

Human–AI Collaborative Decision-Making Under Uncertainty: A Conceptual Framework and Proposed Benchmark Specification

Lanz CWJ Chan

Finamatrix Quant-Lab, Singapore

Email: lanz@finamatrix.net

Shin-Hung Pan

Department of Information Management,
Chaoyang University of Technology, Taiwan

**Corresponding author* Email: vincentpan@cyut.edu.tw

Published on October 13, 2026

Abstract

Purpose: Human–AI collaboration under uncertainty remains under-theorized across policy, energy, finance, crisis response, and strategic planning. This conceptual article develops HACDU (Human–AI Collaborative Decision-Making Under Uncertainty), a framework that integrates Bayesian decision theory, bounded rationality, trust dynamics, explanation policies, and conflict resolution, and proposes HACDU-Bench, a benchmark specification for future cross-domain evaluation.

Design/methodology/approach: The study is conceptual and combines literature synthesis, formal modeling, and a design-science approach. Collaboration is modeled as a partially observable, utility-coupled stochastic process, from which a five-layer core decision pipeline, optional extensions, and five testable propositions are derived. No human-subject data were collected, and no confirmatory experiment was conducted.

Findings: As a conceptual study, the article reports theoretical rather than empirical findings. It offers (C1) a unified formalism centered on a core pipeline of belief fusion, trust dynamics, and override control; (C2) a modular algorithmic stack that makes its two strongest assumptions operational through a post-hoc calibration wrapper for AI uncertainty and a behavioral-telemetry estimator of human reliance; and (C3) HACDU-Bench as a proposed benchmark specification with deployment prerequisites. A characterization of positive collaborative value added (CVA) and five testable propositions complete the framework. A single energy-dispatch example shows how belief fusion, trust-dependent thresholds, value aggregation, and override logic interact in one decision cycle; it illustrates the mechanics of the framework and is not empirical validation.

Research limitations/implications: The propositions remain untested. Human-subject experiments, cross-domain replication, and field deployment are needed before any claim about performance can be made.

Practical implications: Once validated, HACDU could serve as a template for designing, auditing, and governing human–AI decision systems in policy analysis, energy operations, finance, and crisis management. HACDU-Bench offers a transparent structure for evaluating collaborative rather than purely predictive performance.

Originality/value: The article treats the collaborative human–AI system, rather than the standalone model, as the unit of analysis. It unifies decision theory, explainable AI, human factors, calibration, and benchmarking in one formal architecture and sets out a research agenda for hybrid intelligence.

Keywords: Human–AI collaboration; decision-making under uncertainty; Bayesian belief fusion; post-hoc calibration; explainable AI; collaborative value added.

JEL Classifications: C44, D81, D91, G11, G41, M51

1 Introduction

Human–AI collaboration is no longer a peripheral topic in decision sciences. In high-stakes environments — policy design, energy dispatch, crisis management, algorithmic trading, clinical triage, and geopolitical analysis — neither unaided human judgment nor fully autonomous AI is sufficient. The operative decision unit is a hybrid system in which human contextual expertise interacts with machine prediction, optimization, and uncertainty quantification.

This development has outpaced theory integration. The present article addresses this gap by proposing HACDU, grounded in prior work on risk-cybernetics and automated decision systems (Chan & Wong, 2012, 2013) and collaborative research on intelligent systems, IoT reliability, and AI-driven cyber-attack detection (Gaurav et al., 2025; Karim et al., 2023; Ly et al., 2024; Pan & Wang, 2020, 2021). All notation is defined in Table 1 (Appendix A).

The article is conceptual. It builds and formalizes a framework, derives testable propositions, and specifies a benchmark for future use, but it reports no empirical tests. Its numerical content is limited to one illustrative example (Section 11).

1.1 Research Questions

The article addresses four research questions:

RQ1: How can human–AI collaborative decision-making under uncertainty be formalized in a single framework (Sections 3 and 4)?

RQ2: Which modules make up an auditable collaborative decision architecture (Sections 4 to 9)?

RQ3: How can collaborative value added (CVA) be defined and measured (Sections 4.4 and 10)?

RQ4: Which testable propositions follow from the framework (Section 12)?

1.2 Contributions

- **C1 (theory): Unified formalism.** Human–AI collaborative decision-making is modeled as a partially observable, utility-coupled stochastic process, centered on a core pipeline of belief fusion, trust dynamics, and override control (Sections 3 to 7).
- **C2 (method): Modular algorithmic stack.** Each subsystem can be replaced and verified independently, and the stack makes its two strongest assumptions operational: a post-hoc calibration wrapper for AI uncertainty (Section 8) and a telemetry-based estimator of human reliance (Section 9).
- **C3 (evaluation): HACDU-Bench (proposed).** A transparent, multi-metric evaluation structure with mandatory weight disclosure, a disjoint reliance partition (ARR, OTR, UTR), and deployment prerequisites (Section 10). No benchmark results are reported.

These contributions are supported by a characterization of positive collaborative value added (Proposition 1, Section 4.4) and five testable propositions (Section 12). The collaborative POMDP, reinforcement-learning updates, robust optimization, fairness regularization, and variational inference are presented as theoretical completeness and future extensions (Appendix D), not as core contributions.

1.3 Paper Structure

Section 2 reviews prior work. Section 3 formalizes the problem. Section 4 sets out the conceptual research design and the core architecture, and Sections 4 to 7 develop the core mathematical framework. Section 8 specifies the calibration wrapper, and Section 9 the behavioral telemetry used to estimate reliance. Section 10 specifies the proposed HACDU-Bench. Section 11 gives a numerical illustration. Section 12 states testable propositions and a future empirical research agenda. Section 13 discusses implications and limitations, and Section 14 concludes. Appendix A provides a notation reference (Table 1), Appendix B an implementation checklist, Appendix C the architectural schematic (Figure 1), and Appendix D the optional extensions.

2 Theoretical Background

2.1 Decision Theory and Bounded Rationality

The normative foundation begins with expected utility theory (Savage, 1972; von Neumann & Morgenstern, 1944). The rational decision-maker solves

$$a^* = \arg \max_a E_\theta[u(a, \theta)]. \quad (1)$$

Equation 1 states the classical expected-utility rule: an ideal decision-maker chooses the action that maximizes expected utility over unknown states.

Real decision-makers face bounded rationality (Kahneman et al., 1982; Simon, 1957). A bounded human policy restricts attention to the cognitively available set A_j and penalizes effort $c(a)$:

$$a^H = \arg \max_{a \in A_j} [E_\theta[u(a, \theta)] - \lambda c(a)]. \quad (2)$$

Equation 2 introduces bounded rationality by restricting attention to a cognitively manageable subset, A_j , and subtracting a cognitive effort cost, $c(a)$.

2.2 Trust, Calibration, and Explanation

Reliance on automation depends on trust, calibration, explanation quality, and joint experience (Bansal et al., 2021; Guo et al., 2017; Lee & See, 2004; Parasuraman & Riley, 1997). HACDU embeds trust dynamics, calibration, and explanation choice inside the decision model rather than treating them as HCI add-ons.

2.3 Hybrid Intelligence and Complementarity

Team value depends on non-redundant errors: the team must exceed the best component (Bansal et al., 2021; Dellermann et al., 2019; Kamar et al., 2012). HACDU formalizes this through collaborative value added (CVA, Section 4.4).

2.4 Positioning

By defining the collaborative process as the scientific object, HACDU bridges operations research, decision theory, explainable AI, and human factors, consistent with established decision-science methodology (Chan & Wong, 2012, 2013; Ly et al., 2024).

3 Problem Formulation

Consider a finite-horizon sequential environment over $t = 1, \dots, T$ with latent state θ_t . All notation is defined in Table 1 (Appendix A). Human and AI maintain belief distributions:

$$b_t^H = p(\theta_t | o_{1:t}^H, z_t^H), \quad b_t^A = p(\theta_t | o_{1:t}^A, z_t^A), \quad (3)$$

$$I_t = (b_t^H, b_t^A, x_t, r_t, \delta_t, m_t), \quad (4)$$

$$a_t^C \sim \pi_C(\cdot | I_t). \quad (5)$$

The collaborative objective is to maximize the expected cumulative discounted net utility over a finite horizon T :

$$J_C = \mathbb{E} \left[\sum_{t=1}^T \gamma^{t-1} \left(u(a_t^C, \theta_t) - c_{proc}(I_t) - c_{err}(a_t^C, \theta_t) - c_{time}(\Delta t) \right) \right], \quad (6)$$

where $\gamma \in (0, 1]$ is the discount factor (with $\gamma^0=1$ at $t=1$). The term $u(a_t^C, \theta_t)$ represents the primary task payoff, c_{err} penalizes severe safety or domain violations, and c_{time} reflects execution latency. The process overhead $c_{proc}(I_t)$ isolates the non-outcome interaction frictions:

$$c_{proc}(I_t) = \alpha_e E_t + \alpha_c C_t + \alpha_d D_t, \quad (7)$$

where $E_t \geq 0$ is the explanation generation burden, $C_t \geq 0$ is the human cognitive load, and $D_t \geq 0$ is the dialogue or communication latency, parameterized by non-negative penalty weights $\alpha_e, \alpha_c, \alpha_d \geq 0$. Equations 3 to 7 formally decouple task outcome utility from human–AI interactive overheads, preventing double-counting of decision quality.

4 Conceptual Research Design

To establish theoretical lineage, Table 0 categorizes the equations in the HACDU stack into: (i) foundational decision-theoretic formulations adopted from prior literature, and (ii) novel structural extensions developed in this work.

Table 0. Lineage and Theoretical Foundations of HACDU Equations.

Mathematical Subsystem	Equation	Status	Theoretical Origin / Literature Extension
Classical Utility & Rationality	Eq. 1–2	Foundational	Expected Utility (Savage, 1972; von Neumann & Morgenstern, 1944); Bounded Rationality (Kahneman et al., 1982; Simon, 1957).
Collaborative Objective	Eq. 6–7	Novel Extension	Extends sequential utility by decoupling task payoffs from explicit interactive and cognitive friction costs (c_{proc}).
Belief Divergences	Eq. 9–12	Foundational	Symmetrized KL (Jeffreys, 1946); Wasserstein Metric (Villani, 2009); Spearman's rank correlation.
Trust Dynamics & Beta Filter	Eq. 13–16	Novel Extension	Extends Lee and See's (2004) trust theory into a dual-track state-space innovation filter and a non-stationary conjugate Beta Bayesian filter.
Collaborative Action-Value	Eq. 17–19	Novel Extension	Extends multi-attribute utility theory into context-adaptive convex utility fusion with execution friction c_{exec} .
Collaborative Value Added	Eq. 20–21	Novel Extension	Formalizes hybrid intelligence synergy metrics (CVA, NCG) with absolute stability guarantees.
Variational ELBO	c- Eq. D.1–D.2'	Novel Extension	Extends Auto-Encoding Variational Bayes (Kingma & Welling, 2014) to Log-Linear Product-of-Experts belief posteriors.
Risk-Sensitive LCB & Depth	Eq. 26–29'	Novel Extension	Extends Lower Confidence Bound RL (Srinivas et al., 2010) and Cognitive Load Theory (Sweller, 2011) to derive analytic closed-form explanation depth.
Dynamic Override Rule	Eq. 32–34	Novel Extension	Extends Parasuraman and Riley (1997) supervisory control into uncertainty-aware, stakes-sensitive monotonic override boundaries.
Calibration Wrapper	Eq. 38–46	Foundational + Extension	Adopts temperature and Platt scaling, isotonic regression, and conformal prediction (Guo et al., 2017; Platt, 1999; Zadrozny & Elkan, 2002; Angelopoulos & Bates, 2023), and adds ECE-triggered recalibration and a calibration-aware override threshold.
Telemetry Trust Estimator	Eq. 47–50	Novel Extension	Links behavioral proxies of reliance (Hoff & Bashir, 2015) to the Beta trust filter, calibrated against self-reported trust (Jian et al., 2000).
Proposed Benchmark Specification (HBS)	Eq. 51–58	Novel Extension	Original multi-metric composite evaluation benchmark with disjoint calibration partitioning (ARR, OTR, UTR).

