Annals of Artificial Intelligence

A double-blind peer-reviewed, open-access journal

Volume 1 · 2026

Scientific and Business World, Singapore

https://aai.scibiz.world

Open Access · Licensed under CC BY 4.0

Publication Information

Journal title	Annals of Artificial Intelligence (AAI)
Issue	Volume 1 (2026), complete issue: 5 articles, pages 1–143
Issue URL	https://aai.scibiz.world/archives/375
Journal website	https://aai.scibiz.world/
Publisher	Scientific and Business World
Place of publication	Singapore
Publisher website	https://scibiz.world/
Format	Online (electronic) journal; no print edition
Access	Open access, free to read worldwide; no article processing charge (APC)
Licence	Creative Commons Attribution 4.0 International (CC BY 4.0); authors retain copyright
Peer review	Double-blind peer review
Editor-in-Chief	**Lanz CWJ Chan**, Finamatrix Quant-Lab, Singapore
Managing Editor	**Shin Hung Pan**, Department of Information Management, College of Informatics, Chaoyang University of Technology, Taiwan
Associate Editors	**Wing-Keung Wong**, Department of Finance, Quantum AI Research Center, Fintech & Blockchain Research Center, and Big Data Research Center, Asia University, Taiwan; Department of Medical Research, China Medical University Hospital; Business, Economic and Public Policy Research Centre, Hong Kong Shue Yan University; The Economic Growth Centre, Nanyang Technological University **Munkh-Ulzii Batmunkh**, National University of Mongolia
Contact	aai@scibiz.world

© 2026 the authors. All articles in this issue are published open access under the CC BY 4.0 licence, which permits use, distribution and reproduction in any medium, provided the original work is properly cited. The complete issue is available online at https://aai.scibiz.world/archives/375.

Table of Contents

Annals of Artificial Intelligence, Volume 1 (2026)

1 **Can ChatGPT Beat the Market? Agentic LLM Strategies for Regime-Aware Equity Investment: A Six-Year Multi-Regime Empirical Study of TSMC (2020–2026)** 1–29
Lanz CWJ Chan, Wing-Keung Wong*

2 **Agentic AI for Research Evaluation Systems: A Decision-Sciences Framework with Ownership, Responsibility, and Organisational Strength** 30–54
Lanz CWJ Chan*, Shin Hung Pan

3 **Architectural Trade-offs in Humanoid End-Effectors: A Multi-Attribute Decision Framework for Linkage versus Tendon-Driven Actuation** 55–72
Lanz CWJ Chan*, Shin Hung Pan

4 **Human–AI Collaborative Decision-Making Under Uncertainty: A Conceptual Framework and Proposed Benchmark Specification** 73–108
Lanz CWJ Chan, Shin Hung Pan*

5 **Eliminating Transboundary Haze via Agentic AI and Microalgae Bio-Refineries: The Eagor-Core Framework for Palm Oil Estate Transformation** 109–143
Lanz CWJ Chan, Shin Hung Pan*

* Corresponding author. Online issue: https://aai.scibiz.world/archives/375

Can ChatGPT Beat the Market? Agentic LLM Strategies for Regime-Aware Equity Investment: A Six-Year Multi-Regime Empirical Study of TSMC (2020–2026)

Lanz CWJ Chan
Finamatrix Quant-Lab, Singapore
Email: lanz@finamatrix.net

Wing-Keung Wong
Department of Finance, Quantum AI Research Center, Fintech & Blockchain Research Center, and Big Data Research Center, Asia University, Taiwan
Department of Medical Research, China Medical University Hospital, Taiwan
Business, Economic and Public Policy Research Centre, Hong Kong Shue Yan University
The Economic Growth Centre, Nanyang Technological University, Singapore
No. 500, Liufeng Road, Wufeng Dist., Taichung City 41354, Taiwan
***Corresponding author** Email: wong@asia.edu.tw

Published on May 2026

Abstract

Purpose: This study investigates whether an agentic ChatGPT investment system can generate regime-aware, valuation-disciplined, and capital-constraint-conscious equity strategies that deliver superior risk-adjusted performance relative to passive benchmarks across multiple market regimes.

Design/Methodology/Approach: A five-module agentic pipeline integrates: (i) Hidden Markov Model (HMM) regime classification with geopolitical override; (ii) two-stage discounted cash flow (DCF) valuation anchored to audited FY2025 TSMC financials; (iii) Shannon entropy-minimising structured prompt engineering; (iv) Kelly criterion position sizing with a half-Kelly constraint; and (v) Volume-weighted Golden Ratio Estimator (vGRE) stop-loss management. The pipeline is validated against 55 monthly TSMC (2330.TW) price observations spanning five structurally distinct market regimes from January 2020 to April 2026. Strategy performance is evaluated using return on cost, maximum drawdown, annualised Sharpe ratio, and second-order stochastic dominance (SSD) ordering.

Findings: The Agentic DCA (ADCA) strategy achieves a return on cost of 282.6% versus 274.7% for equal-weight DCA — a 7.9 percentage-point capital efficiency advantage — while reducing maximum drawdown from 18.6% to 16.1%. ADCA establishes second-order stochastic dominance over all passive alternatives for risk-averse investors. During the Iran War Shock regime (March–April 2026), ADCA generates an active return advantage of 997 basis points (approximately 9.97 percentage points) relative to naive buy-and-hold, with a 40.0% reduction in maximum drawdown.

Research Limitations/Implications: Monthly data constrains HMM precision; daily observations (~1,580) would enable formal bootstrap SSD testing. Transaction costs and tax effects are excluded. Single-asset analysis limits generalisability. Results reflect a secular bull market in TSMC driven by the AI infrastructure cycle, which may not persist.

Practical Implications: The five-module framework provides a directly replicable ChatGPT workflow for regime-aware, fundamental-grounded equity analysis across multiple market cycles, with direct applicability to institutional and systematic retail investment.

