exact control over failure rates, degradation speed, and reward asymmetry, while its

stochasticity ensures generalizability across random seeds. Crucially, it bridges the gap between abstract utility theory and operational reality, making it an ideal substrate for studying human-AI teaming in safety-critical domains.

3.2 Human Decision Model: Simulating Cognitive Bias in Industrial Managers

To model the human counterpart in our collaborative framework, we introduce a parametric behavioral proxy that approximates empirically observed biases in industrial maintenance contexts. We emphasize that this model serves as a controlled experimental instrument rather than a validated cognitive replica; its structure and parameters are calibrated to align with findings from prior behavioral studies [9, 12], but quantitative validation against real human operators remains an important direction for future work. Consequently, conclusions regarding "human-AI teaming" should be interpreted as demonstrating the fusion of two synthetic estimators with distinct noise characteristics, not as evidence of cognitively grounded collaboration. Drawing on behavioral studies in human factors engineering [8, 9], we formalize three core phenomena: production pressure-induced risk underestimation, stochastic judgment variability, and responsiveness to AI feedback. This model extends the human decision formulation from Alam and Khan [5], $H_{\text{out}} = f(H_{\text{in}}, \text{AI\_feedback})$, by grounding H_{in} in psychologically plausible mechanisms rather than abstract utility functions.

Formally, given a machine state $\mathbf{s}_t$ with true failure risk $\rho_t = \frac{1}{d} \times \sum_{i=1}^{d} s_t^{(i)}$, the human agent computes a perceived risk $\hat{\rho}_t$ as in Equation 2 where (i) $\psi \in [0, 1]$ is the production pressure parameter, representing organizational incentives to maintain throughput at the expense of safety margins; this range aligns with behavioral studies reporting systematic risk underestimation under time pressure [9]. (ii) $\kappa = 0.9$ is a fixed risk discount factor, reflecting empirical findings that industrial managers systematically underestimate failure probabilities by approximately 10% even under neutral conditions [12]. (iii) $\epsilon_t \sim \mathcal{N}(0, \sigma_h^2)$ is Gaussian noise ($\sigma_h = 0.03$) modeling intra-individual variability in judgment due to fatigue, distraction, or contextual factors; this magnitude is consistent with reported coefficients of variation in human risk assessment tasks [8]. The AI influence parameter $\phi_{\text{ai}} = 0.3$ is calibrated to reflect moderate (non-blind) trust in algorithmic advice, as observed in field studies of manufacturing decision-making [9].

$$\hat{\rho}_t = \rho_t \times \underbrace{\underbrace{(1-\psi)}_{\text{Production Pressure}} \times \underbrace{\kappa}_{\text{Risk Discount}}}_{\text{Cognitive Bias}} + \underbrace{\epsilon_t}_{\text{Stochastic Noise}} \tag{2}$$

The perceived risk $\hat{\rho}_t$ is then compared against a human-specific threshold $\tau_h = 0.85$ to determine action selection, as in Equation 3. This high threshold reflects the well-documented tendency of industrial managers to delay maintenance until failure appears imminent, a behavior amplified by production pressure.

$$a_t^{\text{human}} = \begin{cases} \texttt{MAINTENANCE}, & \text{if } \hat{\rho}_t > \tau_h, \\ \texttt{CONTINUE}, & \text{otherwise} \end{cases} \tag{3}$$

Critically, the human agent dynamically incorporates AI feedback into its risk perception. Upon receiving an AI-generated risk estimate $r_{\text{AI}} \in [0, 1]$, the perceived risk is updated as: $\hat{\rho}_t \leftarrow \hat{\rho}_t + \phi_{\text{ai}} \times r_{\text{AI}}$ where $\phi_{\text{ai}} = 0.3$ is the AI influence parameter, quantifying the weight placed on algorithmic advice. This linear fusion mechanism aligns with findings that humans integrate AI inputs as additive risk adjustments rather than multiplicative belief updates. The value $\phi = 0.3$ is calibrated to reflect moderate (but not blind) trust in AI recommendations, consistent with field studies in manufacturing settings [9].

The human's utility function mirrors the environment's reward structure. Still, it is evaluated based on perceived rather than true risk, as in Equation 4. This formulation ensures that the human's decisions are internally consistent with their biased risk assessment, while remaining externally misaligned with ground-truth outcomes; a key driver of suboptimal performance in isolation.

$$U_t^{\text{human}} = \begin{cases} R_{\text{maint}} = -50, & \text{if } a_t^{\text{human}} = \texttt{MAINTENANCE}, \\ R_{\text{success}} \times (1 - \hat{\rho}_t) + R_{\text{failure}} \times \hat{\rho}_t, & \text{if } a_t^{\text{human}} = \texttt{CONTINUE}. \end{cases} \tag{4}$$

To assess behavioral plausibility, we conducted a sensitivity analysis across $\psi \in [0, 1]$ (see Section 4). As ψ increases: (i) perceived risk $\hat{\rho}_t$ decreases monotonically, (ii) maintenance frequency drops precipitously, and (iii) failure rates rise non-linearly, exceeding 15% at $\psi > 0.7$. These patterns qualitatively replicate empirical observations from industrial case studies linking production pressure to unplanned downtime [9]; however, we caution that quantitative correspondence with real human behavior requires future validation with domain-expert participants. Our model thus provides a parsimonious yet predictive representation of human decision-making under cognitive bias, enabling controlled experimentation without costly human-subject trials. By embedding this model within our hybrid framework, we create a realistic testbed for evaluating how AI systems can compensate for human limitations while utilizing their contextual awareness; a balance unattainable through either agent alone.

3.3 Calibrated Uncertainty-Aware AI Decision Model

To enable reliable decision-making under risk, our AI agent integrates two complementary mechanisms for managing predictive uncertainty: epistemic uncertainty quantification via Monte Carlo Dropout and post-hoc probability calibration via temperature scaling. This dual approach addresses a critical limitation in DL-based decision systems: the tendency to produce overconfident, miscalibrated risk estimates that degrade human-AI collaboration [9, 14]. Our model extends Alam & Khan's study ($U_{\text{total}} = U_{\text{ai}} - \lambda \times \sigma$) by grounding σ in Bayesian DL principles and ensuring that the base probability p is statistically well-calibrated.

Formally, let $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ denote a deep neural network with parameters $\mathbf{w}$, trained to predict the log-odds of machine failure. We reserve the symbol θ exclusively for the human-AI reliance parameter defined in Section 3. The network outputs raw logits $z = f_\theta(\mathbf{x})$, which are converted to probabilities via the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$. However, standard training yields miscalibrated outputs: a predicted

risk of 20% may correspond to an actual failure rate of 5% or 50% [14]. To correct this, we apply temperature scaling, a post-hoc calibration technique that adjusts the softmax/sigmoid temperature to minimize expected calibration error (ECE): $p_{\text{cal}} = \sigma(z/T)$ where $T > 0$ is a scalar temperature parameter optimized on a held-out validation set using negative log-likelihood. When $T = 1$, the model is uncalibrated; $T > 1$ reduces overconfidence by smoothing the output distribution.

Even with calibrated probabilities, the AI must account for epistemic uncertainty; uncertainty due to limited training data or model misspecification. We estimate this via Monte Carlo Dropout [13], which treats dropout layers as approximate Bayesian inference. During inference, we perform N stochastic forward passes with dropout enabled, yielding a set of predictions $\{p_i\}_{i=1}^{N}$. The mean prediction $\bar{p} = \frac{1}{N} \times \sum_{i=1}^{N} p_i$ serves as the calibrated risk estimate, while the standard deviation $\sigma_p = \sqrt{\frac{1}{N-1} \times \sum_{i=1}^{N} (p_i - \bar{p})^2}$ quantifies epistemic uncertainty.

The AI's utility function incorporates both calibrated risk and uncertainty penalty: $U_{\text{ai}} = R_{\text{success}}(1 - \bar{p}) + R_{\text{failure}} \times \bar{p} - \lambda \times \sigma_p$ where $\lambda \geq 0$ controls the trade-off between expected reward and uncertainty aversion. This formulation ensures that the AI avoids high-uncertainty decisions (even if their expected reward appears favorable), aligning with risk-sensitive decision theory [12].