This article follows a conceptual research design and performs no empirical tests. The design has five steps.

Literature synthesis. Constructs were drawn from expected-utility decision theory and Bayesian updating (Savage, 1972; von Neumann & Morgenstern, 1944), bounded rationality (Kahneman et al., 1982; Simon, 1957), trust and automation research (Lee & See, 2004; Parasuraman & Riley, 1997), interpretable and calibrated machine learning (Doshi-Velez & Kim, 2017; Guo et al., 2017), hybrid-intelligence work on human–AI teams (Bansal et al., 2021; Dellermann et al., 2019), and partially observable sequential decision-making (Russell & Norvig, 2021). Table 0 records which equations are adopted from this literature and which are new.

Formal modeling. The collaborative process is represented as a partially observable, utility-coupled stochastic process (Section 3), so that trust, explanation, override, and escalation become explicit model quantities rather than implicit design choices.

Framework construction. Following design-science logic, the formal model is organized into a five-layer core decision pipeline, a calibration wrapper, and a telemetry-based trust estimator, with optional extensions stated separately (Table 2; Figure 1 in Appendix C). Each component can be replaced by a domain-specific instantiation.

Proposition derivation. Formal results (Propositions 1 to 3) are proved from stated assumptions, and testable propositions (P1 to P5, Section 12) are derived from the structure of the framework.

Conceptual evaluation criteria. Following the design-science framework for evaluating conceptual artifacts (Hevner et al., 2004), the framework is assessed against five criteria: logical consistency, internal coherence across equations, explicit boundary conditions, testability, and operationalizability. Statistical significance testing is not used, because no data are analyzed.

Core pipeline and optional extensions. To keep the contribution focused, the framework is organized as a core decision pipeline, developed in full in the body of the article, plus optional extensions stated briefly in Appendix D (Table 2). The core runs from observation and context inputs, through belief fusion and the calibration wrapper, to the collaborative decision engine, the override, escalation, and arbitration layer, and HACDU-Bench evaluation. Variational inference, the collaborative POMDP, reinforcement-learning updates, robust optimization, and fairness regularization are extensions: they show theoretical completeness and future directions but are not needed to operate the core.

Table 2. Core Pipeline and Optional Extensions of HACDU.

Layer	Function	Main Equations	Status
L1 Observation and context inputs	Human and AI observations, priors, stakes, behavioral telemetry stream	Eq. 3–5; 47	Core
L2 Belief fusion and disagreement	Log-linear pooling; KL, Wasserstein, and rank divergence	Eq. 8–12, 22–25	Core
Calibration wrapper (between L2 and L3)	Recalibrates AI confidence and σ^A ; monitors drift	Eq. 38–46	Core
L3 Collaborative decision engine	Utility fusion, risk-sensitive scoring, explanation depth, trust dynamics	Eq. 13–19, 26–31	Core

Telemetry estimator	trust	Infers reliance from behavioral proxies for L3 and L4	Eq. 47–50	Core
L4 Override, escalation, and arbitration		Dynamic override threshold, escalation probability, arbitration policy	Eq. 32–37	Core
L5 HACDU-Bench evaluation		Benchmark metrics, composite score, collaborative value added	Eq. 20–21, 51–58	Core
Optional extensions		Variational inference, collaborative POMDP, reinforcement learning, robust optimization, fairness regularization	Eq. D.1–D.2', D.3–D.10, D.11–D.13	Extension (Appendix D)

Modular design principles. Design-science research treats design as a search process over candidate artifacts (Hevner et al., 2004), and classical software design decomposes systems into modules that hide their internal decisions (Parnas, 1972). In that spirit, each HACDU module follows five principles: single responsibility (one decision function per module); well-defined interfaces (modules exchange data only through the information state I_t of Equation 4); replaceability (each module can be swapped for a domain-specific implementation, such as a parametric, particle, or variational belief representation); independent verifiability (each module can be unit-tested and simulated on its own); and loose coupling (modules depend on I_t , not on one another's internals).

Unified Mathematical Framework. This Section develops the mathematical architecture across five subsystems. All symbols are defined in Table 1 (Appendix A).

4.1 Collaborative Belief State and Disagreement

$$b_t^C = \varphi(b_t^H, b_t^A, r_t, e_t), \quad (8)$$

$$D_{KL}^s(b_t^H \parallel b_t^A) = \frac{1}{2} [D_{KL}(b_t^H \parallel b_t^A) + D_{KL}(b_t^A \parallel b_t^H)], \quad (9)$$

$$D_{KL}(p \parallel q) = \sum_{\theta \in \Theta} p(\theta) \ln \frac{p(\theta)}{q(\theta)}, \quad (10)$$

$$W_1(b_t^H, b_t^A) = \inf_{\gamma \in \Gamma} E_{\gamma}[\|\theta_1 - \theta_2\|], \quad (11)$$

$$D_{rank} = 1 - \rho_s(b_t^H, b_t^A). \quad (12)$$

Equations 8 to 12 jointly define the collaborative belief b_t^C and three disagreement measures: D_{KL}^s captures symmetric information divergence, W_1 measures distance in outcome space, and D_{rank} captures rank-order disagreement between human and AI beliefs.

4.2 Reliance and Trust Dynamics

Human reliance propensity $r_t \in [0,1]$ represents the dynamic probability that the human decision-maker defers to AI recommendations. Rather than assuming static trust, HACDU formalizes trust evolution through two complementary, mathematically unified lenses: (i) an adaptive state-space innovation filter for fast algorithmic tracking, and (ii) a non-stationary Bayesian belief model over trust uncertainty.

At each decision epoch t , the human receives an instantaneous trust feedback signal (innovation) $\psi_t \in [0,1]$ driven by observed system performance, cognitive burden, and situational risk:

$$\psi_t = \sigma(\theta_0 + \theta_Q Q_t - \theta_\Lambda |\Lambda_t| - \theta_S S_t - \theta_f f_t - \theta_C C_t), \quad (13)$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ maps performance signals onto the unit interval. Here, trust is reinforced by decision quality ($\theta_Q > 0$), but penalized by calibration divergence ($|\Lambda_t|$, $\theta_\Lambda > 0$), severe failures ($f_t \in \{0,1\}$, $\theta_f > 0$), and excessive cognitive workload (C_t , $\theta_C > 0$). Critically, high situational stakes (S_t , $\theta_S > 0$) induce epistemic conservatism, suppressing uncritical reliance.

The dynamic reliance state r_{t+1} updates via an exponential smoothing filter with learning rate $\lambda_r \in (0,1]$:

$$r_{t+1} = (1 - \lambda_r) r_t + \lambda_r \psi_t. \quad (14)$$

To account for second-order uncertainty in human trust, the latent reliance is modeled as a Beta-distributed belief state $r_t \sim \text{Beta}(a_t, b_t)$, where a_t and b_t denote pseudo-counts of successful and unsuccessful collaborative interactions. To capture non-stationarity and recency bias, the parameters update sequentially via discounted conjugate updating:

$$a_{t+1} = \gamma_r a_t + \psi_t, \quad b_{t+1} = \gamma_r b_t + (1 - \psi_t), \quad (15)$$

where $\gamma_r \in (0,1]$ is an exponential memory retention factor. The expected reliance $E[r_t]$ and associated epistemic variance $\text{Var}[r_t]$ are analytically tractable:

$$E[r_t] = \frac{a_t}{a_t + b_t}, \quad \text{Var}[r_t] = \frac{a_t \cdot b_t}{(a_t + b_t)^2 (a_t + b_t + 1)}. \quad (16)$$

In the operational stack, the point estimate $E[r_t]$ drives utility weighting (Eq. 17) and override boundaries (Eq. 33), while the variance $\text{Var}[r_t]$ provides an uncertainty penalty to prevent premature convergence to over-trust or under-trust.

4.3 Utility Fusion and Collaborative Action-Value

The collaborative action-value function fuses individual assessments while deducting action-specific operational costs:

$$Q^C(a \mid I_t) = w^H(x_t) Q^H(a) + w^A(x_t) Q^A(a) - c_{\text{exec}}(a, I_t), \quad (17)$$

$$a_t^C = \arg \max_{a \in A} Q^C(a \mid I_t), \quad (18)$$

$$w^A(x_t) = \sigma(\beta_0 + \beta_1 E^H + \beta_2 S_{\text{shift}} + \beta_3 \Delta_{\text{cal}}), \quad (19)$$

where $c_{\text{exec}}(a, I_t)$ represents the direct execution and communication cost of dispatching action a given information state I_t .

4.4 Collaborative Value Added

To quantify the relative synergy generated by the human–AI partnership over standalone or static configurations, we define Collaborative Value Added (CVA) and Normalized Collaborative Gain (NCG):

$$CVA = J_C - \max(J^H, J^A, J_{static}), \quad (20)$$

$$NCG = \frac{CVA}{|\max(J^H, J^A, J_{static})| + \epsilon_0}, \quad (21)$$

where $\epsilon_0 > 0$ is a small stability constant (e.g., 10^{-5}). Taking the absolute value in the denominator keeps the ratio numerically stable and preserves its sign in domains with strictly negative utilities (e.g., loss minimization or crisis mitigation).

Proposition 1 (Necessary Condition and Characterization of Positive Collaborative Value Added).

Consider a one-step version of Equation 6. Let $\mathcal{E}^H, \mathcal{E}^A \subseteq \theta$ be the error sets of the human and the AI, let $\theta_1 = \mathcal{E}^A \setminus \mathcal{E}^H$ and $\theta_2 = \mathcal{E}^H \setminus \mathcal{E}^A$ be the regions where only one agent errs, and let $\bar{c}_{proc} > 0$ be the expected process overhead per decision. Assume that (a) the collaborative policy executes either a^H or a^A in each state; (b) on states where both or neither agent errs, the two recommendations yield equal utility; and (c) without loss of generality, the AI is the best single baseline, $J^A = \max(J^H, J^A, J_{static})$. Define $\Delta_1 = \mathbb{E}[u(a^H, \theta) - u(a^A, \theta) \mid \theta \in \theta_1] > 0$ and $\Delta_2 = \mathbb{E}[u(a^A, \theta) - u(a^H, \theta) \mid \theta \in \theta_2] > 0$, and let q_1 and e_2 be the probabilities that the arbitration defers to the human on θ_1 and on θ_2 , respectively. Then (i) $CVA > 0$ if and only if $q_1 P(\theta_1) \Delta_1 > e_2 P(\theta_2) \Delta_2 + \bar{c}_{proc}$; and (ii) error non-redundancy, $P(\theta_1) > 0$, is necessary but not sufficient for $CVA > 0$.

Proof. Under (a) and (b), the collaborative policy differs from the AI-only policy only when it defers to the human. It gains Δ_1 when it does so on θ_1 , loses Δ_2 when it does so on θ_2 , and pays the overhead in every state, so $J_C - J^A = q_1 P(\theta_1) \Delta_1 - e_2 P(\theta_2) \Delta_2 - \bar{c}_{proc}$. By (c), $CVA = J_C - J^A$, which gives (i). If $P(\theta_1) = 0$, that is, if the AI's errors are nested in the human's, the first term vanishes and $CVA \leq -\bar{c}_{proc} < 0$, so non-redundancy is necessary. It is not sufficient: with poor routing (small q_1 or large e_2) or high overhead, the inequality in (i) fails even when $P(\theta_1) > 0$. The case in which the human or the static hybrid is the best baseline follows by relabeling.

Equations 17 to 21 define how human and AI value functions are fused into a collaborative value Q^C and how collaborative value added (CVA) and normalized collaborative gain (NCG) quantify whether the team outperforms its best individual member.