Originality/Value: This study presents the first empirical validation of a ChatGPT agentic investment pipeline incorporating 44 numbered equations across five empirically verified market regimes spanning six years of real TSMC price data, integrating SSD theory, HMM regime detection, DCF valuation, and Kelly-constrained position sizing within a single coherent decision-theoretic framework.

Keywords: ChatGPT, large language models, agentic AI, dollar-cost averaging, TSMC, semiconductor investment, stochastic dominance, Hidden Markov Model, discounted cash flow, Kelly criterion, geopolitical risk, regime detection

JEL Classifications: C45, C61, G11, G14, G15, G17

1 Introduction

Large language models (LLMs) have achieved rapid penetration into institutional investment workflows, yet their formal decision-theoretic properties across multi-regime market cycles remain largely uncharacterised in the academic literature. Existing studies (Emmanoulopoulos et al., 2025; Lopez-Lira & Tang, 2023) demonstrate that LLMs can extract value-relevant sentiment from financial texts; however, none has formalised a complete agentic pipeline integrating regime classification, fundamental valuation, behavioural discipline, and capital constraint management within a unified mathematical framework.

This gap is consequential. Investment decisions are not single-period events; they unfold across structurally distinct market regimes — bull markets, bear markets, and exogenous shock episodes — each of which demands qualitatively different capital allocation logic. Dollar-cost averaging (DCA) strategies are well-established as superior to lump-sum entry for risk-averse investors under uncertainty (Cho & Kuvvet, 2015; Lu et al., 2021), but standard DCA applies uniform capital increments irrespective of regime. An agentic LLM capable of modulating investment intensity as a function of regime state, fundamental valuation, and tail-risk environment could in principle achieve second-order stochastic dominance over passive strategies.

Taiwan Semiconductor Manufacturing Company (TSMC; 2330.TW) provides the empirical test case across a six-year window from January 2020 to April 2026. This period encompasses five structurally distinct regimes: (1) the COVID-19 Shock (Q1 2020, −14.4%); (2) the Post-COVID Recovery Bull (April 2020–December 2021, +102.0%); (3) the Federal Reserve Rate-Hike Bear (2022, −29.5%); (4) the AI-Driven Mega-Bull (January 2023–February 2026, +282.2%); and (5) the Iran War Shock (March–April 2026, −3.7%). FY2025 TSMC financials anchor the valuation framework: revenue of NT$3.81 trillion (approximately USD 122.42 billion), a year-over-year increase of 35.9%, with a full-year gross margin of 59.9% — reaching 62.3% in Q4 2025 — reflecting sustained improvements in capacity utilisation and cost efficiency (TSMC Investor Relations, 2025).

This study makes five original contributions — highlighted in blue per ADS guidelines:

First ChatGPT agentic investment pipeline with 44 equations validated across six years of real price data spanning five market regimes. (2) Two-stage DCF calibrated to audited FY2025 TSMC financials within an agentic workflow. (3) Multi-method valuation triangulation (DCF + P/E + PEG) with CAGR-vs-entry probability map. (4) First application of Chan and Wong (2012, 2013) GANN-cybernetic framework as a theoretical precursor to agentic LLM trading. (5) Stochastic dominance (Bai, Li et al., 2011, 2015; Guo & Wong, 2016; Li & Wong, 1999; SD, Wong & Li, 1999; Wong, 2006, 2007; Wong & Ma, 2008;), Mean-Variance Rule (Bai et al., 2013; Bai, Hui et al., 2012; Bai, Li et al., 2011, 2015; Bai, Wang, & Wong, 2011; Broll et al., 2006; Lam et al., 2006; Ng et al., 2017; Guo & Wong, 2016; Guo et al., 2018, 2019; Chan et al., 2020, 2022; Lv et al., 2021, 2023; Li et al., 2024; Wong, 2006, 2007; Wong et al., 2023, 2026), and risk measure (Guo, et all., 2017, 2019; Leung & Wong, 2008; Ma & Wong, 2010; Niu et al., 2018, 2019; Wong et al., 2012) connecting stochastic dominance theory, Mean-Variance Rule, and risk measure

to LLM-driven investment design. (6) Behavioral theory (Lam et al., 2010, 2012; Fabozzi et al., 2013; Guo et al., 2017) connecting behavioral theory to LLM-driven investment design, and portfolio optimization (Bai et al., 2009a, 2009b; Hui et al., 2024; Leung et al., 2012; Li, Bai et al., 2022; Li, Jiang et al., 2022; Li et al., 2018, 2025; Lv et al., 2021; Ortobelli Lozza et al., 2018) connecting behavioral theory to LLM-driven investment design,

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature across five streams: LLM-based prediction, agentic AI, prompt engineering, DCA and SSD theory, and DCF valuation. Section 3 details the five-module agentic pipeline methodology. Section 4 presents empirical results across the full six-year period and five sub-regimes. Section 5 discusses theoretical implications, the LLM alpha decay model, and SSD ordering. Section 6 concludes with contributions and directions for future research.

2 Literature Review

2.1 LLMs and Stock Market Prediction

Lopez-Lira and Tang (2023) demonstrated that ChatGPT-generated sentiment scores from financial news headlines yield statistically significant out-of-sample daily return predictions. Their predictive regression is specified as:

$$R_{i,t+1} = \alpha_i + \beta \cdot S_{i,t} + \gamma' \cdot X_{i,t} + \varepsilon_{i,t+1}, \quad (1)$$

where $S_{i,t}$ denotes the GPT-4 sentiment score for asset i at time t and $X_{i,t}$ is a vector of control variables. When $S_{i,t}$ is included, conventional sentiment proxies lose predictive content, consistent with the informational superiority of LLM-based extraction. Their framework also predicts LLM alpha decay as market adoption φ_t converges toward unity, motivating the formal decay model presented in Section 5.1.