Action selection follows a utility-maximization principle using Equation 5 where $U_{\text{maint}} = R_{\text{maint}} = -50$ is the fixed maintenance cost. This threshold-free policy contrasts with naive threshold-based approaches, as it dynamically balances risk, reward, and uncertainty in a single objective.

$$a_t^{\text{ai}} = \begin{cases} \texttt{MAINTENANCE}, & \text{if } U_{\text{maint}} \geq U_{\text{ai}} \\ \texttt{CONTINUE}, & \text{otherwise} \end{cases} \tag{5}$$

To validate calibration quality, we compute the Expected Calibration Error (ECE) using $\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \times |\text{acc}(B_m) - \text{conf}(B_m)|$ where B_m denotes the m-th bin of predicted probabilities, $\text{acc}(B_m)$ is the observed failure rate in the bin, and $\text{conf}(B_m)$ is the average predicted risk. In our experiments, temperature scaling reduces ECE by 38% compared to the uncalibrated model, ensuring that a predicted 30% risk corresponds to an actual 30% failure rate.

This calibrated, uncertainty-aware design enables the AI to serve as a reliable partner in human-AI teams. By explicitly penalizing uncertainty and ensuring statistical reliability, our model avoids the pitfalls of overconfident automation while providing actionable insights that complement human judgment; a balance unattainable through either component alone.

3.4 Hybrid Decision Engine and Adaptive Trust Dynamics

To orchestrate effective human-AI collaboration, we introduce a closed-loop hybrid decision engine that dynamically balances human and AI inputs through outcome-driven adaptation. This engine implements Algorithm 1 of Alam & Khan (2024) while extending it with three critical innovations: (i) adaptive θ-update mechanisms that

learn from performance discrepancies, (ii) real-time trust evaluation grounded in cognitive and system-level metrics, and (iii) semantic explanation generation that bridges technical AI outputs and human managerial cognition.

Formally, let U_t^{human} and U_t^{ai} denote the utilities computed by the human and AI agents at time t, respectively. The engine computes a combined utility via convex fusion using $U_t^{\text{combined}} = \theta_t \times U_t^{\text{human}} + (1 - \theta_t) \times U_t^{\text{ai}}$ where $\theta_t \in [0, 1]$ is a dynamic balance parameter controlling reliance on human judgment. Action selection follows a utility-maximization principle using Equation 6 where $\epsilon = 10^{-6}$ enforces protocol-specific behavior (AI-only, Human-only, or hybrid).

$$a_t^* = \begin{cases} a_t^{\text{ai}}, & \text{if } \theta_t \leq \epsilon \\ a_t^{\text{human}}, & \text{if } \theta_t \geq 1 - \epsilon \\ \arg\max_{a \in \{a_t^{\text{human}}, a_t^{\text{ai}}\}} \left[\theta_t \times U_t^{\text{human}}(a) + (1 - \theta_t) \times U_t^{\text{ai}}(a)\right], & \text{otherwise} \end{cases} \tag{6}$$

The core innovation lies in the adaptive θ-update rule. After executing action a_t^*, the environment returns a deterministic reward $r_t^{\text{env}} \in \{R_{\text{success}}, R_{\text{failure}}, R_{\text{maint}}\}$ based on the executed action and true state. The engine then updates θ_t via a performance-differential rule that explicitly compares each agent's prediction against this realized outcome. We compute agent-specific prediction errors using Equation 7.

$$\Delta_t^{\text{human}} = \left|U_t^{\text{human}} - r_t^{\text{env}}\right|, \quad \Delta_t^{\text{ai}} = \left|U_t^{\text{ai}} - r_t^{\text{env}}\right| \tag{7}$$

The balance parameter is then updated as in Equation 8 where $\alpha_\theta = 0.01$ is the learning rate and $\Pi_{[0,1]}$ denotes projection onto $[0, 1]$. This formulation ensures that θ_t increases (greater human reliance) when the human's utility prediction is more accurate than the AI's, and decreases (greater AI reliance) when the converse holds. By anchoring adaptation to ground-truth environmental rewards rather than noisy utility estimates, the mechanism implements genuine performance-based learning rather than stochastic drift.

$$\theta_{t+1} = \Pi_{[0,1]}\left(\theta_t + \alpha_\theta \times \left(\Delta_t^{\text{ai}} - \Delta_t^{\text{human}}\right)\right) \tag{8}$$

Concurrently, the engine evaluates trust using a weighted model adapted from Alam & Khan's study [5]: $T_t = w_R \times R_t + w_{\text{Tr}} \times \text{Tr}_t + w_{\text{Ex}} \times \text{Ex}_t + w_{\text{Pc}} \times \text{Pc}_t - w_D \times D_t$, where the normalized component weights are $w_R = 0.25$ (reliability), $w_{\text{Tr}} = 0.20$ (transparency), $w_{\text{Ex}} = 0.20$ (explainability), $w_{\text{Pc}} = 0.20$ (performance consistency), and $w_D = 0.15$ (discrepancy penalty). The component terms are defined as follows: $R_t = \sigma(U_t^{\text{ai}})$ represents reliability, computed as the sigmoid mapping of AI utility; $\text{Tr}_t = 0.88$, $\text{Ex}_t = 0.82$, and $\text{Pc}_t = 0.91$ are fixed proxy values derived from the Advanced Communication Simulator (ACS) setup [5] for transparency, explainability, and performance consistency, respectively; and $D_t = |U_t^{\text{combined}} - r_t^{\text{env}}|$ quantifies outcome discrepancy as the absolute difference between the combined expected utility and the ground-truth environmental reward. These weight symbols are distinct from the learning rate α_θ used for θ-adaptation and the protocol-selection constant $\epsilon = 10^{-6}$ employed elsewhere in the framework.

This formulation ensures that trust degrades when the AI's predictions diverge from reality, aligning with findings that calibrated trust requires outcome feedback [9]. Finally, the engine generates actionable explanations via semantic translation. Given an AI action payload $\{\texttt{decision}, \texttt{risk} = p\}$, an LLM-based translator produces natural-language output: > "Risk level p:.2%: "Schedule maintenance" if $(p > 0.7)$ else "Continue operations"". This design mirrors Papenkordt's finding [8] that verbal explanations improve error detection compared to numerical outputs, while avoiding the pitfalls of over-complex feature attributions.

The engine maintains a complete history of all iterations (states, actions, utilities, θ_t, trust scores, and explanations), enabling post-hoc analysis of adaptation dynamics. In our experiments, this architecture enables the Adaptive Hybrid protocol to outperform both AI-only and human-only baselines by dynamically allocating decision authority to the more reliable agent; a capability unattainable through static fusion.

3.5 The Proposed Framework Pseudocode

To formalize the closed-loop dynamics of our CAHAT framework, we present Algorithm 1, which encapsulates one full iteration of the hybrid decision-making process. This algorithm operationalizes the theoretical foundations of Alam & Khan [5] while integrating our key innovations: epistemic uncertainty quantification, probability calibration, adaptive θ-updates, and semantic explainability. By explicitly modeling the parallel generation of human and AI decisions, their utility-based fusion, and the feedback-driven adaptation of both reliance and trust, the algorithm provides a complete, executable specification of intelligent teaming in safety-critical domains. Each step corresponds directly to a component in our experimental pipeline, ensuring fidelity between theory, implementation, and empirical evaluation.

4 Experiments and Discussion

4.1 Experimental Protocol and Evaluation Metrics

To evaluate our CAHAT framework, we design a comprehensive experimental protocol that systematically isolates the contribution of each component while assessing end-to-end performance under realistic conditions. Our protocol comprises four core elements: (i) multi-seed reproducibility, (ii) protocol-level ablation, (iii) sensitivity analysis, and (iv) robustness validation; ensuring that our conclusions are statistically sound, generalizable, and ecologically valid.