5 Belief Fusion and Collaborative Inference

5.1 Bayesian and Log-Linear Fusion

$$b_t^C(\theta) \propto b_t^H(\theta) \cdot b_t^A(\theta), \quad (22)$$

$$b_t^C(\theta) = \frac{[b^H(\theta)]^{w^H} \cdot [b^A(\theta)]^{w^A}}{Z}, \quad (23)$$

$$\ln b_t^C(\theta) = w^H \ln b^H(\theta) + w^A \ln b^A(\theta) - \ln Z, \quad (24)$$

$$Z = \sum_{\theta \in \Theta} [b^H(\theta)]^{w^H} \cdot [b^A(\theta)]^{w^A}. \quad (25)$$

Equations 22 to 25 show that collaborative belief fusion can be viewed as a log-linear pooling of human and AI beliefs, where the weights w^H and w^A act as exponents controlling how much influence each agent has over the combined posterior.

6 Recommendation and Explanation Policies

6.1 Risk-Sensitive Collaborative Score

Under epistemic uncertainty, relying strictly on point expectations can lead to catastrophic edge-case failures. HACDU incorporates Lower Confidence Bound (LCB) principles to penalize high-variance recommendations:

$$LCB^A(a) = \hat{u}^A(a) - \lambda_{risk} \sigma^A(a), \quad LCB^H(a) = \hat{u}^H(a) - \lambda_{risk} \sigma^H(a), \quad (26)$$

$$S^C(a) = w^H LCB^H(a) + w^A LCB^A(a) - c_{exec}(a), \quad (27)$$

where $\hat{u}^A(a)$ and $\hat{u}^H(a)$ are expected utility estimates, $\sigma^A(a)$ and $\sigma^H(a)$ capture predictive standard deviations, $\lambda_{risk} \geq 0$ is a global risk-aversion parameter, and $c_{exec}(a)$ denotes action execution overhead. Equations 26 and 27 prioritize actions that achieve robust performance across ambiguous states while penalizing speculative, high-variance choices.

6.2 Explanation Depth Policy

Providing machine explanations under uncertainty involves a fundamental trade-off: detailed rationales improve mental model calibration but incur substantial computational overhead and human cognitive fatigue. To prevent degenerate corner solutions (where the system either provides no explanation or dumps maximal detail), HACDU balances marginal epistemic gain against strictly convex cognitive load.

The optimal explanation depth $d^* \in [0, d_{max}]$ solves the following concave utility maximization problem:

$$d^* = \arg \max_{d \in [0, d_{max}]} [G_{expl}(d) - c_{gen}(d) - c_{cog}(d)], \quad (28)$$

The expected decision-support gain $G_{expl}(d)$ is driven by the baseline necessity γ_1 , human contextual expertise demand E^H , and machine predictive uncertainty σ^A :

$$G_{expl}(d) = (\gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A) d, \quad (29)$$

where $\gamma_1, \gamma_2, \gamma_3 \geq 0$.

Proposition 2 (Closed-Form Optimal Explanation Policy).

Let the epistemic gain $G_{\text{expl}}(d)$ be an affine function of depth $d \in [0, d_{\text{max}}]$ modulated by machine uncertainty σ^A and human expertise demand E^H , and let cognitive processing cost be strictly convex: $c_{\text{total}}(d) = c_1 d + \frac{1}{2} c_2 d^2$ with $c_2 > 0$. Then, the optimal explanation depth d^* is unique, globally optimal, and admits the closed-form solution:

$$d^* = \text{clip} \left(\frac{\gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A - c_1}{c_2}, 0, d_{\text{max}} \right).$$

Proof.

The objective function $\Phi(d) = (\gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A - c_1)d - \frac{1}{2} c_2 d^2$ is twice continuously differentiable with second derivative $\frac{\partial^2 \Phi}{\partial d^2} = -c_2 < 0$. Since $c_2 > 0$, $\Phi(d)$ is strictly concave on the convex compact set $[0, d_{\text{max}}]$. By Weierstrass' Extreme Value Theorem, a unique global maximum exists. The unconstrained first-order necessary and sufficient condition $\frac{\partial \Phi}{\partial d} = 0$ yields $d^{\text{raw}} = \frac{\gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A - c_1}{c_2}$. Projecting d^{raw} onto $[0, d_{\text{max}}]$ via the Euclidean projection $\text{clip}(\cdot)$ completes the proof.

In accordance with Cognitive Load Theory (Miller, 2019; Sweller, 2011), while generation latency scales linearly ($c_{\text{gen}}(d) = c_1 d$), cognitive processing burden exhibits superlinear escalation due to working-memory limitations ($c_{\text{cog}}(d) = \frac{1}{2} c_2 d^2$, with $c_2 > 0$). Substituting these into Equation 28 yields the strictly concave net objective:

$$\Phi(d) = (\Gamma(E^H, \sigma^A) - c_1)d - \frac{1}{2} c_2 d^2$$

where $\Gamma(E^H, \sigma^A) \triangleq \gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A$. Setting the first-order condition $\partial \Phi / \partial d = 0$ and projecting onto the admissible depth set $[0, d_{\text{max}}]$ yields the unique, closed-form optimal explanation policy:

$$d^* = \text{clip} \left(\frac{\gamma_1 + \gamma_2 E^H + \gamma_3 \sigma^A - c_1}{c_2}, 0, d_{\text{max}} \right). \quad (29')$$

Equation 29' shows that optimal explanation depth expands continuously with machine uncertainty (σ^A) and human expertise requirements (E^H), while automatically contracting when cognitive fatigue or processing costs dominate.

6.3 Explanation Efficiency

$$\eta_e = \frac{\Delta \text{Decision Quality}}{c_{\text{gen}} + c_{\text{cog}}}, \quad (30)$$

$$H_e = \frac{1}{T} \sum_{t=1}^T \eta_{e,t}. \quad (31)$$

Equations 28 to 31 formalize the explanation policy: the system selects an explanation depth d that maximizes expected benefit minus generation and cognitive costs, and explanation efficiency η_e measures the performance gain per unit explanation effort.

7 Override, Escalation, and Arbitration

7.1 Uncertainty-Aware Override Rule

Rather than enforcing a static division of labor, HACDU regulates human intervention through a dynamic, uncertainty-aware override mechanism. An override is triggered when the epistemic divergence between the human and AI belief distributions exceeds a context-sensitive threshold τ :

$$a_t^* = \begin{cases} a_t^H, & \text{if } D_{KL}^S(b_t^H \parallel b_t^A) > \tau_t, \\ a_t^C, & \text{otherwise} \end{cases}, \quad (32)$$

The override threshold τ is dynamically parameterized as a function of the operational stakes S_t , current reliance r_t , and contextual divergence:

$$\tau = g(S_t, D_{KL}, r_t), \quad (33)$$

To ensure behavioral and decision-theoretic validity, the threshold function $g(\cdot)$ must satisfy the following monotonicity constraints:

1. **Trust-Driven Discretion:** $\partial g / \partial r_t > 0$. As human trust and calibrated reliance increase, the threshold τ rises, widening the AI's autonomous operational envelope and suppressing unnecessary human micro-interventions.
2. **Stakes-Sensitive Precaution:** $\partial g / \partial S_t \leq 0$. Under higher stakes or mission-critical consequences, the threshold τ contracts, lowering the barrier for human supervisory oversight and encouraging active human verification even under moderate disagreement.

An illustrative linear-clamped instantiation is:

$$\tau_t = \text{clip}(\tau_0 - \theta_S S_t + \theta_r r_t, \tau_{\min}, \tau_{\max})$$

where $\tau_0 \in (0,1)$ is the baseline tolerance, and $\theta_S, \theta_r > 0$ are domain-specific sensitivity parameters.

Proposition 3 (Safety-Preserving Override Monotonicity).

Let the dynamic override boundary be governed by $\tau_t = g(S_t, D_{KL}, r_t)$. Under the structural constraints $\frac{\partial g}{\partial r_t} > 0$ and $\frac{\partial g}{\partial S_t} \leq 0$, the probability of triggering human supervisory intervention $P(\text{Override})$ is monotonically increasing with operational stakes S_t and monotonically decreasing with calibrated reliance r_t .

Proof.

An override occurs when the epistemic divergence exceeds the threshold: $O = \{\omega: D_{KL}^S > g(S_t, D_{KL}, r_t)\}$. For any two stake levels $S_1 < S_2$, the condition $\frac{\partial g}{\partial S_t} \leq 0$ implies $g(S_2, \cdot) \leq g(S_1, \cdot)$. Therefore, the intervention event set satisfies $O_{S_1} \subseteq O_{S_2}$, which implies $P(O_{S_2}) \geq P(O_{S_1})$. By identical reasoning, since $\frac{\partial g}{\partial r_t} > 0$, higher reliance $r_1 < r_2$ yields $g(\cdot, r_2) > g(\cdot, r_1)$, resulting in $O_{r_2} \subseteq O_{r_1}$ and $P(O_{r_2}) \leq P(O_{r_1})$.

When severe conflict persists or when collaborative value is compromised, the system invokes an arbitration escalation policy. The escalation probability $p_{esc}(I_t)$ is modeled as:

$$p_{esc}(I_t) = \sigma(v_0 + v_1 D_{KL} + v_2 S_t - v_3 Q_t). \quad (34)$$

Equations 32 to 34 jointly formalize dynamic authority handovers: overrides are mediated by the balance between stakes and calibrated trust, while escalation to higher-level arbitration grows with belief disagreement ($v_1 > 0$) and consequence severity ($v_2 > 0$), but is mitigated by expected collaborative decision quality ($v_3 > 0$).

7.2 Arbitration Policy

$$P_t = \{(I_\tau, a_\tau^C, u_\tau, r_\tau) : \tau < t\}, \quad (35)$$

$$\pi_s(I_t) = \text{softmax}(f_\theta(I_t, P_t)), \quad (36)$$

$$a^{arb} = \arg \max_{a \in A} \pi_s(a | I_t). \quad (37)$$

Equations 35 to 37 describe the arbitration process: P_t is a history of past interactions, π_s is a learned policy over who should decide, and a^{arb} selects the action source (human, AI, or committee) that maximizes long-run performance given that history.

8 Calibration Wrapper Integration

HACDU assumes that the AI supplies calibrated uncertainty: the predictive standard deviation σ^A in the risk-sensitive score (Equation 26) and the confidence assessed by the expected calibration error (Equation 45). Modern neural networks, including large foundation models, are often overconfident, so their raw probabilities do not match their empirical accuracy (Guo et al., 2017). If this assumption fails, the lower-confidence-bound penalty, the override threshold, and the trust update all inherit the error. This section therefore treats calibration as an explicit module: a post-hoc wrapper placed between belief fusion (Layer 2) and the collaborative decision engine (Layer 3), which recalibrates the AI's outputs without retraining the underlying model.

8.1 A Three-Tier Wrapper Architecture

The wrapper has three tiers of increasing assumption strength and complexity. All tiers are fitted on a held-out calibration set $\mathcal{D}_{cal}$ that the base model has not seen.

Tier 1: parametric post-hoc calibration. Temperature scaling learns a single scalar $T_c > 0$ that rescales the logits $z_k(x)$ before the softmax (Guo et al., 2017):

$$\hat{p}_k(x) = \frac{\exp(z_k(x)/T_c)}{\sum_{j=1}^K \exp(z_j(x)/T_c)}, \quad T_c^* = \operatorname{argmin}_{T_c > 0} - \sum_{i \in \mathcal{D}_{cal}} \ln \hat{p}_{y_i}(x_i), \quad (38)$$

Because one scalar divides all logits, temperature scaling preserves the ranking of classes and leaves accuracy unchanged; it is cheap and, in Guo et al.'s (2017) comparison, among the most effective post-hoc methods for deep networks. For binary scores, Platt scaling fits a logistic map (Platt, 1999):

$$\hat{p}(x) = \sigma(A s(x) + B), \quad (39)$$

where $s(x)$ is the raw score and A and B are fitted by maximum likelihood on $\mathcal{D}_{cal}$. Isotonic regression instead fits a nondecreasing, piecewise-constant map from scores to probabilities (Zadrozny & Elkan, 2002). It is more flexible and often better calibrated under complex miscalibration, but it can overfit small calibration sets (Niculescu-Mizil & Caruana, 2005).