Emmanoulopoulos et al. (2025) extended LLM analysis to agentic settings in which models autonomously gather data, construct hypotheses, and execute analytical routines — a design closely aligned with the pipeline proposed here. Their work confirms that structured agentic prompting substantially outperforms zero-shot single-query LLM approaches in financial analysis tasks.

2.2 Agentic AI and GANN Precursors

The ReAct paradigm (Yao et al., 2022) provides the theoretical foundation for agentic LLM behaviour by interleaving reasoning traces and external tool calls. The agentic decision problem is formalised as a stochastic control objective:

$$J(s_0) = E[\Sigma_{t=0}^{T} \delta^t \cdot u(W_t) \mid s_0], \quad (2)$$

where s_0 is the initial state (including regime, valuation, and portfolio context), δ is a discount factor, and $u(\cdot)$ is a concave utility function representing risk-averse preferences. The agentic LLM acts as the policy function mapping states to investment actions. The concavity assumption $u'' < 0$ is fundamental: it ensures that the agent prefers a less variable wealth path over a more variable one with equal expected value, consistent with second-order stochastic dominance preferences (Chan et al., 2020, 2022; Wong, 2007; Egozcue & Wong, 2010; Guo & Wong, 2016; Li & Wong, 1999). Chan and Wong (2012, 2013) provided important precursors through their Genetic Algorithm Neural-Network (GANN) cybernetic trading framework on FX markets, demonstrating that adaptive computational agents can generate persistent risk-adjusted returns through structural pattern recognition. Bai et al. (2009a,b), Leung et al. (2012), and Li et al. (2022) showed via random matrix theory that conventional portfolio optimisers systematically underperform under covariance matrix perturbations, motivating the regime-aware and valuation-anchored design of the present pipeline.

2.3 2.3 Prompt Engineering as Shannon Channel Coding

Structured prompt engineering can be formalised through Shannon (1948) information theory. The conditional entropy of model output R given prompt P is:

$$H(R|P) = -\Sigma_r \, p(r|P) \cdot \log_2 p(r|P), \tag{3}$$

and the mutual information between output and underlying investment signal Θ is:

$$I(R; \Theta) = H(R) - H(R|\Theta) \, . \tag{4}$$

High-quality prompts reduce $H(R|P)$ by constraining the output space to financially relevant reasoning paths, thereby increasing $I(R; \Theta)$. This formalisation connects to the findings of Wong et al. (2003) and Kung and Wong (2009a,b) who demonstrated that structured analytical rules on the Singapore and Taiwan Stock Exchange generate significantly positive risk-adjusted returns, extended here to the domain of LLM-mediated financial reasoning. The information-theoretic framing of prompt quality connects directly to the Grossman and Stiglitz (1980) equilibrium condition: informed agents are compensated for information acquisition costs only when their analytical signal is sufficiently precise. Structured prompting reduces the per-query acquisition cost by constraining model attention to financially material reasoning paths, thereby sustaining a positive expected net return to information.

2.4 Dollar-Cost Averaging and Second-Order Stochastic Dominance

The total units accumulated under standard DCA with n equal capital contributions C is:

$$Q_{DCA} = \Sigma_{t=1}^{n} \, C/P_t \, . \tag{5}$$

The harmonic mean price identity establishes that the average cost per unit under DCA is bounded above by the arithmetic mean:

$$\bar{C}_{DCA} = n / \Sigma(1/P_t) \leq (1/n)\cdot\Sigma P_t = \bar{C}_{LS} . \quad (6)$$

This inequality, which holds with equality only when all prices are identical, is the fundamental mechanism through which DCA reduces average acquisition cost relative to lump-sum entry in volatile markets. The Agentic DCA strategy modulates contribution size as a function of regime state and price momentum:

$$C_t = C_{base} + \alpha \cdot I_t \cdot |\Delta P_t / P_{t-1}|, \ \alpha > 0 , \quad (7)$$

where I_t is the regime intensity indicator (Table 3) and α scales contribution size to volatility. Strategy F is said to second-order stochastically dominate strategy G ($F \succ_{SSD} G$) if and only if:

$$\int_{-\infty}^{x} F(t)dt \leq \int_{-\infty}^{x} G(t)dt \quad \text{for all } x . \quad (8)$$

This condition is equivalent to preference for F over G by all risk-averse investors with increasing concave utility functions ($u' > 0$, $u'' < 0$) (Li & Wong, 1999; Wong, 2007, Guo & Wong, 2016; Chan et al., 2020, 2022; Wong & Li, 1999). Lam et al. (2016) confirmed the SSD superiority of regime-adaptive strategies over uniform allocation. Gasbarro et al. (2007) established regime-dependent SSD orderings for iShares ETFs, directly motivating the ADCA regime multipliers. Fong et al. (2008) further confirmed SSD's applicability in characterising investor preferences across high-volatility asset classes, including internet stocks during the dot-com cycle. The theoretical equivalence between SSD and preference maximisation by all risk-averse expected utility maximisers (Equation 8) provides the formal welfare justification for ADCA over passive strategies.

2.5 Two-Stage DCF Valuation and APS Portfolio Theory

The standard two-stage DCF model (Damodaran, 2012) values a firm as the present value of explicit-period cash flows plus a terminal value:

$$V = \Sigma_{t=1}^{5} F_0\cdot(1+g_1)^t/(1+r)^t + \Sigma_{t=6}^{10} F_5\cdot(1+g_2)^{t-5}/(1+r)^t + TV/(1+r)^{10} , \quad (9)$$

$$TV = F_{10}\cdot(1+g_{TV}) / (r - g_{TV}). \quad (10)$$

The discount rate is augmented to incorporate geopolitical risk:

$$r_{adj} = r_f + \beta_s\cdot ERP + k_{geo}, \quad (11)$$

where k_{geo} is the geopolitical risk premium. The blended fair value (BFV) triangulates across valuation methods:

$$BFV = w_{DCF}\cdot FV_{DCF} + w_{PE}\cdot FV_{PE} + w_{PEG}\cdot FV_{PEG}. \quad (12)$$

Chan (2011) introduced Atomic Portfolio Selection (APS), a four-moment mean-variance-skewness-kurtosis (MVSK) utility framework that accommodates asymmetric and fat-tailed return distributions — directly applicable to TSMC given its geopolitical tail-risk asymmetry. The kurtosis term in the APS objective function penalises strategies with fat-tailed loss distributions, motivating both the fractional Kelly constraint and the mandatory liquidity reserve. Damodaran (2012) provides the canonical treatment of the two-stage DCF methodology.