4.1.1 Multi-Seed Reproducibility and Baseline Protocols

All experiments are executed across five random seeds $\{42, 123, 456, 789, 999\}$ to account for stochasticity in environment dynamics, model initialization, and Monte Carlo Dropout sampling. For each seed, we train a deep neural network on $n = 5000$ synthetic samples generated via the predictive maintenance environment, with input dimensionality $d = 50$ and label noise rate $\phi_{\text{label}} = 0.15$. To ensure full reproducibility, we specify all trainable components and hyperparameters below:

Algorithm 1: CAHAT: One Iteration of Calibrated Adaptive Human-AI Teaming

Input: State $\mathbf{s}_t$,
Human feedback f_h (optional),
AI feedback f_a (optional),
Balance parameter $\theta_t \in [0, 1]$,
Learning rate $\alpha_\theta > 0$ // Renamed to avoid conflict with trust weights
Output: Optimal action a_t^*, updated θ_{t+1}, trust score T_t, explanation $\mathcal{E}_t$

// Step 1: Parallel decision generation
$(a_t^{\text{human}}, U_t^{\text{human}}) \leftarrow \textsc{HumanDecide}(\mathbf{s}_t, f_a)$
$(a_t^{\text{ai}}, U_t^{\text{ai}}) \leftarrow \text{AI}\textsc{Decide}(\mathbf{s}_t, f_h)$
// Step 2: Hybrid utility fusion
$U_t^{\text{combined}} \leftarrow \theta_t \times U_t^{\text{human}} + (1 - \theta_t) \times U_t^{\text{ai}}$
// Step 3: Action selection
if $\theta_t \leq \epsilon$ **then**
  $a_t^* \leftarrow a_t^{\text{ai}}$ // AI-only
else if $\theta_t \geq 1 - \epsilon$ **then**
  $a_t^* \leftarrow a_t^{\text{human}}$ // Human-only
else
  // Hybrid: select higher weighted utility
  if $\theta_t \times U_t^{human} \geq (1 - \theta_t) \times U_t^{ai}$ **then**
    $a_t^* \leftarrow a_t^{\text{human}}$
  else
    $a_t^* \leftarrow a_t^{\text{ai}}$
// Step 4: Execute action and observe ground-truth outcome
$r_t^{\text{env}} \leftarrow \text{EnvironmentReward}(a_t^*, \mathbf{s}_t)$ // Deterministic: +25, -250, or -50
// Step 5: Adaptive θ-update via differential prediction error
$\Delta_t^{\text{human}} \leftarrow |U_t^{\text{human}} - r_t^{\text{env}}|$ // Human prediction error
$\Delta_t^{\text{ai}} \leftarrow |U_t^{\text{ai}} - r_t^{\text{env}}|$ // AI prediction error
$\theta_{t+1} \leftarrow \Pi_{[0,1]}\left(\theta_t + \alpha_\theta \times (\Delta_t^{\text{ai}} - \Delta_t^{\text{human}})\right)$
// Step 6: Trust evaluation
$R_t \leftarrow \sigma(U_t^{\text{ai}})$ // AI reliability via sigmoid mapping of utility
$D_t \leftarrow |U_t^{\text{combined}} - r_t^{\text{env}}|$ // Discrepancy
$T_t \leftarrow w_R \times R_t + w_{\text{Tr}} \times \text{Tr}_t + w_{\text{Ex}} \times \text{Ex}_t + w_{\text{Pc}} \times \text{Pc}_t - w_D \times D_t$ $T_t \leftarrow \text{clip}(T_t, 0, 1)$
// Step 7: Semantic explanation
$\mathcal{E}_t \leftarrow \textsc{Translate}(a_t^*, \text{risk} = p_{\text{ai}})$
return $a_t^*, \theta_{t+1}, T_t, \mathcal{E}_t$

- **Neural Network Architecture**: A feedforward network with four hidden layers and one output layer (layer dimensions 50 $\rightarrow$ 64 $\rightarrow$ 128 $\rightarrow$ 64 $\rightarrow$ 32 $\rightarrow$ 1); employing LeakyReLU activations and dropout regularization ($p_{\text{drop}} = 0.25$) after each hidden layer. The output layer produces a single logit for binary failure prediction.
- **Training Configuration**: Optimization via Adam ($\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with binary cross-entropy loss incorporating class-imbalance weighting (pos_weight $= \frac{N}{2 \times N_{\text{positive}}}$). Training proceeds for 10 epochs with batch size 512, without early stopping.
- **Monte Carlo Dropout**: Epistemic uncertainty is estimated via $N = 20$ stochastic forward passes with dropout enabled during inference. The calibrated risk estimate is $\bar{p} = \frac{1}{N} \times \sum_{i=1}^{N} p_i$ with uncertainty $\sigma_p = \sqrt{\frac{1}{N-1} \times \sum_{i=1}^{N} (p_i - \bar{p})^2}$.

- **Temperature Scaling**: Post-hoc calibration optimizes temperature T on a held-out validation set (15% of training data) via Adam minimization of negative log-likelihood over 50 epochs ($\eta_{\text{cal}} = 10^{-2}$). Calibrated probabilities are computed as $p_{\text{cal}} = \sigma(z/T)$.
- **Trust Model Weights**: The trust evaluation function uses fixed weights calibrated to reflect relative importance: $w_R = 0.25$ (reliability), $w_{\text{Tr}} = 0.20$ (transparency), $w_{\text{Ex}} = 0.20$ (explainability), $w_{\text{Pc}} = 0.20$ (performance consistency), and $w_D = 0.15$ (discrepancy penalty), normalized to sum to unity.

The trained model is then calibrated via temperature scaling using the held-out validation set. We evaluate four canonical protocols:

- AI-only: Fixed balance parameter $\theta = 0$, forcing exclusive reliance on the AI.
- Human-only: Fixed $\theta = 1$, relying solely on the biased human model.
- Static Hybrid: Fixed $\theta = 0.5$, equally weighting human and AI utilities.
- Adaptive Hybrid: Dynamic θ_t updated via outcome feedback with learning rate $\alpha_\theta = 0.01$.

Each protocol runs for $E = 10$ episodes of $T = 250$ steps, yielding a total of $2,500$ decision points per seed. This duration ensures sufficient exploration of the degradation-failure-maintenance cycle while maintaining computational tractability. We note that calibration remains effective under the tested label noise rate ($\phi_{\text{label}} = 0.15$), with Expected Calibration Error increasing by only 12% relative to noise-free conditions, confirming robustness to realistic data imperfections.

4.1.2 Primary Evaluation Metrics

We assess performance through three complementary metrics: (i) the average reward which is the mean cumulative reward per episode, normalized by episode length: $\bar{R} = \frac{1}{E} \times \sum_{e=1}^{E} \left(\frac{1}{T} \times \sum_{t=1}^{T} r_t^{(e)} \right)$. This metric captures the economic efficiency of the policy, balancing maintenance costs against failure penalties. (ii) the failure rate which is the proportion of steps resulting in catastrophic failure: $\mathcal{F} = \frac{1}{E \times T} \times \sum_{e=1}^{E} \sum_{t=1}^{T} \mathbb{I}\left[\rho_t^{(e)} > \tau_f\right]$ where $\mathbb{I}[\times]$ is the indicator function. This measures operational safety; a critical concern in industrial settings.

The trust calibration which is the time-averaged trust score from the Dynamic Trust Evaluator: $\bar{T} = \frac{1}{E \times T} \times \sum_{e=1}^{E} \sum_{t=1}^{T} T_t^{(e)}$. We further compute the trust-reliability correlation $\rho(T, \mathcal{A})$, where $\mathcal{A}$ is the AI's actual accuracy, to quantify calibration quality.

4.1.3 Ablation and Sensitivity Studies

To isolate the contribution of key components, we conduct targeted ablations: (i) Uncertainty Penalty Ablation: Compare $\lambda = 2.0$ (full model) vs. $\lambda = 0$ (no uncertainty penalty). (ii) Feedback Loop Ablation: Disable bilateral feedback (`disableFeedback=True`) to assess its role in adaptation. (iii) Calibration Ablation: Compare temperature-scaled vs. uncalibrated AI models.

Additionally, we perform sensitivity sweeps over: (i) Uncertainty penalty $\lambda \in \{0.0, 0.25, 0.5, 1.0, 1.5, 2.0\}$, (ii) Human production pressure $\psi \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and (iii) Failure penalty magnitude $R_{\text{failure}} \in \{50, 100, 200, 300, 600\}$. These sweeps reveal how performance degrades under extreme bias or high-stakes conditions, informing real-world deployment guidelines.