Tier 2: distribution-free calibration. Split conformal prediction wraps any score function and gives a finite-sample coverage guarantee (Angelopoulos & Bates, 2023). With nonconformity scores $s_i = s(x_i, y_i)$ on n calibration points and a miscoverage level α_{cp} ,

$$\hat{q} = \text{the } [(n+1)(1-\alpha_{cp})]\text{-th smallest of } \{s_i\}_{i=1}^n, \quad \mathcal{C}(x) = \{y: s(x, y) \leq \hat{q}\}, \quad (40)$$

and, when calibration and test points are exchangeable,

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha_{cp}. \quad (41)$$

In HACDU, the size of the conformal set can drive an accept, escalate, or abstain decision: a singleton set supports acceptance, a larger set signals the need for human review, and a very large or empty set raises the escalation probability of Equation 34. Adaptive temperature scaling with conformal prediction (ATS-CP) goes one step further and assigns instance-specific calibrated probabilities to the labels inside the conformal set by choosing an input-dependent temperature, without retraining the base model (Kotelevskii et al., 2025).

For the continuous utility estimates in Equation 26, the same logic applies to normalized residuals. Let R_j be the normalized residual of calibration decision j and $\hat{q}_R$ its conformal quantile; the calibrated standard deviation is then

$$R_j = \frac{|u_j - \hat{u}^A(a_j)|}{\sigma^A(a_j)}, \quad \tilde{\sigma}^A(a) = \kappa_{cp} \sigma^A(a), \quad \kappa_{cp} = \frac{\hat{q}_R}{z_{1-\alpha_{cp}/2}}, \quad (42)$$

so that $\hat{u}^A(a) \pm \hat{q}_R \sigma^A(a)$ is a conformal interval with marginal coverage of at least $1 - \alpha_{cp}$, and $\tilde{\sigma}^A$ replaces σ^A in Equation 26; the factor κ_{cp} expresses that interval in standard-deviation units.

Tier 3: drift-aware calibration. A wrapper fitted before deployment can lose its calibration once the input distribution shifts. Post-hoc methods designed for domain drift calibrate against a range of

perturbed inputs (Tomani et al., 2021), and adaptive conformal inference updates the miscoverage level online so that long-run coverage holds without exchangeability (Gibbs & Candès, 2021). HACDU adds a monitoring rule: the expected calibration error is recomputed over a sliding window of the h most recent labeled decisions, and the wrapper is refitted when it exceeds a tolerance ϵ_{cal} :

$$\text{refit the wrapper at } t \quad \text{if } \widehat{ECE}_t^{(h)} > \epsilon_{cal}. \quad (43)$$

8.2 Calibration Assessment Protocol

Calibration quality is assessed with three complementary tools. The Brier score is the mean squared error of probabilistic predictions and a strictly proper scoring rule (Brier, 1950; Gneiting & Raftery, 2007). The expected calibration error bins predictions by confidence and averages the gap between accuracy and confidence, weighted by bin size (Pakdaman Naeini et al., 2015):

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2, \quad (44)$$

$$ECE \approx \sum_{k=1}^K \frac{|B_k|}{N} |acc(B_k) - conf(B_k)|, \quad (45)$$

where p_i is the predicted probability, y_i the observed outcome, and B_k the k -th confidence bin. Reliability diagrams plot accuracy against confidence by bin and show where the miscalibration lies (Niculescu-Mizil & Caruana, 2005).

Leakage-safe fitting. For in-distribution data, the wrapper is fitted by five-fold cross-fitting: each evaluation instance receives a calibrated prediction from a wrapper fitted on the other four folds, so no calibrator is ever scored on a label it has seen. For out-of-distribution data, the wrapper is fitted once on the calibration set and applied unchanged to the test predictions, so that the evaluation shows how calibration degrades under shift.

8.3 Integration into the HACDU Pipeline

The wrapper changes the framework in four places. First, $\tilde{\sigma}^A$ (Equation 42) replaces σ^A in the risk-sensitive score (Equation 26). Second, the windowed calibration error enters the override threshold of Equation 33, so that a poorly calibrated AI receives less autonomy:

$$\tau_t^{cal} = \tau_t - \theta_c \widehat{ECE}_t^{(h)}, \quad \theta_c \geq 0. \quad (46)$$

Because the adjustment does not depend on S_t or r_t , it preserves the monotonicity constraints of Proposition 3. Third, the calibration-divergence term Λ_t in the trust innovation (Equation 13) is set to the windowed calibration error, so that visible miscalibration lowers reliance. Fourth, HACDU-Bench requires a pre-deployment calibration audit (Section 10.4). Table 3 gives a default choice of wrapper.

Table 3. Default Choice of Calibration Wrapper.

Setting	Default Wrapper	What It Provides	Main Risk
Binary scores	Platt scaling	Smooth, monotone probability map	Misfits if miscalibration is not sigmoid-shaped

Multiclass logits	Temperature scaling	Lower ECE; accuracy and class ranking unchanged	One global factor misses class-specific miscalibration
Ample data, complex miscalibration	Isotonic regression	Flexible monotone map	Overfits small calibration sets
Coverage guarantee required	Split conformal; ATS-CP for set-level probabilities	Marginal coverage $\geq 1 - \alpha_{cp}$ under exchangeability	Holds only at the chosen level and only marginally
Expected distribution shift	Drift-aware calibration with ECE monitoring; adaptive conformal inference	Refitting triggered by monitored error; long-run coverage	Needs labeled feedback; lags after abrupt shifts

8.4 Limitations and Failure Modes

Each tier has known limits. Temperature scaling applies one global factor and cannot correct class-specific or nonlinear miscalibration, which is common under class imbalance. Isotonic regression can overfit when the calibration set is small (Niculescu-Mizil & Caruana, 2005). Conformal coverage is marginal and holds only at the chosen level α_{cp} ; it guarantees neither conditional coverage nor calibration at other levels (Angelopoulos & Bates, 2023), and it breaks when drift violates exchangeability. Calibration also does not transfer across agents. When language models express uncertainty in words, the probabilities attached to verbal markers differ from human readings: a mapping fitted to GPT-4o's outputs assigned 0.677 to “almost certain,” whereas the human value is 0.92 (Duan et al., 2026). Human-derived mapping tables therefore cannot be applied to model-generated verbal uncertainty without recalibration. Finally, the wrapper is estimated from finite data and adds its own error, which the override and escalation layers must absorb.

9 Behavioral Telemetry Operationalization

HACDU treats reliance r_t as a scalar that drives utility weighting and override thresholds, but trust is a latent construct that cannot be observed directly. Without a measurement model, the benchmark metrics ARR, OTR, and UTR (Equations 53 to 55) cannot be computed in a real system. This section specifies how reliance would be inferred from behavioral telemetry, how the inference would be validated, and which privacy safeguards it requires. All elements are proposals for future empirical work.

9.1 Telemetry Indicator Matrix

Table 4 organizes candidate indicators by interaction phase. Acceptance of automated advice is the most widely used behavioral proxy for reliance (Hoff & Bashir, 2015; Lee & See, 2004); the other indicators add information about how a decision was reached. Their relation to trust is expected to be indirect and often nonlinear: a long acceptance lag, for example, may reflect careful verification by a well-calibrated operator rather than low trust.

Table 4. Behavioral Telemetry Indicators by Interaction Phase.

Phase	Indicator	Expected Relation to Reliance	Setting
Acceptance	Recommendation acceptance lag	Longer lag suggests more scrutiny; likely nonlinear	Field and laboratory

Acceptance	Acceptance rate	Standard behavioral proxy of reliance	Field and laboratory
Verification	Interface inspection latency	Dwell time on reasoning, confidence, or explanation panels	Field and laboratory
Verification	Mouse-trajectory features	Track decision confidence more than trust; weak alone	Laboratory and web
Intervention	Manual override frequency	Frequent overrides signal under-trust or vigilance	Field and laboratory
Intervention	Edit distance to the AI recommendation	Partial acceptance and negotiation	Field and laboratory
Escalation	Escalation request rate	Requests for higher-level arbitration	Field
Escalation	Post-decision confidence rating	Self-report anchor for validating proxies	Laboratory; periodic in field

9.2 Eye-Tracking Indicators in Laboratory Studies

In controlled studies, eye tracking provides finer-grained proxies: fixation duration and fixation count on the AI recommendation, revisits to that area, and scanpath length. Longer and more frequent fixations and revisits are expected to indicate verification or uncertainty, fixation counts are expected to rise after trust violations, and shorter scanpaths may indicate higher trust. Eye-movement measures have been used to infer trust as it changes with automation reliability (Lu & Sarter, 2019), which makes them candidates for continuous trust measurement and dynamic trust calibration. Their cost and intrusiveness limit them to laboratory protocols and validation studies.

9.3 Real-Time Trust Estimation Pipeline

The proposed estimator has five stages. (i) *Input*: multimodal telemetry streams, namely interaction logs, timestamps, and, in the laboratory, gaze data. (ii) *Features*: sliding-window aggregates of the indicators in Table 4 over the last h interactions,

$$f_t = \frac{1}{h} \sum_{k=t-h+1}^t x_k, \quad (47)$$

where x_k is the standardized indicator vector of interaction k . (iii) *Estimation*: a measurement model maps the features to an observed trust signal,

$$\hat{\psi}_t = \sigma(\vartheta_0 + \Theta^\top f_t), \quad (48)$$

which is combined with the model-based innovation ψ_t of Equation 13,

$$\bar{\psi}_t = (1 - \kappa_{tel}) \psi_t + \kappa_{tel} \hat{\psi}_t, \quad \kappa_{tel} \in [0, 1], \quad (49)$$

and $\bar{\psi}_t$ replaces ψ_t in the Beta-filter updates of Equation 15. (iv) *Calibration*: the weights are refitted periodically against self-reported trust,

$$\hat{\Theta} = \operatorname{argmin}_{\Theta} \sum_{m \in \mathcal{M}_{SR}} \left(y_m^{SR} - \sigma(\vartheta_0 + \Theta^\top f_{t_m}) \right)^2 + \lambda_{\Theta} \|\Theta\|_2^2, \quad (50)$$

where y_m^{SR} is a self-report, such as the trust-in-automation scale of Jian et al. (2000), rescaled to $[0, 1]$ and collected at sparse probe points $\mathcal{M}_{SR}$. (v) *Output*: the filter returns $\mathbb{E}[r_t]$ and $\operatorname{Var}[r_t]$ (Equation 16),

which feed utility weighting and the uncertainty penalty, respectively; $\mathbb{E}[r_t]$ is also the reliance value used to compute ARR, OTR, and UTR. Other estimators can be substituted at stage (iii), including the quantum-inspired probabilistic trust state of Badhwar et al. (2026), which tracked 349 self-reported ratings from 15 participants in a pilot study.

Reference tools. Existing work suggests that such a pipeline is feasible. VizTrust tracks trust turn by turn in human–AI conversation along four dimensions (competence, integrity, benevolence, and predictability) and displays its evolution on a dashboard (Wang et al., 2025). In the Overcooked-AI environment, models built only on gameplay behavior predicted human trust in AI partners about as accurately as questionnaire-based models (Meimandi et al., 2024). Mouse-trajectory features, by contrast, relate more to decision confidence than to trust and explain too little variance to be used alone (Dechant et al., 2023).