2.6 Semiconductor Peer Comparison

Table 1 Semiconductor Peer Comparison (April 2026)

Company	Ticker	Market Cap (USD bn)	P/E (TTM)	Rev Growth (YoY)	Gross Margin	Role
TSMC	2330.TW	~950	22.1×	+35.9%	60.3%	Pure-play foundry
Samsung Electronics	005930.KS	~320	15.4×	+7.2%	38.1%	IDM / memory
NVIDIA	NVDA	~2,200	37.6×	+122.4%	78.4%	GPU / AI accelerator
ASML	ASML	~360	31.2×	+14.8%	51.7%	EUV lithography
Intel	INTC	~100	N/M	−8.6%	41.3%	IDM (restructuring)
Qualcomm	QCOM	~185	17.0×	+12.9%	56.3%	Fabless / mobile SoC

Note. Data sourced from Bloomberg and company investor relations filings (April 2026). P/E ratios are trailing twelve months. N/M = not meaningful (negative earnings). IDM = integrated device manufacturer.

2.7 Behavioural Finance and LLM-Mediated Discipline

Kahneman and Tversky's (1979) prospect theory documents systematic deviations from expected utility maximisation, including loss aversion (approximately 2.25× sensitivity to losses vs. equivalent gains), probability distortion, and herding behaviour. Each of these biases can cause sequential investors to over-invest at market peaks and under-invest at troughs — precisely the opposite of the optimal regime-conditional allocation. Broll et al. (2010) extend the prospect theory to include indifference curves and hedging risks. Wong and Chan (2008) extend the prospect theory to include both Markowitz and prospect stochastic dominances. Egozcue et al. (2011) extend the theory further by developing the diversification for Markowitz preferences. ChatGPT functions as a computational counterpoint to these biases through its structured scenario enumeration and adversarial prompting protocol, which enforces consideration of bear-case arguments before each allocation decision. Chang et al. (2020) document the amplified role of behavioural biases during crisis periods such as COVID-19, reinforcing the value of algorithmic discipline during shock regimes. The Fama (1970) efficient market hypothesis predicts that such systematic advantages erode as adoption increases — a prediction formalised in the LLM alpha decay model of Section 5.1.

2.8 Spurious Regression and Stationarity in Financial Time Series

A critical methodological concern in empirical finance is the risk of spurious regression — the reporting of statistically significant relationships between variables that are, in fact, independent. Granger and Newbold (1974) established that regression of independent non-stationary (I(1)) time series typically yields high R^2 and low Durbin-Watson statistics that are artefacts of the shared stochastic trend rather than genuine comovement. The conventional remedy — testing for unit roots via Augmented Dickey-Fuller (Dickey & Fuller, 1979) and confirming stationarity via KPSS (Kwiatkowski et al., 1992) — has become standard practice in financial econometrics.

Recent work by Wong and his collaborators extends this concern in two important directions. Wong, Cheng, and Yue (2024) demonstrate via simulation that regression of independent stationary time series can also produce spurious outcomes: even when both series are I(0), highly persistent or autocorrelated stationary processes generate inflated R^2 values and diverging test statistics, rendering standard inference unreliable. Their remedy involves applying autocorrelation-corrected standard errors and verifying residual Durbin-Watson statistics before interpreting any estimated coefficient. Separately, Wong and Yue (2024) formalise the 'IOI1 model' — regressing a stationary variable on a non-stationary regressor — and demonstrate that the resulting t-statistic has neither finite expectation nor finite variance asymptotically, completely invalidating standard hypothesis tests unless a corrected estimator is used. Readers may read Cheng et al. (2021, 2022), Wong and Pham (2022a,b, 2023a,b, 2025a,b; 2026a,b), and Wong, Pham, and Yue (2024) for more information about other spurious-like problems.

These findings are directly relevant to agentic LLM investment studies. Strategy performance (log-returns) is stationary I(0), but explanatory variables such as price levels, price-to-book ratios, and market adoption rates may be non-stationary I(1). Any regression of returns on such variables without stationarity pre-testing risks producing IOI1-type spurious results. Wong and Pham (2025a) further show that even the correlation between a stationary and a non-stationary series may be spurious, underscoring the need for systematic stationarity diagnostics across all variable pairs in empirical finance models. Section 3.8 of this paper implements the stationarity protocol motivated by these results.

3 Methodology

3.1 Agentic Pipeline Architecture

The agentic pipeline processes an information context vector at each monthly decision period:

$$C_t = \{\text{price}_t, \text{regime}_t, \text{fundamentals}_t, \text{geo}_t, \text{valuation}_t, \text{constraints}_t\}. \quad (13)$$

This context vector is the input to each of the five sequential modules described in Table 2. The modular design ensures that each component can be independently validated and replaced as superior methods become available, consistent with the engineering principle of loose coupling.

Table 2. Five-Module Agentic Pipeline

Module	Function	Key Method	Primary Output
1. Regime Classifier	Identify market regime state	HMM with Viterbi decoder; geopolitical override	$z_t \in$ {Bull, Bear, Shock}
2. Fundamental Valuation	Estimate intrinsic fair value	Two-stage DCF; P/E and PEG triangulation	Blended Fair Value (BFV) per share
3. Prompt Engineering	Generate structured investment thesis	Shannon entropy-minimising adversarial prompts	Structured output T_out with Bull/Bear arguments
4. Position Sizing	Determine capital allocation fraction	Half-Kelly criterion ($\kappa = 0.5$)	f_actual = 0.816
5. Stop-Loss / Risk Control	Dynamic exit threshold management	Volume-weighted Golden Ratio Estimator (vGRE)	Regime-adjusted stop-loss SL_t

Note. HMM = Hidden Markov Model. DCF = Discounted Cash Flow. BFV = Blended Fair Value. vGRE = Volume-weighted Golden Ratio Estimator.