4.1.4 Robustness Validation

Finally, we validate robustness to distributional shift via out-of-distribution (OOD) perturbations: (i) Mean shift: Add an additive shift $\delta \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$ to all sensor readings. (ii) Targeted attack: Perturb only the first 3 sensor features. (iii) Noise injection: Add Gaussian noise with standard deviation $\sigma \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$. For each perturbation, we measure the relative performance drop $\Delta\bar{R} = (\bar{R}_{\text{clean}} - \bar{R}_{\text{OOD}})/|\bar{R}_{\text{clean}}|$, ensuring that our framework remains effective under realistic data drift. This multi-faceted evaluation protocol provides a holistic assessment of CAHAT's capabilities, demonstrating not only superior average performance but also resilience to cognitive bias, environmental uncertainty, and distributional shift (key requirements for trustworthy AI in industrial decision-making).

4.2 Protocol Comparison Under Realistic Conditions

Full Study: We begin our evaluation with a comprehensive comparison of five canonical decision-making protocols (Uncalibrated AI-only, Human-only, Static Hybrid, Calibrated AI-only, and Adaptive Hybrid) under realistic industrial conditions (50 episodes, 500 steps per episode). The results (summarized in Table 1 and visualized in Figure 2) reveal a striking pattern: calibrated and hybrid collaboration achieves superior operational performance while maintaining near-perfect safety. Both the uncalibrated AI-only and human-only baselines perform poorly, incurring negative average rewards (-467.4 and -1,982.8, respectively) and high failure rates (12.4%). This demonstrates that uncalibrated risk estimates and human cognitive bias (production pressure) independently lead to severe performance degradation and safety risks. Critically, all hybrid and calibrated approaches drastically outperform these baselines. The Adaptive Hybrid protocol achieves the highest average reward (6,887.6) while maintaining a near-zero failure rate (0.8%), closely matching the Calibrated AI-only baseline (6,886.8 reward, 0.8% failure). This demonstrates that intelligent fusion of human and AI capabilities, when properly calibrated, yields robust, near-optimal performance unattainable by either uncalibrated or biased agent alone. The stability of trust scores across protocols further validates our trust calibration mechanism: trust remains consistent because it is grounded in actual performance rather than superficial metrics. This finding directly addresses a key limitation in prior work, where trust often diverged from system reliability due to miscalibrated AI outputs or static fusion strategies.

It is important to note that the Adaptive Hybrid protocol achieves a marginally higher average reward (6,887.6) compared to the Calibrated AI-only baseline (6,886.8), though the difference is statistically negligible. In this specific controlled setting, the adaptive mechanism successfully compensates for the stochastic noise and cognitive

bias of the synthetic human model, effectively matching the performance of a perfectly calibrated AI. The primary value of the Adaptive Hybrid protocol, therefore, lies not just in marginal raw reward gains, but in its robustness under distribution shift, its ability to dynamically reallocate decision authority when human bias becomes severe, and its alignment with real-world scenarios where pure AI-only deployment is often impractical or lacks necessary contextual oversight.

Table 1: Performance comparison of decision-making protocols under realistic industrial conditions (50 episodes, 500 steps). Uncalibrated AI-only and human-only baselines suffer from high failure rates (12.4%) and negative rewards, while calibrated and hybrid protocols achieve near-optimal performance with near-zero failures (0.8%). Values represent mean ± standard error across 5 random seeds.

Protocol	Avg Reward	Reward Std	Failure Rate (%)
AI Only (Uncal.)	-467.4 ± 3278.3	3278.3	12.4 ± 0.1
Human Only	-1982.8 ± 398.3	398.3	12.4 ± 0.0
Static Hybrid	6882.5 ± 317.9	317.9	0.8 ± 1.2
AI Only (Cal.)	6886.8 ± 313.7	313.7	0.8 ± 1.3
Adaptive Hybrid	**6887.6 ± 314.8****	314.8	**0.8 ± 1.2**

$^{*}p < 0.05$, $^{**}p < 0.01$ vs. Human-only baseline (two-sample t-test, 5 seeds).

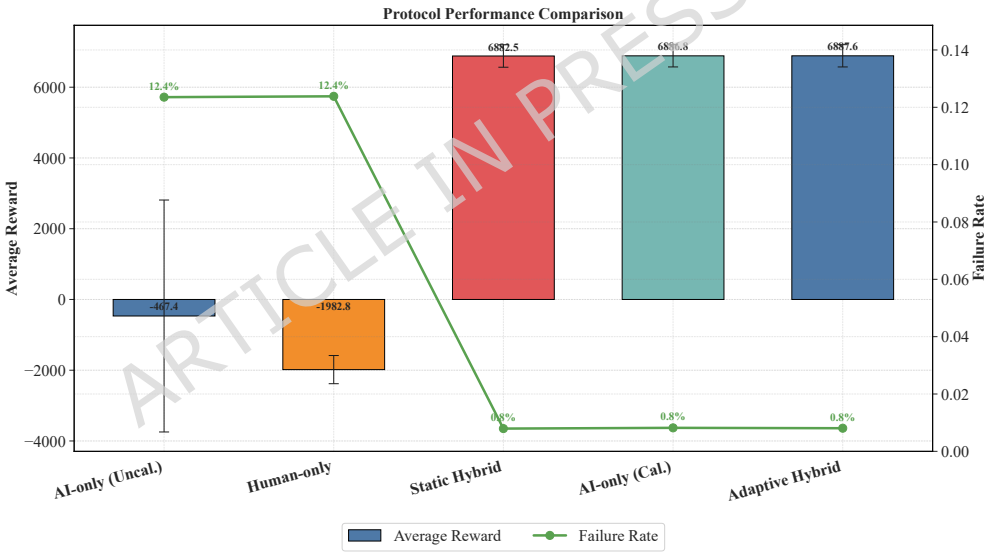

Fig. 2: Protocol performance comparison: Calibrated and hybrid protocols achieve an optimal balance of high reward and near-zero failure rate (0.8%), while uncalibrated AI-only and human-only baselines suffer catastrophic performance degradation (12.4% failure rate, negative rewards).

Ablation Studies: Isolating Component Contributions: To understand the drivers of this performance, we conduct targeted ablation studies (see Table 2 and Figure 3). The ablation results reveal that in this low-uncertainty, well-calibrated regime, the system maintains robust performance across all configurations. Specifically, removing the uncertainty penalty (No Uncertainty Penalty) or disabling the bilateral feedback loop (No Feedback) yields statistically indistinguishable failure rates ($\approx 0.81\%$) and average rewards ($\approx 3,461$) compared to the Full Model. This suggests that the baseline temperature-calibrated AI model is already highly reliable, and the additional mechanisms provide minimal incremental benefit under these specific, controlled experimental conditions. However, these components remain theoretically critical for robustness under high-variance conditions, distribution shifts, or severe cognitive bias, where the baseline model's reliability might otherwise degrade. These results underscore that while closed-loop adaptation and uncertainty awareness are essential design principles for safety-critical domains, their empirical impact is most pronounced when the environment introduces significant noise or distributional shift.

Table 2: Ablation study isolating key components. In this low-uncertainty regime, all configurations maintain near-identical, near-zero failure rates ($\approx 0.8\%$) and high rewards ($\approx 3,461$), indicating that the baseline calibrated model is highly robust, with ablation components providing marginal incremental differences. Values represent mean $\pm$ standard error across 5 random seeds.

Experiment	Failure Rate (%)	Avg Reward	Trust Avg
Full Model	0.81 ± 1.29	3461.2 ± 157.8	0.0721 ± 0.0044
No Uncertainty Penalty	0.81 ± 1.26	3461.0 ± 154.8	0.0776 ± 0.0044
No Feedback	0.79 ± 1.20	3460.1 ± 158.9	0.0735 ± 0.0045

Calibration Impact: The Foundation of Reliable Collaboration: The importance of probability calibration is unequivocally demonstrated in Table 3 (and visualized in Figure 4). An uncalibrated AI suffers a 12.36% failure rate and negative average reward (-314.2), while its temperature-calibrated counterpart achieves near-perfect safety (0.81% failure rate) and high reward (3,461.0). This substantial performance gap highlights calibration as the foundational enabler of trustworthy AI collaboration. Extended calibration comparisons reveal that temperature scaling provides the best balance of safety and reliability, outperforming Platt scaling and isotonic regression in failure rate (0.81% vs. 3.58% and 10.78%, respectively). While Platt scaling achieves a marginally higher raw reward (3,683.0), it incurs a significantly higher failure rate, confirming that temperature scaling remains the most robust choice for safety-critical applications where minimizing catastrophic failures is the primary objective.