9.4 Validity Protocol

Before telemetry proxies enter any benchmark report, their validity would be established on five criteria. *Convergent validity*: proxies should correlate with self-reported trust, such as the trust-in-automation scale (Jian et al., 2000). *Discriminant validity*: they should be separable from cognitive load, measured for example with the NASA-TLX (Hart & Staveland, 1988), and from situation awareness. *Predictive validity*: they should predict later reliance behavior, including overrides and escalations. *Test–retest reliability*: they should be stable for the same participant under similar task conditions. *Cross-task generalization*: the mapping should transfer across the four task families (PolicyArena, EnergySim, CrisisAgent, and StrategyLab), or task-specific mappings should be reported.

9.5 Privacy and Ethics

Behavioral telemetry is personal data. Logs would be pseudonymized at collection and aggregated before analysis, with retention limited to the study period. Gaze and other physiological signals may count as biometric or health-related data under data-protection law and would require ethics-board approval and explicit informed consent. Monitoring can also change the behavior it measures: operators who know they are observed may verify more or defer less, so the effect of monitoring on trust and acceptance should itself be measured, for example by comparing announced and minimally intrusive logging conditions.

10 HACDU-Bench: A Proposed Benchmark Specification

10.1 Benchmark Objective

A benchmark task is defined by the tuple $\langle \mathcal{E}_{env}, \iota, \mathcal{M}, \mathcal{O} \rangle$: environment $\mathcal{E}_{env}$, interface ι , metric set $\mathcal{M}$, and outcome structure $\mathcal{O}$. The interaction interface, including explanations, uncertainty displays, and escalation options, is treated as first-class, unlike standard machine-learning benchmarks that evaluate AI systems in isolation. Everything in this section is a design proposal: the task families, metrics, and weights have not been instantiated or calibrated in any domain, and no benchmark results are reported.

HACDU-Bench is offered as an open specification for future benchmarking, and its contribution is a transparent, auditable evaluation structure.

10.2 Task Families

PolicyArena: Resource allocation under uncertain social and economic outcomes, with ex-post regret and fairness scoring.

EnergySim: Stochastic dispatch under renewable variability, including small-modular-reactor operations.

CrisisAgent: Rapid-response coordination with incomplete information, including variants emulating spear-phishing and digital-arrest attacks (Gaurav et al., 2025).

StrategyLab: Longer-horizon planning under path dependence — capacity buildout, R&D sequencing, geopolitical scenario analysis.

10.3 Metric Suite, Weight Rationale, and Composite Score

The Human–AI Benchmark Score (HBS) composite score (Eq. 58) is a weighted linear combination of seven metrics. Weights $\omega_1 \dots \omega_7$ must be disclosed with any benchmark report. As a proposed default, $\omega = (0.35, 0.20, 0.15, 0.10, 0.05, 0.10, 0.05)$. Rationale: DQ ($\omega_1 = 0.35$) receives the largest weight because collaborative utility is the primary purpose of any decision system. Reg ($\omega_2 = 0.20$) penalizes systematic loss of value relative to the optimal. The appropriate-reliance rate (ARR) ($\omega_3 = 0.15$) rewards appropriate trust calibration, which drives long-run collaboration quality. The over-trust rate (OTR) ($\omega_4 = 0.10$) and under-trust rate (UTR) ($\omega_5 = 0.05$) penalize miscalibrated reliance, with over-trust weighted higher given its greater safety risk (Parasuraman & Riley, 1997). OP ($\omega_6 = 0.10$) rewards high-quality human overrides as a signal of genuine complementarity. EE ($\omega_7 = 0.05$) rewards efficient escalation. These defaults are a starting point that has not been calibrated empirically; researchers instantiating HACDU-Bench must preregister their weight vector and justify deviations from these defaults.

$$DQ = \frac{1}{N} \sum_{i=1}^N u(a_i, \theta_i), \quad (51)$$

$$Reg = \frac{1}{N} \sum_{i=1}^N (\max_a u(a, \theta_i) - u(a_i, \theta_i)), \quad (52)$$

$$ARR = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{|r_i - r_i^*| \leq \varepsilon\}, \quad (53)$$

where ε is a pre-disclosed tolerance threshold (default $\varepsilon = 0.10$) and r_i^* is the normative reliance level derived from a domain expert panel or from the oracle policy. The three rates partition the decisions, so $ARR + OTR + UTR = 1$.

$$OTR = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{r_i > r_i^* + \varepsilon\}, \quad (54)$$

$$UTR = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{r_i < r_i^* - \varepsilon\}, \quad (55)$$

$$OP = P(\text{correct} \mid \text{override}), \quad (56)$$

$$EE = \frac{\Delta DQ \mid \text{escalation}}{c_{esc}}, \quad (57)$$

$$HBS = \omega_1 DQ - \omega_2 Reg + \omega_3 ARR - \omega_4 OTR - \omega_5 UTR + \omega_6 OP + \omega_7 EE. \quad (58)$$

Equations 51 to 58 define the HACDU-Bench metrics and composite score: DQ and Reg capture outcome quality, appropriate-reliance rate (ARR)/over-trust rate (OTR)/under-trust rate (UTR) capture trust calibration, OP measures the value of human overrides, EE measures escalation efficiency, and Human–AI Benchmark Score (HBS) aggregates them with transparent weights $\omega_1 \dots \omega_7$.

10.4 Deployment Prerequisites

Two checks are prerequisites for any HACDU-Bench report. **Pre-deployment calibration audit:** the AI component's ECE, Brier score, and reliability diagram are reported before and after the wrapper of Section 8, on in-distribution and shifted data, together with the drift tolerance ϵ_{cal} . **Proxy validity check:** the telemetry-based reliance estimator of Section 9 must meet the validity criteria of Section 9.4 in the target task family before ARR, OTR, and UTR are reported. Results that skip either check should be labeled provisional.

11 Numerical Illustration of the HACDU Framework

This section gives one simplified numerical example to show how belief fusion, trust-dependent thresholds, value aggregation, and override logic interact. All values are chosen for exposition. The example is didactic, carries no statistical or inferential weight, and is not empirical evidence for the framework.

Consider a single stylized decision in an energy-dispatch setting. The system must choose among three actions: a_1 = maintain current dispatch, a_2 = increase reserve margin, and a_3 = curtail industrial demand. The latent state is $\theta \in \{\text{low risk, medium risk, high risk}\}$. The human operator, drawing on contextual judgment, holds the belief $b_t^H = (0.20, 0.50, 0.30)$, while the AI system, using forecast models and calibrated uncertainty estimates, holds $b_t^A = (0.10, 0.30, 0.60)$.

At time t , let reliance be $r_t = 0.58$ and the stake level $S_t = 0.80$ on a normalized $[0, 1]$ scale. Suppose the context-adaptive fusion weights are $w^H = 0.42$ and $w^A = 0.58$. Applying the log-linear pooling rule in Equation 23 and normalizing gives the collaborative belief $b_t^C \approx (0.14, 0.39, 0.47)$, which moves toward the AI while keeping a substantial human contribution. The symmetric KL divergence between the two beliefs (Equation 9) is $D_{KL}^S(b_t^H \parallel b_t^A) \approx 0.19$ nats.

Assume the estimated action values are $Q^H = \{a1: 0.62, a2: 0.74, a3: 0.49\}$ and $Q^A = \{a1: 0.55, a2: 0.68, a3: 0.71\}$, with process costs $c_{exec}(a1)=0.02$, $c_{exec}(a2)=0.04$, and $c_{exec}(a3)=0.07$. The collaborative values from Equation 17 are therefore $Q^C(a1)=0.42 \times 0.62 + 0.58 \times 0.55 - 0.02 = 0.56$,

$Q^C(a_2)=0.42 \times 0.74 + 0.58 \times 0.68 - 0.04 = 0.67$, and $Q^C(a_3)=0.42 \times 0.49 + 0.58 \times 0.71 - 0.07 = 0.55$. The collaborative recommendation is thus a_2 , increasing the reserve margin.

Now consider the override logic. Let the threshold function in Equation 33 take the illustrative form $\tau_t = 0.40 - 0.30S_t + 0.40r_t$. At the baseline condition ($S_t = 0.80$, $r_t = 0.58$), the threshold is $\tau_t = 0.40 - 0.24 + 0.232 = 0.392$. Because $D_{KL}^S \approx 0.19 < 0.392$, the disagreement stays below the threshold, and Equation 32 retains the collaborative recommendation a_2 without human override.

Next, consider a counterfactual in which recent AI prediction errors have lowered reliance to $r_t = 0.20$ and the situation has escalated to maximum stakes, $S_t = 1.00$. The threshold falls to $\tau_t = 0.40 - 0.30 + 0.08 = 0.18$. Because $D_{KL}^S \approx 0.19 > 0.18$, Equation 32 now assigns authority to the human, whose preferred action is $a^H = \operatorname{argmax}_a Q^H(a) = a_2$. (A drop in reliance alone, with $S_t = 0.80$, would give $\tau_t = 0.24$ and no override.) Here the override changes who holds decision authority but not the action, because the human's choice and the collaborative recommendation coincide; with wider disagreement, escalation under Equation 34 would become more likely. The example illustrates how the components interact within one decision cycle; it does not show that the framework improves decisions.

12 Propositions and Future Empirical Research Agenda

12.1 Testable Propositions

The framework implies five propositions. Each is derived from the formal structure in Sections 3 to 10, is stated in falsifiable form, and has not been tested in this article. Propositions 1 to 3 in Sections 4 to 7 are formal results that hold under their stated assumptions; P1 to P5 below are empirical predictions.

- P1: Holding AI accuracy and human expertise constant, adaptive reliance dynamics (Equations 13 to 16) are expected to yield higher collaborative value added (CVA) than static-weighting hybrids. *Status: to be tested empirically.*
- P2: Averaged over stakes and shift severity, depth-optimized explanation policies (Equation 28 and Proposition 2) are expected to achieve a higher Human–AI Benchmark Score (HBS) than fixed-depth explanations. *Status: to be tested empirically.*
- P3: Relative to fixed-threshold override, uncertainty-aware override (Equations 32 to 34) is expected to lower the over-trust rate (OTR) without raising the under-trust rate (UTR). *Status: to be tested empirically.*
- P4: Fairness-regularized policies (Equation D.13) are expected to reduce utility and reliance disparities at a bounded cost in mean decision quality (DQ). *Status: to be tested empirically.*
- P5: Collaborative value added is expected to be positive only when human and AI error sets are sufficiently non-redundant and arbitration routes authority accurately enough to offset process overhead (Proposition 1). *Status: to be tested empirically.*

12.2 Proposed Human-Subject Validation Protocol

To support future replication, this subsection proposes a human-subject validation protocol for HACDU-Bench. The protocol has not been executed. Before any data collection, it would be preregistered on the Open Science Framework (OSF), fixing task sequences, interface configurations, and statistical tests so that trials across research groups remain comparable.

Participants would be recruited from graduate students and early-career professionals with quantitative training (target $N \approx 90$) and randomly assigned to three between-subject conditions: Human-only, AI-only, and HACDU-assisted. Each participant would complete 40 resource-allocation decisions under uncertainty, with payoffs defined by a PolicyArena task instance. In the HACDU condition, the AI component would be a calibrated probabilistic model trained on held-out synthetic data that participants never see directly.

The interface would log all recommendations, explanations, response times, overrides, and confidence ratings. Primary outcomes would be (i) mean decision quality (DQ) per participant, (ii) regret relative to an oracle benchmark, (iii) the appropriate-reliance rate (ARR), over-trust rate (OTR), and under-trust rate (UTR) computed with $\varepsilon = 0.10$, and (iv) subjective trust and workload scales administered after each block of tasks. All metrics would use the HACDU-Bench definitions (Equations 51 to 57).