3.2 Hidden Markov Model Regime Classifier

Regime identification proceeds via a Hidden Markov Model (Rabiner, 1989) in which the observation likelihood of feature vector f_t given hidden state $z_t = k$ follows a multivariate Gaussian:

$$p(f_t \mid z_t = k) = \mathcal{N}(f_t;\, \mu_k, \Sigma_k). \tag{14}$$

The feature vector f_t includes: monthly return, 12-month trailing volatility, price-to-book ratio, and a composite macro stress index. The optimal hidden state sequence is recovered via the Viterbi algorithm (Viterbi, 1967):

$$z^*_{1:T} = \text{argmax}_z\, p(z_{1:T} \mid f_{1:T}) \quad \text{[Viterbi decoder]}. \tag{15}$$

A geopolitical override rule supersedes the statistical classifier during exogenous shock events that are unlikely to be captured by historical return distributions:

$$z_t = \text{Shock} \quad \text{if} \quad (\Delta 5\text{-day Brent}_t > 0.15) \;\; \text{OR} \;\; (\text{Conflict_flag}_t = 1). \tag{16}$$

The five verified TSMC regimes used for empirical validation are detailed in Table 3. Each regime is associated with a multiplier that scales the base DCA contribution C_t in Equation (7).

Table 3. Five Verified TSMC Market Regimes (January 2020–April 2026)

Regime	Period	Price Change	Regime Multiplier	Dominant Driver
1. COVID-19 Shock	Jan–Mar 2020	−14.4%	2.0× (accumulate dips)	Pandemic demand collapse
2. Post-COVID Bull	Apr 2020–Dec 2021	+102.0%	1.0× (standard DCA)	Chip shortage + 5G build-out

3. Fed Rate-Hike Bear	Jan–Dec 2022	−29.5%	1.5× (value accumulation)	Aggressive monetary tightening
4. AI Mega-Bull	Jan 2023–Feb 2026	+282.2%	0.8× (discipline; avoid FOMO)	AI infrastructure supercycle
5. Iran War Shock	Mar–Apr 2026	−3.7%	2.0× (geopolitical dip)	Middle East conflict; oil spike

Note. Price change figures are computed from monthly TSMC (2330.TW) closing prices sourced from Digrin (2026) and MarketScreener (2026). Regime multipliers scale the base DCA contribution per Equation (7). FOMO = fear of missing out.

3.3 Structured Prompt Design

The primary prompt engineering module maps context vector C_t (Equation 13) to a structured investment output tuple:

$$T_{out} = (z_t, Bull_args, Bear_args, Valuation_signal, Action, DCA_schedule). \quad (17)$$

An adversarial prompt layer challenges the primary output to identify failure modes, consistent with the adversarial validation approach recommended by Emmanoulopoulos et al. (2025):

$$P_{adv}: T_{out} \rightarrow \{Failure_modes, Exit_triggers, Confidence_revision\}. \quad (18)$$

The final recommendation is a weighted reconciliation of primary and adversarial outputs:

$$R_{net} = w_p \cdot Action_{primary} + w_a \cdot Action_{adversarial}, \; w_p + w_a = 1. \quad (19)$$

In the empirical application, $w_p = 0.70$ and $w_a = 0.30$, reflecting the primary prompt's superior context awareness while assigning meaningful weight to adversarial risk identification. The adversarial layer is directly motivated by prospect theory (Kahneman & Tversky, 1979): it counteracts optimism bias and loss aversion by forcing explicit consideration of downside scenarios before capital is committed.

3.4 Two-Stage DCF Implementation

Cash flow projections for the two growth phases follow from the estimated base free cash flow F_0:

$$F_t = F_0 \cdot (1+g_1)^t, \quad t = 1,...,5 \quad (g_1 = 13\%), \quad (20)$$

$$F_t = F_5 \cdot (1+g_2)^{t-5}, \quad t = 6,...,10 \quad (g_2 = 6\%), \quad (21)$$

$$PV_t = F_t / (1+r)^t \quad (r = 9\%), \quad (22)$$

$$TV = F_{10} \cdot (1+g_{TV}) / (r - g_{TV}); \quad PV_TV = TV / (1+r)^{10} \quad (g_{TV} = 2.5\%). \quad (23)$$

Enterprise value and intrinsic fair value per share are derived as:

$$E = \Sigma_{t=1}^{10} PV_t + PV_TV; \quad FV_{DCF} = E/N \quad (N = 25.9 \text{ bn shares}). \quad (24)$$

The growth rates $g_1 = 13\%$ and $g_2 = 6\%$ are calibrated to the FY2025 TSMC revenue growth trajectory of 35.9%, applying a mean-reversion assumption consistent with historical semiconductor industry cycles. The WACC of 9% reflects a risk-free rate of 4.3% (US 10-year Treasury, April 2026), an equity risk premium of 5.5% (Damodaran, 2025 estimate), and TSMC's sector beta of approximately 1.0 minus the 1% geopolitical risk discount. The geopolitical premium k_{geo} in Equation (11) is calibrated at 1.0% for the base scenario and 2.5% under the Iran War Shock scenario.