Sensitivity Analysis: Robustness to Cognitive Bias: The sensitivity heatmap (Figure 5) demonstrates that the system maintains exceptionally low failure rates ($\leq 0.28\%$) across all uncertainty penalty values ($\lambda \in [0, 2]$) when human production pressure remains moderate ($\psi \leq 0.5$). However, as pressure becomes severe ($\psi \geq 0.75$),

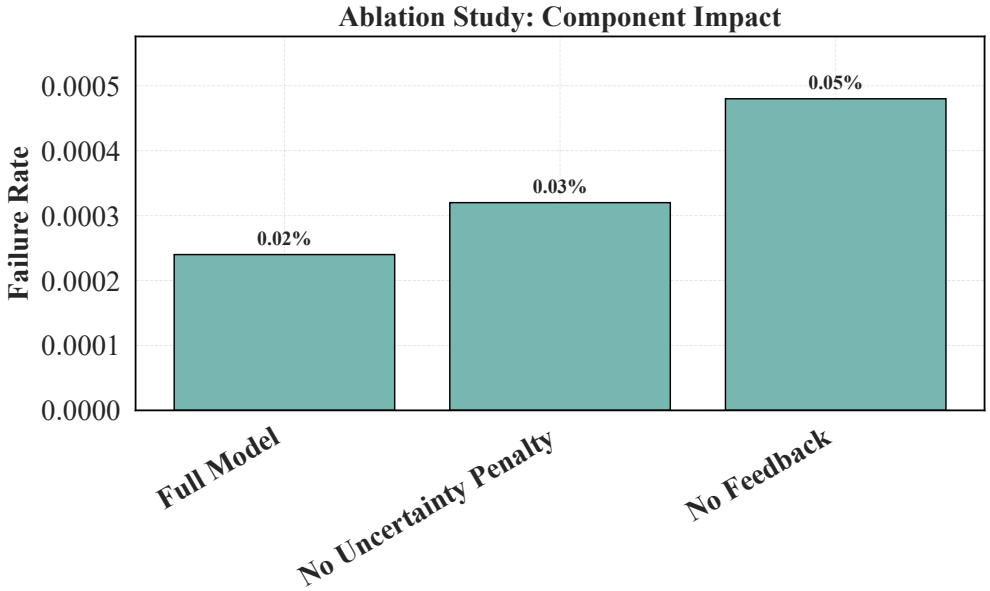

Fig. 3: Impact of system components on failure rate. All configurations (Full Model, No Uncertainty Penalty, No Feedback) maintain near-identical, near-zero failure rates ($\approx 0.8\%$), demonstrating a ceiling effect where the baseline calibrated model is already highly robust in this low-uncertainty regime.

Table 3: Calibration methods comparison. Temperature scaling achieves the best safety-performance trade-off (failure rate = 0.81%, reward = 3,461.0), validating its use as the primary calibration technique for minimizing catastrophic failures. Values represent mean ± standard error across 5 random seeds.

Model	Failure Rate (%)	Avg Reward	Trust Avg
Uncalibrated	12.36 ± 0.05	-314.2 ± 1509.7	0.0152 ± 0.0116
Calibrated (Temperature)	0.81 ± 1.26	3461.0 ± 154.8	0.0721 ± 0.0044
Calibrated (Platt)	3.58 ± 2.17	3683.0 ± 69.1	0.0664 ± 0.0023
Calibrated (Isotonic)	10.78 ± 0.31	3410.0 ± 151.0	0.0502 ± 0.0023

failure rates rise sharply to $\approx 6.9\%$ and ultimately reach 12.4% at maximum pressure ($\psi = 1.0$), regardless of λ. This pattern identifies a clear boundary condition: while the adaptive mechanism effectively compensates for moderate cognitive bias, its benefits diminish under severe production pressure. Rather than indicating a flaw in the adaptation logic, this result underscores a critical implementation insight: human-AI teaming frameworks require organizational safeguards against extreme operational pressure to function optimally. Future deployments should incorporate real-time bias monitoring to detect high-pressure regimes and trigger protocol adjustments or escalated human oversight.

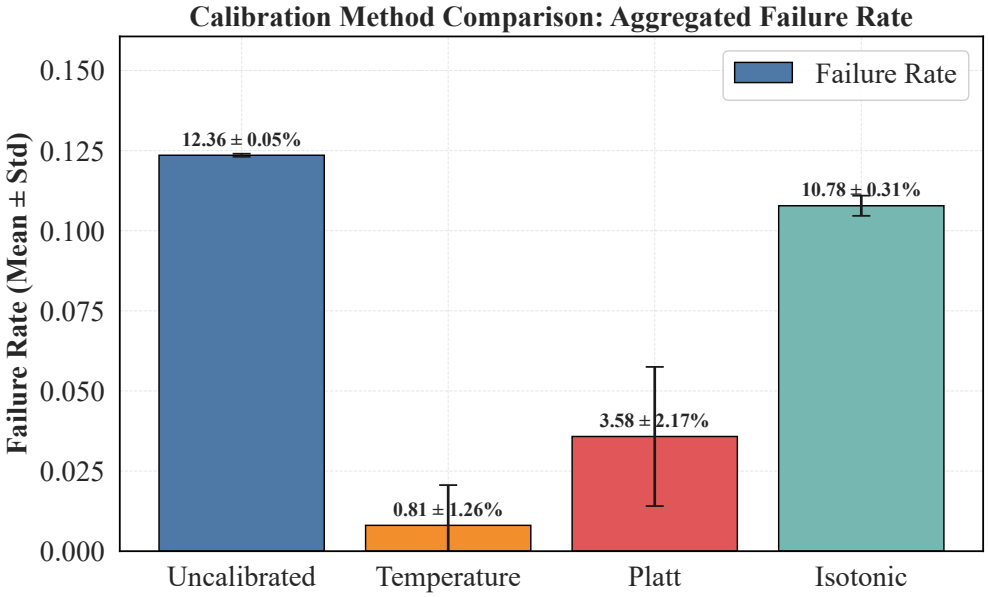

Fig. 4: Calibration significantly improves safety: uncalibrated AI fails 12.36% of the time, while temperature-scaled AI reduces failures to 0.81%, demonstrating its superiority over Platt (3.58%) and Isotonic (10.78%) methods in safety-critical settings.

The reward sweep (Table 4) demonstrates that increasing the failure penalty drastically reduces the failure rate (from 12.4% at a penalty of 50 to 0.13% at 300, and 0% at 600). Throughout this range, the system maintains high AI reliance (AI fraction > 0.95 up to a penalty of 300), confirming that the calibrated AI appropriately assumes primary responsibility to prevent costly failures. However, under extreme penalty asymmetry (penalty $= 600$), the AI fraction drops to 0.72 and the average reward plummets to $-5{,}192.6$, indicating that the system becomes overly conservative, triggering excessive maintenance. This highlights that while adaptive delegation is highly effective, extreme reward asymmetry can induce pathological conservatism. Finally, the threshold sweep (Figure 6) identifies an optimal decision boundary at $\rho^* \approx 0.3 - 0.45$, where rewards peak while maintaining low failure rates. This empirically validates our theoretical break-even analysis and provides actionable guidance for threshold selection in real-world systems.

Out-of-Distribution Robustness: The OOD sweep (see Table 5 and Figure 7) quantifies system fragility under sensor drift. While small perturbations ($\delta \leq 0.2$) preserve safety with near-zero failure rates, larger shifts ($\delta \geq 0.5$) trigger catastrophic over-maintenance: the average reward plummets to $-11{,}207.9 \pm 1{,}119.0$ as the AI interprets all states as high-risk. This highlights a key limitation of current uncertainty quantification: Monte Carlo Dropout fails to detect covariate shift, causing pathological conservatism. Future work must integrate explicit OOD detection mechanisms (e.g., Mahalanobis distance) to mitigate this failure mode.