P1 to P3 could be tested with the mixed-effects specifications in Section 12.3. Decision quality and regret would be modeled with linear mixed models (Equation 59), and binary reliance choices and override triggers with logistic generalized linear mixed models (Equation 60), each with crossed random intercepts for participants (u_i) and tasks (v_s). Planned contrasts would compare the HACDU-assisted condition with the Human-only and AI-only conditions, and the Holm–Bonferroni method would adjust for multiple comparisons across the primary metrics (DQ, regret, ARR, OTR, and UTR). By Equation 61, $N = 90$ ($n = 30$ per condition, 40 tasks each, 3,600 decisions) would give about 82% power to detect a standardized effect of $\Delta \geq 0.35$ at $\alpha = 0.05$ before multiplicity adjustment, assuming $\rho_{ICC} \leq 0.20$. This is a planning calculation, not a result.

12.3 Proposed Statistical Models

To account for repeated measures (each participant i evaluates K scenarios), the sample size per arm would be adjusted with the cluster design effect $DE = 1 + (K - 1)\rho_{ICC}$:

$$Y_{isc} = \beta_0 + \beta_c + u_i + v_s + \varepsilon_{isc}, \quad (59)$$

$$r_{isc} \sim \text{Bernoulli}(\sigma(\eta_0 + \eta_c + \varphi_i + \psi_s)), \quad (60)$$

$$N_{arm} \approx \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\Delta^2 K} (1 + (K - 1)\rho_{ICC}). \quad (61)$$

where Δ is the minimal detectable effect size (standardized mean difference in DQ or ARR), σ^2 is the total outcome variance, $\rho_{ICC} = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2) \in [0, 1]$ is the intraclass correlation coefficient that captures between-participant heterogeneity, and α would be adjusted across the primary endpoints with the Holm – Bonferroni procedure, as in Section 12.2.

Under these assumptions, the proposed allocation ($N \approx 90$; $n = 30$ per arm, $K = 40$ tasks, 3,600 decisions) would reach about 82% power for $\Delta \geq 0.35\sigma$ at $\rho_{ICC} = 0.20$ before multiplicity adjustment, and more under weaker clustering. These are planning values for future studies; no data have been analyzed.

13 Discussion

The main contribution of HACDU is to move the focus of hybrid-intelligence research from isolated system accuracy to collaborative value under uncertainty. By making the collaborative process the scientific object, HACDU turns previously implicit design decisions, such as how much to trust, when to explain, when to override, and when to escalate, into explicit, estimable, and optimizable quantities. The article does not show that HACDU improves decisions; it supplies the formal structure, propositions, and benchmark specification needed to test whether it does.

13.1 Pathways for Empirical Validation

1. **Controlled laboratory studies.** Operationalize HACDU-Bench as a laboratory protocol with synthetic but ecologically valid tasks (Section 12.2). P1 to P3 can each be tested in within-subject or between-group designs, sized with the planning calculation in Section 12.3 and preregistered before data collection.
2. **Field deployment in regulated domains.** Deploy in energy dispatch centers, clinical triage units, or financial trading desks. Field studies can test P4 (fairness) and P5 (complementarity) under realistic workload and institutional incentive structures.
3. **Multi-site cross-domain replication.** A fully powered validation requires replication across at least three distinct domains, with code, seeds, and interaction traces shared under the HACDU reporting protocol, enabling meta-analytic synthesis.
4. **Longitudinal adaptive studies.** The trust dynamics (Equations 13 to 16) predict how reliance evolves over time. Longitudinal designs that track operator–AI dyads over weeks or months can test the Beta-filter updates of a_t and b_t (Equation 15) against observed reliance trajectories.

13.2 Limitations

No empirical data. This article collects and analyzes no data. The only numbers it contains are in the illustrative example of Section 11; they were chosen for exposition and carry no inferential weight. No benchmark comparison, significance test, confidence interval, or effect size is reported, and none of the propositions in Section 12 has been tested.

No human participants. Trust, reliance, cognitive load, and explanation benefit enter the framework as theoretical constructs. Their functional forms (Equations 13 to 16, 28, and 29) follow prior literature but have not been validated against human behavior, and synthetic agents cannot substitute for such validation.

No cross-domain validation. The four task families of HACDU-Bench are proposals. None has been instantiated, calibrated, or deployed, so no claim is made about which domain would show the largest CVA, how quickly trust converges across operator populations, or whether the families adequately cover real collaborative environments.

Scope of the mathematical model. The framework assumes beliefs representable as distributions over a common state space Θ and utilities $u(a, \theta)$ that are well defined and measurable. In practice, human beliefs may be non-probabilistic, partly verbal, or culturally embedded. Domain-specific instantiation will require substantial engineering judgment; the equations are structural templates, not plug-and-play specifications.

Latent-space and semantic asymmetry. A related assumption is that human and AI beliefs, b_t^H and b_t^A , map cleanly onto a shared, mutually interpretable state space Θ . Modern foundation models, however, represent information in high-dimensional, opaque latent spaces whose relational structure may have no counterpart in human cognitive frameworks. Projecting such representations onto Θ may discard or distort information, and log-linear pooling (Equations 22 to 25) may then force numerical agreement between structurally incompatible representations, producing semantic dissonance rather than genuine complementarity.

Reliance and latent state measurement. The trust dynamics model (Equations 13 to 16) treats r_t as a scalar. Section 9 proposes how it would be inferred from behavioral telemetry, but the proxies' validity is bounded: they capture reliance behavior rather than trust itself, their mapping to trust may be nonlinear and task-specific, and none has yet been validated against the criteria of Section 9.4. Until that validation is done, ARR, OTR, and UTR computed from telemetry are provisional, and measurement error in $E[r_t]$ propagates into every downstream metric.

Strategic gaming of behavioral telemetry. Because the framework estimates latent reliance r_t from passive telemetry, such as inspection latency and override frequency (Section 9), and uses the estimate to set the autonomy envelope through τ_t (Equation 33), its measures are exposed to Goodhart's law: when a measure becomes a target, it ceases to be a good measure (Strathern, 1997). Operators who recognize that their interaction metrics govern their supervisory workload may game them, for example by inflating verification times or withholding necessary overrides, and thereby steer the Beta-filter updates of a_t and b_t (Equation 15). Such distortion would undermine the validity of the telemetry estimator; the observer effect noted in Section 9, where monitoring changes the behavior it measures, is a milder form of the same problem. Safeguards, such as audit sampling against outcome data or periodic self-report recalibration (Equation 50), have not been designed or tested.

AI calibration assumptions. The framework assumes calibrated uncertainty estimates $\sigma^A(a)$ (Equation 26) with stable expected calibration error (Equation 45), an assumption that modern foundation models often violate. Section 8 adds a post-hoc calibration wrapper, but the wrapper is itself estimated from finite data and introduces additional uncertainty: temperature scaling cannot fix class-specific miscalibration, isotonic regression can overfit, conformal guarantees are marginal and level-specific, and every tier degrades under abrupt drift until it is refitted.

Computational latency and real-time scalability. Although HACDU is modular, running its pipeline in real time, including log-linear belief fusion (Equations 22 to 25), conformal recalibration under drift (Section 8), and, when used, the variational bound of Appendix D, adds computational overhead that compounds across layers. In ultra-low-latency settings, such as algorithmic trading or rapid tactical crisis response, the cumulative processing burden may violate the time constraint captured by c_{time} in Equation 6 and could outweigh the net utility gain of collaboration. The framework does not yet specify latency budgets or anytime approximations for such settings.

Ethical, legal, and fairness constraints. The fairness objective (Equation D.13) addresses distributional but not procedural fairness, and does not cover explainability obligations under AI regulation (e.g., the transparency and human-oversight provisions in Articles 13 and 14 of the EU AI Act) or legal liability allocation for human-override decisions. HACDU-Bench currently includes no regulatory compliance metrics. Institutional deployment in regulated sectors — healthcare, finance, critical infrastructure — requires alignment with sector-specific governance frameworks outside this article's scope. The treatment of fairness and regulation in this article is conceptual, and no mapping to specific compliance requirements is attempted.

Accountability under dynamic override. Proposition 3 implies that, as the stake level S_t rises, the uncertainty-aware override becomes more likely to return control to the human when epistemic divergence exceeds τ_t . This mechanism creates legal and accountability ambiguity. Requiring an operator to take immediate control during an unfolding crisis, particularly when situational awareness is limited by the information compression of the explanation policy (d^* , Equation 29'), raises unresolved questions about negligence, fiduciary and professional liability, and compliance with frameworks such as the EU AI Act. The framework allocates decision authority but not legal responsibility, and how the two should align is left to legal and governance research.

Composite score sensitivity. Human–AI Benchmark Score (HBS) (Equation 58) is a weighted linear combination; results are sensitive to weight choices. Linear aggregation masks trade-offs between metric dimensions. Researchers should report individual metric profiles alongside HBS and conduct sensitivity analyses over plausible weight ranges before drawing comparative conclusions.

13.3 Potential Stakeholder Implications

If its propositions are supported empirically, HACDU could offer value to three stakeholder groups:

(1) For Academics and Decision Scientists:

Unified Analytical Paradigm: If validated, HACDU could bridge the long-standing divide between behavioral human factors (trust, cognitive fatigue) and formal operations research (POMDP, Bayesian inference), providing a common mathematical language for hybrid intelligence.

Replication Standard: If validated, HACDU-Bench (Section 10) could provide standardized task environments, reporting rules, and separate calibration metrics that support benchmark-driven replication across studies.

Testable Theoretical Results: If validated, P1 to P5, which translate the formal results of Propositions 1 to 3 on error complementarity, optimal explanation depth, and override monotonicity into predictions, would give experimental work a tested theoretical basis.

(2) For Policymakers and AI Regulators:

Regulatory Auditability: If validated, the framework could make human oversight explicit and auditable, a concern of governance frameworks such as the EU AI Act and the NIST AI RMF. The dynamic override threshold $\tau = g(S_t, D_{KL}, r_t)$ makes human intervention more likely as stakes rise (Proposition 3); whether it satisfies any specific regulatory requirement has not been assessed.

Trust Calibration Metrics: If validated, the proposed metrics (ARR , OTR , UTR) could help auditors check whether deployed AI decision aids induce automation complacency or disuse in critical public infrastructure.

Distributional Fairness Governance: If validated, Equations D.11 to D.13 could let regulators set bounded disparity constraints (Δ_u , Δ_r) across protected groups; the resulting cost in decision quality is an open empirical question (P4).

(3) For Investors, Asset Managers, and Enterprise Executives:

Return Measured by Collaborative Value Added: If validated, and once the J values are estimated, CVA (Equation 20) could express the net utility of pairing human traders or operators with AI models, relative to standalone deployment, as an auditable economic metric.

Tail-Risk Exposure: If validated, the risk-sensitive collaborative score (LCB , Equation 26) and uncertainty-aware override could limit exposure during severe distribution shifts (e.g., flash crashes or geopolitical shocks); this property remains to be tested.

Operational Cost: If validated, the closed-form explanation policy (Proposition 2) could limit cognitive overload among human analysts by matching explanation depth to uncertainty and expertise.

13.4 Design Principles

The framework follows five design principles:

- (1) **Theoretical Parsimony and Modular Decoupling:** HACDU treats each subsystem as a replaceable module under the principles of Section 4. The body of the article develops only the core decision pipeline (belief fusion, calibration, trust filtering, and override control); variational inference, the collaborative POMDP, reinforcement learning, robust optimization, and fairness regularization are stated in Appendix D.
- (2) **Literature-Grounded Behavioral Assumptions:** The behavioral components draw on established empirical literature rather than stylized rational-actor assumptions. Dynamic trust updates (ψ_t , Equation 13) reflect documented complacency and misuse dynamics (Lee & See, 2004; Parasuraman & Riley, 1997), and cognitive processing cost is modeled as superlinear

$(c_{cog}(d) = \frac{1}{2}c_2d^2)$ following Cognitive Load Theory (Miller, 2019; Sweller, 2011). These forms remain hypotheses to be validated.