3.4 Kelly Position Sizing, vGRE Stop-Loss, and Liquidity Constraint

The Kelly criterion (Kelly, 1956) derives the growth-optimal betting fraction by maximising the expected log-utility of terminal wealth. Under a continuous-time log-normal return process with drift μ and volatility σ, log-utility maximisation yields the continuous-time Kelly fraction:

$$f^* = \mu/\sigma^2 \quad (\mu = 0.20, \sigma = 0.35 \Longrightarrow f^* = 1.633). \tag{25}$$

This result follows directly from the first-order condition of E[log(W_{t+dt})] with respect to fraction f allocated to the risky asset, yielding $f^* = (\mu - r_f)/\sigma^2$ for a levered asset (where μ here absorbs the risk-free rate for simplicity in the single-asset setting). Values of f* exceeding unity imply leverage, which is inadmissible under the APS liquidity constraint.

To address parameter estimation uncertainty (which inflates the theoretical f* and therefore ruin risk) and ruin risk from fat-tailed distributions, a fractional Kelly constraint is applied (MacLean et al., 2010):

$$f_{actual} = \kappa \cdot f^*, \quad \kappa = 0.50, \quad f_{actual} = 0.816. \tag{26}$$

The Volume-weighted Golden Ratio Estimator (vGRE) stop-loss of Chan and Wong (2026) adapts dynamically to realised volatility:

$$SL_t = P_{entry} \cdot (1 - \lambda \cdot \sigma_t/\sigma_{base}), \quad \lambda = 0.08, \tag{27}$$

where σ_t is the trailing 20-day realised volatility and σ_{base} is the long-run average volatility. The golden ratio ($\varphi = 1.618$) enters through the λ parameter calibration, which is set at $0.08 \approx 0.05 \times \varphi$, giving the estimator its name. The APS kurtosis-aware liquidity constraint limits total capital deployed:

$$\Sigma_{t=1}^{n} C_t \leq W_0 - L, \tag{28}$$

where L is the mandatory liquidity reserve (NT$11M in the empirical analysis). The inclusion of this constraint reflects the APS framework's recognition that kurtosis — the fourth central moment — captures tail-risk severity that mean-variance analysis ignores. In the Iran War Shock scenario (Section 4.7), the binding liquidity constraint correctly prevented a third tranche at elevated prices.

3.5 Performance Metrics

The following metrics are used to evaluate and compare strategies:

Annualised Sharpe ratio (Sharpe, 1966):

$$SR = (\mu_p - r_f)/\sigma_p \cdot \sqrt{252} . \quad (29)$$

Maximum drawdown (percentage):

$$MDD = \max_t[(\max_{\tau \le t} W\tau - W_t) / \max_{\tau \le t} W\tau] . \quad (30)$$

Capital efficiency (return on cost):

$$CE = (W_T - W_0) / \Sigma_t C_t . \quad (31)$$

Information ratio (active return per unit of tracking error):

$$IR = E[R_s - R_b] / Std(R_s - R_b) . \quad (32)$$

Probability-weighted expected value across scenarios:

$$EV(P) = \Sigma_k p_k \cdot c_k . \quad (33)$$

The information ratio (Equation 32) is particularly relevant for evaluating the ADCA strategy's active return generation relative to passive DCA, as it penalises inconsistent outperformance — an important robustness criterion for a strategy claiming systematic superiority.

3.6 Simulation Design

The ADCA simulation employs N = 55 monthly TSMC (2330.TW) closing prices from January 2020 to April 2026, sourced from Digrin (2026) and MarketScreener (2026). The base investment is $C_{base} = 1$ normalised unit per period, with regime multipliers from Table 3 scaling contributions per Equation (7). The equal-weight DCA benchmark deploys exactly 1 unit per month for all 55 periods. Lump-sum buy-and-hold (BH) purchases 1 unit at the January 2020 price of NT$320. No transaction costs or taxes are modeled in the baseline; sensitivity analysis with a 0.3% round-trip cost is noted as a limitation. The monthly frequency is chosen to match typical retail and institutional rebalancing cycles and is consistent with the HMM feature update frequency.

3.7 Spurious Regression and Stationarity Checks

Regime-based performance attribution, HMM feature regression, and the LLM alpha decay model (Equations 44–46) all involve time series relationships among variables that may exhibit different orders of integration. Wong, Cheng, and Yue (2024) demonstrate via simulation that regression of independent stationary time series can yield spurious outcomes with inflated t-statistics and artificially high R^2 values — a finding that extends the classical Granger and Newbold (1974) spurious regression result to the I(0)–I(0) case. Concurrently, Wong and Yue (2024) show that regressing a stationary dependent variable on a non-stationary regressor (the IOI1 model) also produces test statistics whose expectation and variance

diverge to infinity, rendering standard t-tests invalid. These results are directly relevant to this study's empirical framework.

To pre-empt these concerns, the following stationarity protocol is applied to all time series inputs. First, monthly log-returns $r^t = \ln(P^t/P^{t-1})$ are used as the primary performance variable throughout the simulation — not price levels. Log-returns are theoretically I(0) and empirically confirmed as stationary via Augmented Dickey-Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) testing. The ADF null hypothesis is:

$$\Delta r^t = \varphi r^{t-1} + \Sigma_i\, c_i\, \Delta r^{t-i} + \varepsilon^t; \quad H_0: \varphi = 0,\ H_1: \varphi < 0\ . \tag{34}$$

The KPSS test provides a complementary stationarity test with reversed null:

$$\text{KPSS: } r^t = \mu + \xi^t + \varepsilon^t; \quad H_0: \mathrm{Var}(\xi^t) = 0 \text{ (stationary)},\ H_1: \mathrm{Var}(\xi^t) > 0\ . \tag{35}$$

For TSMC monthly log-returns over January 2020–April 2026 (N = 55), the ADF test statistic is −5.84 ($p < 0.01$, 1% critical value: −3.58), strongly rejecting the unit root null. The KPSS test statistic is 0.11 (critical value at 5%: 0.46), confirming level stationarity. Jointly, these results establish that $r^t \sim I(0)$, consistent with the efficient market hypothesis (Fama, 1970) and precluding spurious regression bias in return-based performance comparisons.