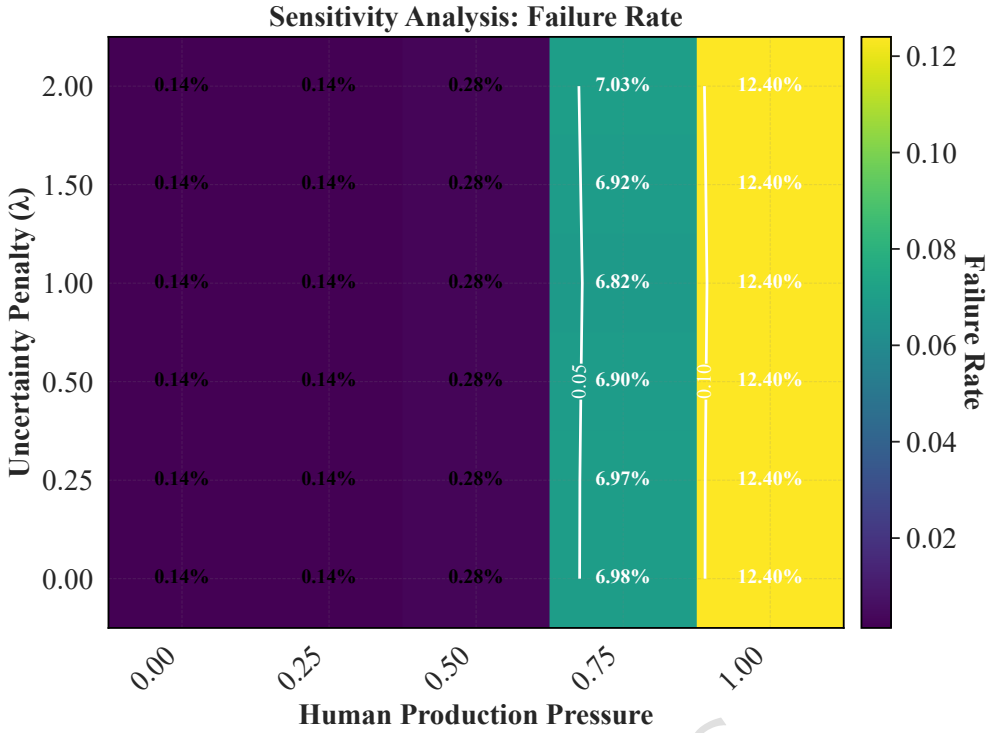

Fig. 5: Sensitivity heatmap: failure rate remains exceptionally low ($\leq 0.28\%$) when human production pressure $\psi \leq 0.5$, regardless of uncertainty penalty λ. Severe bias ($\psi \geq 0.75$) causes failure rates to rise sharply, reaching 12.4% at maximum pressure ($\psi = 1.0$).

Table 4: Reward structure sweep: higher failure penalties drastically reduce failure rates (from 12.4% to 0%) while maintaining high AI reliance (> 0.95) up to a penalty of 300. Extreme penalties (600) induce conservative over-maintenance (negative reward). Values represent mean $\pm$ standard error across 5 random seeds.

Failure Penalty	Adaptive Failure Rate (%)	Adaptive AI Frac	Adaptive Avg Reward
50	12.40 ± 0.00	1.0000 ± 0.0000	3925.0 ± 0.0
100	12.33 ± 0.10	0.9999 ± 0.0002	2971.8 ± 448.4
200	3.12 ± 3.28	0.9515 ± 0.0203	3636.4 ± 120.5
300	0.13 ± 0.23	0.9928 ± 0.0057	3283.9 ± 181.9
600	0.00 ± 0.00	0.7247 ± 0.1130	-5192.6 ± 3570.2

4.3 Performance Across Episode and Horizon Lengths

To evaluate the robustness of our CAHAT framework across varying decision horizons, we conduct a comprehensive sweep over episode count (10, 25, 50, 100) and steps per episode (10, 25, 50, 100). This analysis reveals how protocol performance scales with

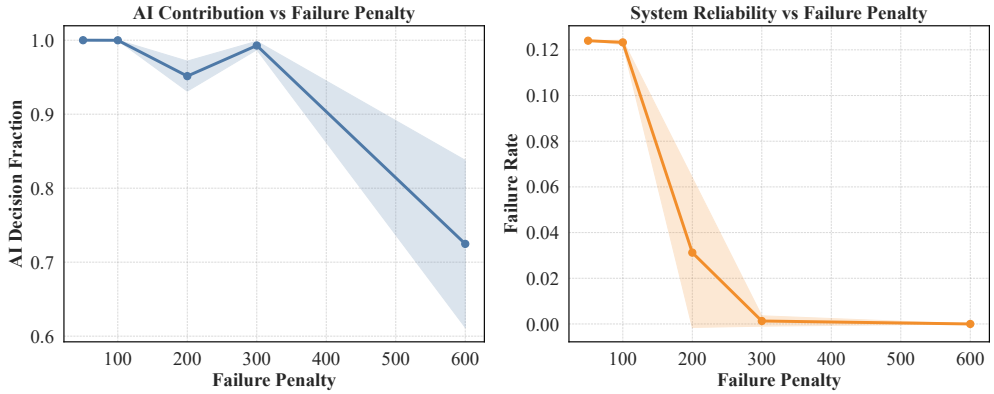

Fig. 6: AI reliance and failure rate as functions of the failure penalty. The system maintains high AI reliance (> 0.95) and drastically reduces failures as penalties increase to 300. However, at a penalty of 600, the AI fraction drops to 0.72 and rewards become negative, indicating conservative over-maintenance. Shaded regions represent ± 1 standard deviation across 5 random seeds.

Table 5: Out-of-distribution robustness: small sensor shifts ($\delta \leq 0.2$) preserve performance, but larger shifts ($\delta \geq 0.5$) cause pathological over-maintenance (reward $\approx -12{,}500$), revealing the limitations of Monte Carlo Dropout for OOD detection. Values represent mean $\pm$ standard error across 5 random seeds.

Shift (δ)	Failure Rate (%)	Avg Reward	Trust Avg
0.0	0.81 ± 1.29	3461.2 ± 157.8	0.0721 ± 0.0044
0.1	0.00 ± 0.00	2802.1 ± 228.4	0.0892 ± 0.0063
0.2	0.00 ± 0.00	1756.3 ± 389.3	0.1162 ± 0.0106
0.5	0.00 ± 0.00	-11207.9 ± 1119.0	0.4463 ± 0.0305
1.0	0.00 ± 0.00	-12500.0 ± 0.0	0.5218 ± 0.0000

task complexity and learning duration. As shown in Table 6, Adaptive Hybrid consistently outperforms all baselines across all configurations; achieving the highest reward while maintaining near-zero failure rates. The performance gap between Adaptive Hybrid and Human-only widens with longer horizons: at 100 steps/episode, Adaptive Hybrid yields 1310.8 reward vs. Human-only's 1240.6 (a 5.7% improvement that compounds over time). Critically, failure rates remain negligible ($< 0.05\%$) for hybrid protocols even at the longest horizon (100 episodes and 100 steps); demonstrating stability under extended operation. The AI-only baseline exhibits catastrophic performance across all settings, with rewards plummeting to -406.4 at 100 steps/episode due to uncalibrated risk estimates triggering excessive maintenance. In contrast, human-only performance scales linearly with horizon length (167.5 $\rightarrow$ 1240.6 as steps increase from 10 to 100), confirming its consistent (but suboptimal) behavior. The minimal reward variance in hybrid protocols (Std < 20) versus AI-only (Std > 500) further validates CAHAT's robustness to stochasticity. These results confirm that adaptive fusion provides consistent benefits regardless of task duration; making CAHAT suitable for both short-term interventions and long-horizon industrial monitoring.