- (3) **Non-Intrusive Behavioral Proxies:** Because internal mental states are unobservable, latent reliance (ψ_t, r_t) would be operationalized in deployed systems through passive telemetry: interface inspection latency, recommendation verification lag, and manual override frequency. Section 9 specifies these proxies and their validation.
- (4) **Transparent Composite Scoring:** Because the Human–AI Benchmark Score (*HBS*, Equation 58) depends on its weights, the specification requires any future benchmark study to disclose its weight vector, report the separate calibration metrics (*ARR*, *OTR*, *UTR*), and analyze rank stability over weight vectors sampled from the unit simplex Δ^6 .
- (5) **Institutional and Regulatory Governance:** The dynamic override boundary $\tau = g(S_t, D_{KL}, r_t)$ (Proposition 3) and the fairness objective $(J_c^{fair}$, Equation D.13) are designed to support human oversight, auditable intervention thresholds, and disparity safeguards of the kind sought by AI governance frameworks (e.g., the EU AI Act and the NIST AI RMF), although no compliance mapping is attempted here.

14 Conclusion

This article has presented HACDU, a conceptual framework for human–AI collaborative decision-making under uncertainty, together with HACDU-Bench, a proposed benchmark specification. The framework formalizes collaborative belief states, trust dynamics, utility fusion, explanation-policy optimization, uncertainty-aware override, escalation logic, and benchmark metrics in one mathematical structure. Its central argument is that human–AI systems should be evaluated by collaborative value added, the degree to which the coupled system exceeds its best individual component, rather than by model accuracy alone. The aim is to give subsequent research a shared language, testable propositions, and an operational evaluation structure. The article provides no empirical validation; its propositions must be tested through human-subject experiments, cross-domain replication, and field deployment.

14.1 Data Availability Statement

No empirical data were collected or analyzed for this study. The numerical illustration in Section 11 is fully specified in the text and can be reproduced from the values given there; no other data or code underlie the article.

References

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/2200000101>

- Badhwar, R., Walji, Z., Dave, P., & Park, J. (2026). A quantum-inspired framework for real-time trust estimation in human–AI interaction: A pilot study. *Scientific Reports*. Advance online publication. <https://doi.org/10.1038/s41598-026-67931-x>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 81, pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445717>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Chan, L., & Wong, W. K. (2012). Automated trading with genetic-algorithm neural-network risk cybernetics: An application on FX markets. *Finamatrix Journal*, February.
- Chan, L., & Wong, W. K. (2013). Expert advisor development on MT4/MT5 for automated algorithmic trading on EURUSD M1 data. *Finamatrix Journal*, September.
- Dechant, M., Poeller, S., Hosp, B., Lukashova-Sanz, O., Sipatchin, A., & Wahl, S. (2023). Assessing the relationship between subjective trust, confidence measurements, and mouse trajectory characteristics in an online task. *arXiv*. <https://doi.org/10.48550/arXiv.2310.16632>
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://doi.org/10.48550/arXiv.1702.08608>
- Duan, J., Liu, Z., Liu, Z., Xu, K., & Chen, T. (2026). “Very likely” means “uncertain”? How LLMs diverge from humans in linguistic uncertainty quantification. *arXiv*. <https://doi.org/10.48550/arXiv.2610.00083>
- Gaurav, A., Pan, V. S.-H., Arya, V., Attar, R. W., Alhazmi, A. H., Alhomoud, A., Gupta, B. B., & Chui, K. T. (2025). Defense-optimized BERT architecture against digital arrest attacks in intelligent systems. *Telecommunication Systems*, 88(3), Article 107. <https://doi.org/10.1007/s11235-025-01337-4>
- Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 1660–1672). Curran Associates.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. <https://proceedings.mlr.press/v70/guo17a.html>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Advances in Psychology, Vol. 52, pp. 139–183). North-Holland. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>

- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007), 453–461. <https://doi.org/10.1098/rspa.1946.0056>
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809477>
- Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 467–474). International Foundation for Autonomous Agents and Multiagent Systems. <https://dl.acm.org/doi/10.5555/2343576.2343643>
- Karim, K. S., Islam, M. E., Ibrahim, A. M., Pan, S.-H., & Rahman, M. M. (2024). Online marketing trends and purchasing intent: Advances in customer satisfaction through PLS-SEM and ANN approach. *Advances in Decision Sciences*, 27(4), 24–54. <https://doi.org/10.47654/v27y2023i4p24-54>
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.1312.6114>
- Kotelevskii, N., Guizani, M., Moulines, E., & Panov, M. (2025). Adaptive temperature scaling with conformal prediction. *arXiv*. <https://doi.org/10.48550/arXiv.2505.15437>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lu, Y., & Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 49(6), 560–568. <https://doi.org/10.1109/THMS.2019.2930980>
- Ly, S., Pho, K.-H., Pan, S.-H., & Wong, W.-K. (2025). Newton-Raphson method: Overview and applications. *Advances in Decision Sciences*, 28(3), 52–78. <https://doi.org/10.47654/v28y2024i3p52-78>
- Meimandi, K. J., Bolton, M. L., & Beling, P. A. (2024). Action over words: Predicting human trust in AI partners through gameplay behaviors. In *Proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 563–568). IEEE. <https://doi.org/10.1109/RO-MAN60168.2024.10731166>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. <https://doi.org/10.1145/1102351.1102430>
- Pakdaman Naeini, M., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1, pp. 2901–2907). <https://doi.org/10.1609/aaai.v29i1.9602>

- Pan, S.-H., & Wang, S.-C. (2020). Enhancing the reliability of cellular internet of things through agreement. *Applied Sciences*, 10(21), Article 7699. <https://doi.org/10.3390/app10217699>
- Pan, S.-H., & Wang, S.-C. (2021). Identifying vehicles dynamically on freeway CCTV images through the YOLO deep learning model. *Sensors and Materials*, 33(5), 1517–1530. <https://doi.org/10.18494/SAM.2021.3236>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12), 1053–1058. <https://doi.org/10.1145/361598.361623>
- Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). Dover.
- Simon, H. A. (1957). *Models of man: Social and rational*. Wiley.
- Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In J. Fürnkranz & T. Joachims (Eds.), *Proceedings of the 27th International Conference on Machine Learning* (pp. 1015–1022). Omnipress. <https://doi.org/10.48550/arXiv.0912.3995>
- Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, 5(3), 305–321. [https://doi.org/10.1002/\(SICI\)1234-981X\(199707\)5:3<305::AID-EURO184>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4)
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., & Buettner, F. (2021). Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10119–10127). <https://doi.org/10.1109/CVPR46437.2021.00999>
- Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Springer. <https://doi.org/10.1007/978-3-540-71050-9>
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.
- Wang, X., Tulk Jesso, S., Kojaku, S., Neyens, D. M., & Kim, M. S. (2025). VizTrust: A visual analytics tool for capturing user trust dynamics in human–AI communication. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–10). ACM. <https://doi.org/10.1145/3706599.3719798>
- Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). ACM. <https://doi.org/10.1145/775047.775151>

Appendix A. Notation and Abbreviations

Table 1. Complete notation reference for HACDU. Symbols listed in order of first appearance.

Symbol	Domain	Meaning
θ, θ_t	State	Latent world state
o_t^H, o_t^A	Observation	Human and AI observations at time t
z_t^H, z_t^A	Context	Human expert priors; AI latent memory
b_t^H, b_t^A, b_t^C	Belief	Human, AI, collaborative belief distributions
$r_t \in [0,1]$	Trust	Dynamic reliance state / probability human defers to machine
$\psi_t \in [0,1]$	Trust	Instantaneous trust feedback innovation signal
$a_t, b_t > 0$	Trust Prior	Beta distribution parameters tracking collaborative history
$\gamma_r \in (0,1]$	Parameter	Trust memory retention (discount) factor
$\lambda_r \in (0,1]$	Parameter	Dynamic reliance learning rate
$\theta_Q, \theta_A, \theta_S, \theta_P, \theta_C$	Weights	Sensitivity coefficients for trust innovation components
e_t	Escalation	Binary escalation state
$\delta_t, D_{KL}, W_1, D_{rank}$	Disagreement	Overall, KL, Wasserstein, rank divergences
I_t	Info state	Full collaborative information state
Q^H, Q^A, Q^C	Value	Human, AI, collaborative action-value functions
w^H, w^A	Weight	Context-adaptive fusion weights ($w^H + w^A = 1$)
π_C, π_s	Policy	Collaborative policy; arbitration policy
$C_{proc}, C_{err}, C_{time}$	Cost	Frictional process, error violation, and execution time costs
E_t, C_t, D_t	Interaction	Explanation burden, cognitive load, and dialogue latency
CVA, NCG	Performance	Collaborative value added; normalized collaborative gain
DQ, Reg, ARR, OTR, UTR, OP, EE	Benchmark	HACDU-Bench component metrics (Equations 51–57)
HBS	Composite	HACDU-Bench composite score (Equation 58)
$\omega_1 \dots \omega_7$	Weights	Composite score weights (mandatory disclosure at submission)
$q_\phi(\theta)$	Variational Inference	Tractable variational approximation of the collaborative belief
ϕ	Parameter	Variational parameters (e.g., mean and covariance μ_ϕ, Σ_ϕ)
L_{c-ELBO}	Objective	Collaborative Evidence Lower Bound (Equations D.2 and D.2')
$H(q_\phi)$	Information Theory	Differential entropy of the variational distribution
ε	Tolerance	Reliance tolerance in ARR, OTR, and UTR (default 0.10)
ϵ_0	Parameter	Stability constant in NCG (Equation 21)
$\mathcal{E}^H, \mathcal{E}^A$	Error sets	Human and AI error sets (Proposition 1)
$\mathcal{T}$	POMDP	Transition kernel of the latent state (Equation D.3)
T_c, Platt A, B	Calibration	Temperature and Platt-scaling parameters (Section 8)
$\hat{q}, \alpha_{cp}, \mathcal{C}(x)$	Conformal	Conformal quantile, miscoverage level, and prediction set
$\tilde{\sigma}^A, \kappa_{cp}$	Calibration	Calibrated AI standard deviation and conformal scale factor

$ECE_t^{(h)}, \epsilon_{cal}, \theta_c$	Calibration	Windowed calibration error, refit tolerance, and threshold sensitivity
x_k, f_i, h	Telemetry	Indicator vector, windowed features, and window length
$\hat{\psi}_i, \bar{\psi}_i, \kappa_{tel}, \vartheta$	Telemetry	Observed and fused trust signals, fusion weight, and measurement weights