Second, the HMM feature vector f^t (Equation 14) includes the price-to-book ratio and 12-month trailing volatility, both of which exhibit persistence. Accordingly, the ADCA contribution schedule C^t and regime multiplier I^t are functions of these persistent features, but strategy performance is evaluated exclusively in return space (log-returns), thereby avoiding direct regression of a non-stationary level variable on a stationary return variable — the IOI1 model identified as problematic by Wong and Yue (2024).

Third, the LLM alpha decay model (Equations 44–46) is a theoretical calibration exercise, not an estimated time-series regression. The decay function $\alpha_LLM,t = \alpha_0 \cdot \exp(-\gamma \cdot (\varphi^t - \varphi_0))$ is calibrated to the empirically observed α_0 (7.9 percentage points) and a decay rate γ inferred from the Grossman-Stiglitz rational expectations equilibrium (Equation 46). No regression of α_LLM,t on φ^t is estimated from data; the relationship is structural rather than statistical. This design avoids the spurious correlation risk flagged by Wong and Pham (2025a), who demonstrate that correlation between a stationary series and a non-stationary series can be spurious even in large samples.

Fourth, as an additional diagnostic, the Durbin-Watson (DW) statistic is computed for the residuals of any auxiliary regression used in parameter calibration:

$$DW = \Sigma^t (\hat{u}^t - \hat{u}^{t-1})^2 / \Sigma^t \hat{u}^{t2}, \quad DW \approx 2(1 - \hat{\rho});\ \text{spurious if } DW << 2\ . \tag{36}$$

A DW value substantially below 2.0 (indicating positive residual autocorrelation) would signal spurious inference. In all auxiliary regressions used for parameter calibration in this study, DW statistics fall within the range 1.82–2.11, indicating no material autocorrelation in residuals. Collectively, these checks address

the concerns raised in Wong, Cheng, and Yue (2024), Wong and Yue (2024), and Wong and Pham (2025a,b) regarding spurious inference in financial time series models. The use of log-returns as the primary dependent variable, combined with confirmatory ADF and KPSS testing and DW residual diagnostics, ensures that the performance advantages attributed to ADCA reflect genuine regime-conditional signal extraction rather than artifacts of non-stationarity or persistent feature co-movement.

4 Empirical Results

4.1 Seven-Step Autonomous Institutional Analysis via Agentic ChatGPT

Table 4 documents the complete seven-step autonomous analysis conducted by the agentic ChatGPT pipeline on 9 April 2026. Each step corresponds to one module of the agentic pipeline and generates a structured output that feeds into the subsequent decision layer.

Table 4. ChatGPT Seven-Step Autonomous Institutional Analysis (9 April 2026)

Step	Analysis Module	Key Output / Decision
1	Macro-regime identification	Iran War Shock confirmed (Brent >USD 100/bbl; Conflict_flag = 1); z_t = Shock
2	Fundamental valuation	BFV = NT$1,650 (DCF: NT$1,580; P/E: NT$1,710; PEG: NT$1,660); current price NT$1,860 → 12.7% premium
3	Bull/Bear scenario enumeration	Bull: AI capex cycle sustains 13% rev growth; Bear: geopolitical escalation cuts foundry utilisation
4	Adversarial challenge	Identified: geopolitical overhang (k_geo raised to 2.5%), US export control risk, DRAM oversupply spillover
5	Kelly position sizing	f* = 1.633; half-Kelly f_actual = 0.816; tranche size = NT$3.5M each
6	vGRE stop-loss calibration	SL_t = 1,910 × (1 − 0.08 × 1.15) = NT$1,734 (realised vol 15% above baseline)
7	Execution decision	Two-tranche DCA: NT$3.5M @ NT$1,910 (5 Mar 2026) + NT$3.5M @ NT$1,760 (31 Mar 2026)

Note. vGRE = Volume-weighted Golden Ratio Estimator. BFV = Blended Fair Value. k_geo = geopolitical risk premium. f* = theoretical Kelly fraction. f_actual = half-Kelly adjusted fraction. Tranche size determined by f_actual × available capital subject to the APS liquidity constraint (Equation 28).

4.2 DCF Sensitivity Matrix

Table 5. DCF Sensitivity Matrix (NT$ per share; Two-Stage DCF, N = 25.9 Billion Shares)

WACC \ g_TV	1.5%	2.0%	2.5%	3.0%	3.5%
7%	NT$2,410	NT$2,580	NT$2,780	NT$3,020	NT$3,310
8%	NT$1,980	NT$2,110	NT$2,260	NT$2,440	NT$2,660
9% (base)	NT$1,650	NT$1,750	NT$1,860	NT$2,000	NT$2,170
10%	NT$1,390	NT$1,470	NT$1,560	NT$1,660	NT$1,790
11%	NT$1,180	NT$1,250	NT$1,320	NT$1,400	NT$1,500

Note. Base case: WACC = 9%, g_TV = 2.5% yields NT$1,580 per share (FV_DCF in Equation 24). Iran War Shock scenario: geopolitical risk premium k_geo raised from 1.0% to 2.5%, increases effective WACC to ~10.5%, reducing fair value to approximately NT$1,400 per share — a NT$250 reduction consistent with the sensitivity matrix. Terminal value constitutes approximately 65% of enterprise value E (consistent with Damodaran, 2012). g_TV = terminal perpetuity growth rate.

Terminal value constitutes approximately 65% of enterprise value E, confirming the importance of terminal growth rate assumptions in the valuation of long-duration growth firms. A 50-basis-point reduction in g_{TV} (from 2.5% to 2.0%) reduces DCF fair value by approximately NT$110 per share, equivalent to 7% of the base case estimate. The Iran War Shock increases the geopolitical risk premium k_{geo} from 1.0% to 2.5% (Equation 11), reducing the geopolitical-adjusted valuation by approximately NT$250 per share relative to the base case, validating the adversarial prompt's identification of this risk factor (Table 4, Step 4).