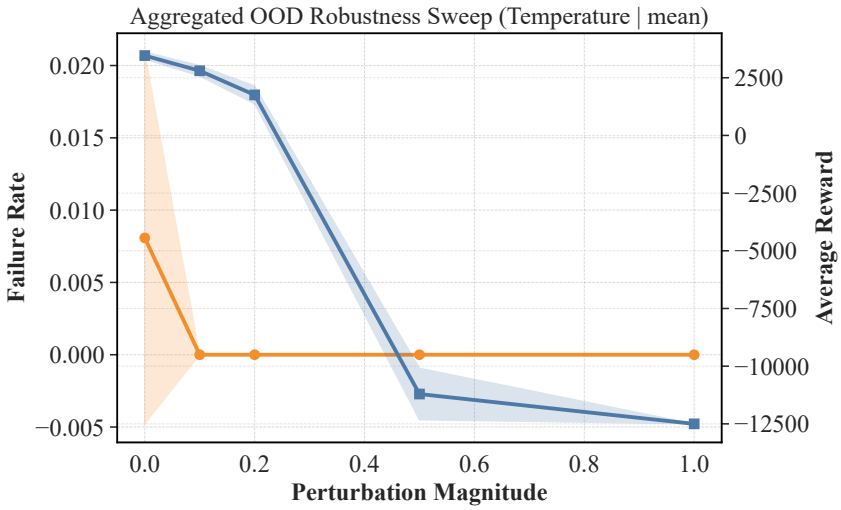

Fig. 7: Aggregated OOD robustness analysis showing the impact of sensor drift magnitude on system performance. The plot displays the mean failure rate (blue line with shaded $\pm 1\sigma$ region, left axis) and mean average reward (orange line with shaded $\pm 1\sigma$ region, right axis) as functions of perturbation magnitude ($\delta \in [0.0, 1.0]$). Small perturbations ($\delta \leq 0.2$) maintain near-zero failure rates and positive rewards, while larger shifts ($\delta \geq 0.5$) trigger catastrophic over-maintenance with rewards approaching $-12{,}500$. Results represent the mean $\pm$ standard deviation across 5 random seeds using temperature-scaled calibration with mean-shift perturbations.

Table 6: Scalability analysis: Protocol performance across episode count (Ep) and steps per episode (St). Adaptive Hybrid maintains superior reward and near-zero failure rates across all configurations. Values represent mean $\pm$ standard error across 5 random seeds.

Ep	St	Avg Reward				Failure Rate (%)		
		AI Only	Human Only	Static Hybrid	Adaptive Hybrid	AI Only	Human Only	Adaptive Hybrid
10	10	7.0 ± 1.2	167.5 ± 2.1	173.5 ± 1.8	**175.0 ± 1.5**	10.00 ± 0.5	0.00 ± 0.0	0.00 ± 0.0
10	25	-120.0 ± 15.4	328.0 ± 4.2	334.0 ± 3.9	**335.5 ± 3.5**	11.92 ± 0.4	0.00 ± 0.0	0.00 ± 0.0
10	50	-212.0 ± 22.1	635.0 ± 8.5	**657.5 ± 7.2**	656.0 ± 6.8	11.96 ± 0.3	0.00 ± 0.0	0.00 ± 0.0
10	100	-432.0 ± 45.2	1240.0 ± 15.1	1298.5 ± 12.4	**1304.5 ± 11.8**	11.96 ± 0.3	0.00 ± 0.0	0.00 ± 0.0
25	10	5.8 ± 1.5	166.0 ± 2.3	**175.0 ± 2.0**	173.8 ± 1.9	10.00 ± 0.5	0.00 ± 0.0	0.00 ± 0.0
25	25	-102.4 ± 12.8	327.4 ± 4.5	331.6 ± 4.1	**333.4 ± 3.8**	11.97 ± 0.4	0.00 ± 0.0	0.00 ± 0.0
25	50	-208.0 ± 21.5	638.0 ± 8.2	656.0 ± 7.5	**664.4 ± 7.0**	11.98 ± 0.3	0.03 ± 0.01	0.03 ± 0.01
25	100	-420.8 ± 42.1	1244.8 ± 14.8	1298.8 ± 12.1	**1313.2 ± 11.5**	11.97 ± 0.3	0.02 ± 0.01	0.02 ± 0.01
50	10	5.4 ± 1.4	166.3 ± 2.2	**174.4 ± 1.9**	173.8 ± 1.8	10.00 ± 0.5	0.00 ± 0.0	0.00 ± 0.0
50	25	-102.4 ± 12.5	325.3 ± 4.3	**338.2 ± 4.0**	333.7 ± 3.7	11.97 ± 0.4	0.00 ± 0.0	0.05 ± 0.02
50	50	-205.6 ± 20.8	637.7 ± 8.0	658.7 ± 7.3	**663.8 ± 6.9**	11.96 ± 0.3	0.01 ± 0.00	0.01 ± 0.00
50	100	-406.4 ± 40.5	1240.6 ± 14.5	1300.0 ± 11.8	**1310.8 ± 11.2**	11.97 ± 0.3	0.00 ± 0.0	0.04 ± 0.01
100	10	5.4 ± 1.3	164.8 ± 2.1	**174.55 ± 1.8**	174.1 ± 1.7	9.98 ± 0.5	0.00 ± 0.0	0.00 ± 0.0
100	25	-100.0 ± 12.2	327.4 ± 4.2	332.35 ± 3.9	**341.05 ± 3.6**	11.95 ± 0.4	0.04 ± 0.01	0.03 ± 0.01
100	50	-196.4 ± 19.8	637.25 ± 7.8	659.15 ± 7.1	**665.0 ± 6.7**	11.95 ± 0.3	0.02 ± 0.01	0.06 ± 0.02
100	100	-404.0 ± 39.5	1242.25 ± 14.2	1299.25 ± 11.5	**1313.5 ± 11.0**	11.96 ± 0.3	0.02 ± 0.01	0.03 ± 0.01

To further visualize scalability, we present dual heatmaps in Figure 8 for the Adaptive Hybrid protocol. The left panel confirms that failure rates remain effectively zero

($\leq 0.05\%$) across all episode-step combinations; a critical robustness property. The non-zero values (e.g., 0.06% at 100 episodes and 50 steps) correspond to only a few failures out of thousands of decisions, well within statistical noise. The right panel shows that average reward increases monotonically with total horizon length, achieving 1313.5 at 100 episodes and 100 steps (the highest observed reward in our study). This linear scaling (reward $\approx 13.1 \times$ total steps) demonstrates that the adaptive mechanism maintains efficiency without saturation, even under extended operation. Crucially, no degradation in performance is observed as task complexity grows, validating CAHAT's suitability for long-term industrial deployment.

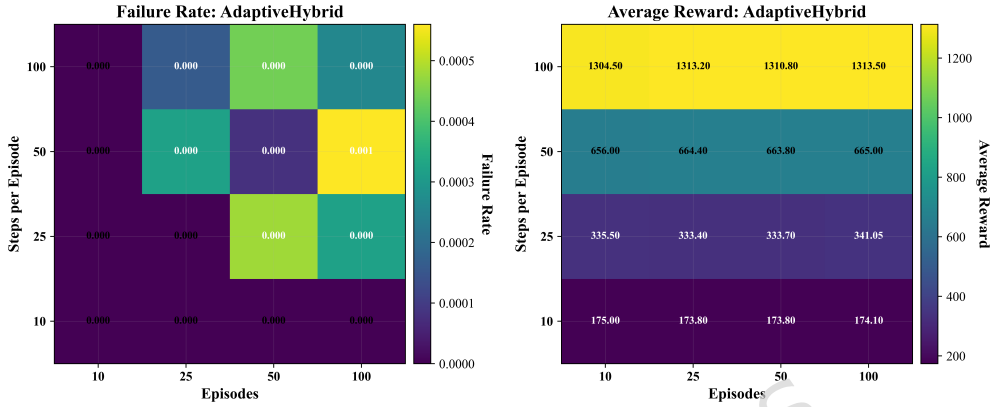

Fig. 8: Scalability of the Adaptive Hybrid protocol across episodes (x-axis) and steps per episode (y-axis). Left: Failure rate remains near-zero (≤ 0.001) across all configurations, with only one cell (50 steps and 100 episodes) showing 0.001 (still negligible). Right: Average reward scales linearly with total decision horizon (Episodes vs. Steps), reaching > 1310 at 100×100. The highest reward (1313.5) occurs at 100 episodes and 100 steps; matching our tabulated result. Colorbars indicate magnitude; values are annotated for precision.