Appendix B. Implementation Checklist

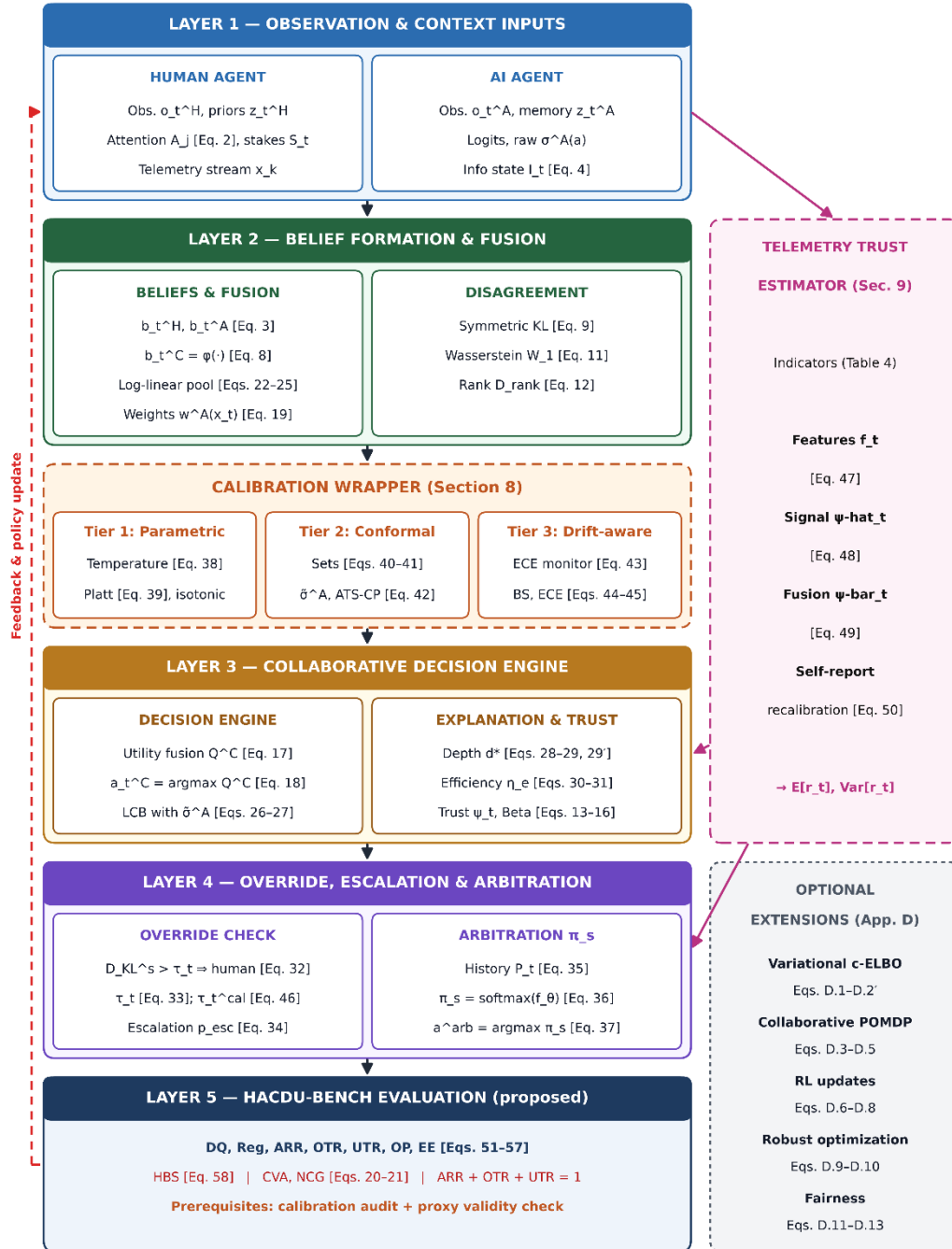
The checklist is ordered by logical dependency: each step uses the outputs of the steps before it. In particular, the override threshold and escalation rules (step 8) are parameterized only after the calibration wrapper (step 6) and the reliance estimator (step 7) are in place, because the calibration-aware threshold τ_t^{cal} (Equation 46) depends on the windowed calibration error and the threshold τ_t (Equation 33) depends on $\mathbb{E}[r_t]$.

1. Define the action space A , joint state space S , and outcome utility $u(a, \theta)$.
2. Specify human observation channels o^H and AI channels o^A and their latent memories.
3. Choose a belief representation for each agent (parametric, particle, or variational).
4. Select a divergence metric (KL, Wasserstein, or rank-based) and trust-update model (Equations 13 to 16).
5. Define the explanation-depth policy (Equation 28) and interaction logging schema.
6. Fit the calibration wrapper on a held-out calibration set, report ECE, Brier score, and reliability diagrams before and after calibration, and set the drift tolerance ϵ_{cal} (Section 8).
7. Specify the telemetry indicators and the reliance estimator, and validate the proxies against self-reported trust before computing ARR, OTR, and UTR (Section 9).
8. Parameterize the dynamic override threshold $\tau = g(S_t, D_{KL}, r_t)$ and escalation rules (Equations 32 to 34). Verify the monotonicity constraints: positive sensitivity to reliance ($\partial g / \partial r_t > 0$), which grants the AI autonomy under high trust, and negative sensitivity to stakes ($\partial g / \partial S_t \leq 0$), which tightens human supervisory control in high-consequence scenarios, with explicit clamping to $[\tau_{min}, \tau_{max}]$.
9. Preregister benchmark metrics, baselines, composite weights $\omega_1 \dots \omega_7$, and tolerance ϵ .
10. Report DQ, Reg, expected calibration error (ECE), appropriate-reliance rate (ARR), over-trust rate (OTR), under-trust rate (UTR), OP, EE, and fairness metrics (Equations D.11 to D.12).
11. Release code, seed, and interaction traces to make results replayable and auditable.

Appendix C. HACDU Architecture Schematic

Figure 1 gives a schematic overview of the HACDU architecture. The five core layers (observation and context inputs, belief fusion, the collaborative decision engine, override and arbitration, and HACDU-Bench evaluation) are developed in Sections 3 to 7 and 10. The calibration wrapper (Section 8) sits between belief fusion and the decision engine, the telemetry-based trust estimator (Section 9) feeds the reliance state, and the optional extensions are collected in Appendix D.

Figure 1. HACDU Architecture: Core Pipeline, Calibration Wrapper, Telemetry-Based Trust Estimation, and Optional Extensions.



Appendix D. Optional Extensions

This appendix collects the extensions of HACDU that are not needed to operate the core decision pipeline. They show the theoretical completeness of the framework and indicate directions for future work: scalable inference in high-dimensional state spaces (D.1), sequential optimization (D.2 and D.3), robustness to distribution shift (D.4), and fairness regularization (D.5).

D.1 Variational Collaborative Inference

In complex, continuous, or high-dimensional state spaces Θ , computing the exact normalizing partition function Z in Equation 25 is analytically intractable. To enable scalable real-time Bayesian fusion, HACDU introduces a variational approximation $q_\phi(\theta)$ parameterized by ϕ (e.g., a Gaussian variational family $N(\mu_\phi, \Sigma_\phi)$).

The optimal variational parameter ϕ^* is obtained by minimizing the reverse Kullback–Leibler (KL) divergence to the target collaborative belief $b_t^c(\theta)$:

$$\phi^* = \arg \min_{\phi} D_{KL}(q_\phi(\theta) || b_t^c(\theta)). \quad (D.1)$$

Expanding the divergence using the log-linear pooled formulation from Equation 24 yields:

$$\begin{aligned} D_{KL}(q_\phi(\theta) || b_t^c(\theta)) &= \int_{\Theta} q_\phi(\theta) \ln \frac{q_\phi(\theta)}{b_t^c(\theta)} d\theta \\ &= E_{q_\phi}[\ln q_\phi(\theta)] - E_{q_\phi}[w^H \ln b_t^H(\theta) + w^A \ln b_t^A(\theta) - \ln Z] \\ &= -H(q_\phi(\theta)) - w^H E_{q_\phi}[\ln b_t^H(\theta)] - w^A E_{q_\phi}[\ln b_t^A(\theta)] + \ln Z \end{aligned}$$

where $H(q_\phi(\theta)) \triangleq -E_{q_\phi}[\ln q_\phi(\theta)]$ is the differential entropy of the variational distribution. Because $D_{KL}(\cdot) \geq 0$, rearranging terms establishes the **Collaborative Evidence Lower Bound (L_{c-ELBO})** on the log partition function $\ln Z$:

$$\ln Z \geq L_{c-ELBO}(\phi) \triangleq w^H E_{q_\phi}[\ln b_t^H(\theta)] + w^A E_{q_\phi}[\ln b_t^A(\theta)] + H(q_\phi(\theta)). \quad (D.2)$$

Equivalently, when individual agent posteriors are factored through a common prior $p(\theta)$ and conditional observation likelihoods $p(o_t^H|\theta)$ and $p(o_t^A|\theta)$, the collaborative lower bound can be rewritten as:

$$L_{c-ELBO}(\phi) = w^H E_{q_\phi}[\ln p(o_t^H|\theta)] + w^A E_{q_\phi}[\ln p(o_t^A|\theta)] - D_{KL}(q_\phi(\theta)||p(\theta)). \quad (D.2')$$

Equations D.1 to D.2 establish a direct bridge between log-linear belief fusion and stochastic variational inference: the variational agent maximizes weighted cross-agent data reconstruction while regularizing against epistemic collapse via the entropy / prior penalty. The bound can be efficiently optimized via Monte Carlo gradient estimation with the reparameterization trick.

D.2 Collaborative Partially Observable Markov Decision Process

$$\langle \mathcal{S}, \mathcal{A}, \Omega, \mathcal{T}, O, R, \gamma, T \rangle, \quad \mathcal{S} = \Theta \times [0,1], \quad (\text{D.3})$$

$$b_{t+1}(\theta') \propto O(o_{t+1} \mid \theta', a_t) \sum_{\theta \in \Theta} \mathcal{T}(\theta' \mid \theta, a_t) b_t(\theta), \quad (\text{D.4})$$

$$V_t^*(b, r) = \max_{a \in \mathcal{A}} \{R(b, a, r) + \gamma \mathbb{E}_{o, r'}[V_{t+1}^*(b^{a,o}, r')]\}, \quad V_{T+1}^* \equiv 0. \quad (\text{D.5})$$

Equations D.3 to D.5 cast HACDU as a collaborative partially observable Markov decision process (Russell & Norvig, 2021). The state $s_t = (\theta_t, r_t) \in \mathcal{S}$ joins the latent world state with the reliance state; $\mathcal{A}$ is the action set, Ω the joint observation space, $\mathcal{T}$ the transition kernel of θ_t (written in calligraphic type to avoid a clash with the horizon T), O the observation model, R the expected net reward implied by Equation 6, and $\gamma \in (0,1]$ the discount factor. Reliance evolves by Equations 14 and 15 and is treated as observed through behavioral proxies, so the belief b_t is maintained over θ only (Equation D.4). The finite-horizon value function V_t^* trades off immediate net reward against future collaborative performance (Equation D.5).

D.3 Reinforcement Learning Updates

Equations D.6 to D.8 connect HACDU to standard reinforcement learning principles (Sutton & Barto, 1998):

$$Q(s, a) \leftarrow Q(s, a) + \eta[r + \gamma \max_{a'} Q(s', a') - Q(s, a)], \quad (\text{D.6})$$

$$\nabla_{\theta} J(\theta) = E_{\pi}[\nabla_{\theta} \ln \pi_{\theta}(a|s) A(s, a)], \quad (\text{D.7})$$

$$\max_{\theta} J(\theta) \text{ s.t. } E_{\pi}[g(s, a)] \leq \delta_{safe}. \quad (\text{D.8})$$

Equations D.6 to D.8 connect HACDU to reinforcement learning: standard TD learning and policy-gradient updates can be used to optimize collaborative policies subject to safety constraints on expected risk signals.

D.4 Robust Optimization

$$\pi^* = \arg \max_{\pi} \min_{P \in U} E_P[J_C(\pi)], \quad (\text{D.9})$$

$$J_C^{shift} = J_C - \lambda_{shift} \ell(P, P_0). \quad (\text{D.10})$$

Equations D.9 and D.10 define a robust optimization objective that seeks policies performing well under worst-case distributional shifts within the uncertainty set U , where $\ell(P, P_0) \geq 0$ measures the divergence of a shifted distribution P from the nominal distribution P_0 and $\lambda_{shift} \geq 0$ sets the penalty on harmful shifts.

D.5 Fairness Regularization

$$\Delta_u = \max_{g \in G} J_g - \min_{g \in G} J_g, \quad (\text{D.11})$$

$$\Delta_r = \max_{g \in G} r_g - \min_{g \in G} r_g, \quad (\text{D.12})$$

$$J_C^{fair} = J_C - \lambda_u \Delta_u - \lambda_r \Delta_r. \quad (\text{D.13})$$

Equations D.11 to D.13 define distributional fairness metrics and a fairness-regularized objective: Δ_u and Δ_r measure the largest utility and reliance gaps across groups $g \in G$, and the regularized objective penalizes them with weights $\lambda_u, \lambda_r \geq 0$. The objective addresses distributional, not procedural, fairness (Section 13.2).