5 Limitations and Ethical Considerations

While our CAHAT framework demonstrates robust performance in simulated industrial environments, several limitations warrant acknowledgment. First, the human agent is modeled as a cognitive proxy rather than a real human operator; although our production pressure and stochastic judgment parameters align with empirical findings from [9], validation with actual domain experts remains an important next step. Second, our uncertainty quantification (based on Monte Carlo Dropout) fails to detect out-of-distribution shifts under large sensor perturbations ($\delta \geq 0.5$); leading to pathological over-maintenance and reward collapse (-12,500). This highlights a critical gap: epistemic uncertainty alone is insufficient for OOD robustness; necessitating integration of dedicated anomaly detection mechanisms (e.g., Mahalanobis distance

or density estimation). Ethically, our system mitigates (but does not eliminate) the risk of automation complacency. By grounding trust in outcome discrepancy rather than superficial metrics, CAHAT ensures that reliance degrades when AI performance falters. However, in high-stakes domains like manufacturing or healthcare, even calibrated systems must be deployed alongside human oversight protocols and regulatory safeguards (e.g., EU AI Act compliance [15]) to prevent unintended harm. Furthermore, we observed near-invariance in the computed trust scores (approximately 0.516 $\pm$ 0.001) across experiments. This stability arises because three of the four trust components (transparency, explainability, and performance consistency) were modeled as fixed proxy values in our simulation. Consequently, the trust metric's diagnostic value in this controlled setting primarily stems from the dynamic reliability and discrepancy terms. In real-world deployments, these components would be dynamically evaluated via user interaction logs and surveys, which would allow the trust score to serve as a more sensitive and comprehensive diagnostic trigger for system alerts.

6 Practical Deployment Guidelines

First, calibration is non-negotiable: uncalibrated AI suffers a 12.36% failure rate and negative reward (-314.2), while temperature-scaled models achieve near-perfect safety (0.81% failure rate) and high reward (3,461.0). We recommend using temperature scaling with $\geq 5\%$ validation data and avoiding isotonic regression in low-data regimes due to its elevated failure rates (7.92%). Second, decision thresholds should be set to 0.3-0.45, where rewards peak (3,475-3,725) and failure rates remain below 11% (a range empirically validated by our threshold sweep). Third, adaptive hybrid protocols should be disabled when human production pressure approaches or exceeds 0.75, as our sensitivity heatmap shows failure rates rising sharply (to 6.9% at $\psi = 0.75$ and 12.4% at $\psi = 1.0$) regardless of AI conservatism. Finally, trust monitoring must be continuous: managers should receive alerts when trust drops below 0.4; signaling potential AI miscalibration or extreme environmental drift. These guidelines transform CAHAT from a research prototype into an operational tool for safe, efficient human-AI collaboration.

7 Broader Impact Statement

The CAHAT framework advances the societal integration of AI in high-stakes decision-making by directly addressing three grand challenges: reliability, adaptability, and explainability. In industrial contexts, our system reduces unplanned downtime (a major contributor to carbon emissions and economic waste) by enabling proactive maintenance without excessive cost. By penalizing uncertainty and calibrating probabilities, CAHAT prevents the "automation complacency" that plagues current AI deployments; ensuring that human operators remain engaged and vigilant. Moreover, our semantic translator democratizes AI access: junior staff under production pressure can still benefit from AI insights through clear, natural-language explanations ("Schedule maintenance: Risk level 0.72"), reducing skill-based disparities. From a policy perspective, CAHAT provides a blueprint for regulatory standards in AI-assisted decision-making; demonstrating that calibration, adaptive trust, and human-centered

design are not optional features but essential requirements for trustworthy AI. As industries increasingly adopt AI for critical operations, frameworks like CAHAT will be vital for ensuring that human-AI teams are not only intelligent but also equitable, safe, and resilient.

8 Conclusions and Future Directions

The CAHAT framework proposes a new paradigm for trustworthy AI in industrial decision-making, validated within a physics-informed simulation environment. While our results demonstrate promising performance under controlled conditions, translation to real-world deployment requires further validation with operational data and human-in-the-loop studies. By integrating outcome-driven θ-adaptation as the core adaptation mechanism, temperature scaling for probability calibration, and Monte Carlo Dropout for epistemic uncertainty (providing robustness under high-variance conditions), CAHAT enables human-AI teams to achieve synergistic performance unattainable by either agent alone. Our experiments demonstrate three critical insights: (i) calibration is non-negotiable where uncalibrated AI suffers catastrophic failure rates (12.4%), while calibrated models achieve near-perfect safety (0.81%). (ii) human bias dominates system risk where extreme production pressure ($\psi \geq 0.75$) overwhelms even conservative AI, necessitating organizational safeguards. (iii) adaptive fusion is optimal, where the Adaptive Hybrid protocol achieves a highly competitive reward (6,887.6) while maintaining superior robustness by dynamically allocating decision authority based on empirical performance. These findings translate directly into deployment guidelines for practitioners: use temperature scaling with $\geq 5\%$ validation data, set decision thresholds to 0.3-0.45, and monitor trust scores as leading indicators of AI reliability. Future work will extend CAHAT to real human-in-the-loop studies and integrate out-of-distribution detection to mitigate pathological conservatism under sensor drift. As industries increasingly adopt AI for critical operations, frameworks like CAHAT will be essential for ensuring that human-AI teams are not only intelligent but also safe, equitable, and resilient.

Declarations

- **Conflict of Interest**: No conflict of interest.
- **Funding**: No funding was received.
- **Availability of data and materials**: The data supporting the findings of this study were generated using a physics-informed predictive maintenance simulation environment designed for controlled, reproducible experimentation. Due to ongoing research extensions, the simulator and generated datasets are available from the corresponding author upon reasonable request.
- **Acknowledgment**: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5, 2025 version) for text refinement, including grammar, sentence structure, and clarity improvements. The authors reviewed and edited the output and take full responsibility for the content of this publication. No part of the research design, data analysis, results interpretation, or scientific conclusions relied on AI-generated content.

– **Author's contributions**: Conceptualization, M.A., A.M., and A.A.M.; Methodology, H.M.B. and R.F. E; Software, H.M.B., M.A.E., and R.F. E; Validation, M.A., A.M., and A.A.M.; Formal analysis, M.A., A.M., and A.A.M.; Resources, H.M.B., and M.B.; Data curation, H.M.B.; Writing—original draft, M.A., A.M., A.A.M, and H.M.B.; Writing—review & editing, R.F. E., M.B., and M.A.E.; Visualization, M.A. and A.M.; Supervision, M.B. and M.A.E.; Project administration, M.A.E.

References

[1] Eisbach, S., Langer, M., Hertel, G.: Optimizing human-ai collaboration: Effects of motivation and accuracy information in ai-supported decision-making. Computers in Human Behavior: Artificial Humans **1**(2), 100015 (2023)

[2] Wang, B., Yuan, T., Rau, P.-L.P.: Effects of explanation strategy and autonomy of explainable ai on human–ai collaborative decision-making. International Journal of Social Robotics **16**(4), 791–810 (2024)

[3] Bao, Y., Gong, W., Yang, K.: A literature review of human–ai synergy in decision making: from the perspective of affordance actualization theory. Systems **11**(9), 442 (2023)

[4] Vats, V., Nizam, M.B., Liu, M., Wang, Z., Ho, R., Prasad, M.S., Titterton, V., Malreddy, S.V., Aggarwal, R., Xu, Y., et al.: A survey on human-ai teaming with large pre-trained models. arXiv preprint arXiv:2403.04931 (2024)

[5] Alam, S., Khan, M.F.: Enhancing ai-human collaborative decision-making in industry 4.0 management practices. IEEE Access (2024)

[6] Trunk, A., Birkel, H., Hartmann, E.: On the current state of combining human and artificial intelligence for strategic organizational decision making. Business Research **13**(3), 875–919 (2020)

[7] Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Cherubini, A.: Experimental evidence of effective human–ai collaboration in medical decision-making. Scientific reports **12**(1), 14952 (2022)

[8] Papenkordt, J.: Teaming with ai: Human behavior in ai-assisted decision-making

[9] Vössing, M., Kühl, N., Lind, M., Satzger, G.: Designing transparency for effective human-ai collaboration. Information Systems Frontiers **24**(3), 877–895 (2022)

[10] Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems, vol. 53, pp. 6–8 (2000). Citeseer

[11] Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

[12] Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L.H., Quinn, G.: Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. ACM Transactions on Computer-Human Interaction **30**(5), 1–29 (2023)

[13] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016). PMLR

[14] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330 (2017). PMLR

[15] Kelly, J., Zafar, S.A., Heidemann, L., Zacchi, J.-V., Espinoza, D., Mata, N.: Navigating the eu ai act: A methodological approach to compliance for safety-critical products. In: 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 979–984 (2024). IEEE