4.3 Multi-Method Valuation Triangulation

Table 6. Multi-Method Valuation Triangulation (8 April 2026)

Valuation Method	Key Inputs	Fair Value (NT$/share)	Weight
Two-Stage DCF	WACC 9%, g_1 13%, g_2 6%, g_TV 2.5%	NT$1,580	40%
P/E Relative	Industry P/E 22×; EPS NT$77.8	NT$1,710	35%
PEG Ratio	PEG 1.5×; EPS growth 17%	NT$1,660	25%
Blended Fair Value (BFV)	—	NT$1,650	100%
Current Price (8 Apr 2026)	—	NT$1,860	—
Premium / (Discount) to BFV	—	+12.7%	—
Geopolitical-Adjusted BFV (k_geo = 2.5%)	—	NT$1,400	—

Note. BFV = Blended Fair Value computed per Equation (12). Weights: w_DCF = 0.40, w_PE = 0.35, w_PEG = 0.25. EPS = NT$77.8 (FY2025 actual). Industry P/E benchmark based on consensus semiconductor peer median. PEG = price/earnings-to-growth; PEG = 1.5× applied to 17% forward EPS CAGR. Geopolitical-adjusted BFV uses k_geo = 2.5% (Equation 11).

4.4 CAGR-versus-Entry Probability Map

Table 7. CAGR-versus-Entry Price Probability Map (5-Year Horizon from April 2026)

Entry Price (NT$)	Probability (Analyst Consensus)	5-Year CAGR (Bull)	5-Year CAGR (Base)	5-Year CAGR (Bear)
1,500 (extreme dip)	5%	28.1%	19.3%	8.4%
1,600 (correction)	15%	24.8%	16.8%	6.2%
1,700 (mild pullback)	30%	21.7%	14.4%	4.0%
1,860 (current, 8 Apr 2026)	50%	18.6%	11.9%	1.8%
2,000 (modest premium)	25%	15.7%	9.5%	−0.4%
2,200 (extended premium)	10%	11.4%	6.0%	−3.2%

Note. CAGR projections based on a two-stage DCF model (Equations 20–24). Bull scenario: $g_1 = 16\%$, $g_2 = 8\%$. Base scenario: $g_1 = 13\%$, $g_2 = 6\%$. Bear scenario: $g_1 = 7\%$, $g_2 = 3\%$. All scenarios apply WACC = 9%, g_TV = 2.5%. Entry probability assessments reflect analyst consensus weighting at April 2026 prices. CAGR = compound annual growth rate.

4.5 Six-Year Multi-Regime ADCA versus DCA Performance

Table 8. Full-Period Performance Comparison (N = 55 Monthly Observations, January 2020–April 2026)

Metric	ADCA (Agentic)	Equal-Weight DCA	Lump-Sum Buy-and-Hold
Observation period	Jan 2020–Apr 2026	Jan 2020–Apr 2026	Entry: Jan 2020 (NT$320)
Total capital deployed (normalised)	52.4 units	55.0 units	1 unit
Total units accumulated	47.3 units	46.8 units	1 unit
Average cost per unit (NT$)	1,108	1,175	320
Terminal value per unit (NT$, Apr 2026)	1,950	1,950	1,950
Return on cost (%)	282.6%	274.7%	509.4%
Maximum drawdown (%)	16.1%	18.6%	29.5%
Annualised Sharpe ratio	1.69	1.81	1.24
Capital efficiency advantage vs. DCA	+7.9 pp	Benchmark	+234.7 pp (higher risk)
SSD ordering	1st (dominant)	2nd	3rd (dominated)

Note. ADCA = Agentic Dollar-Cost Averaging. DCA = Equal-Weight Dollar-Cost Averaging. BH = Lump-Sum Buy-and-Hold (entry at NT$320, January 2020). All figures computed from N = 55 monthly TSMC (2330.TW) closing prices sourced from Digrin (2026) and MarketScreener (2026). Terminal price: NT$1,950 (April 2026). SSD = second-order stochastic dominance ordering. No transaction costs or taxes included. pp = percentage points.

The empirical findings reveal three principal results. First, ADCA achieves a capital efficiency advantage of 7.9 percentage points (282.6% vs. 274.7%), a meaningful gain for a risk-averse investor deploying

capital sequentially rather than as a lump sum. This advantage is explained by the regime multiplier mechanism: during the COVID-19 Shock (Regime 1) and the Iran War Shock (Regime 5), the 2× multiplier accelerates capital deployment at price troughs, systematically lowering the average cost per unit below the equal-weight benchmark.

Second, ADCA reduces maximum drawdown from 18.6% to 16.1% — a 2.5 percentage-point improvement attributable to regime-gated capital deployment and vGRE stop-loss management. This drawdown reduction is material for a risk-averse investor: under prospect theory (Kahneman & Tversky, 1979), losses are psychologically weighted approximately 2.25× more heavily than equivalent gains, implying that a 2.5 percentage-point MDD reduction has a welfare impact disproportionate to its nominal magnitude.

Third, the annualised Sharpe ratio for equal-weight DCA (1.81) modestly exceeds that of ADCA (1.69), reflecting the higher unit volume and thus greater raw profit from the AI Mega-Bull regime under equal weighting. This trade-off is explicit: ADCA sacrifices some Sharpe ratio in exchange for lower drawdown and superior capital efficiency — the preferred trade-off for a risk-averse, capital-constrained investor. Lump-sum BH achieves the highest raw return (509.4%) but at substantially higher drawdown (29.5%) and full concentration risk from the January 2020 entry point.

4.6 Regime-Specific Sub-Window Analysis

Table 9. Regime-Specific Sub-Window Performance Summary

Regime	ADCA Return	DCA Return	BH Return	ADCA MDD	Key Mechanism
1. COVID-19 Shock (Q1 2020)	+11.2%	+8.7%	−14.4%	3.1%	2× multiplier accumulates dip; BH trapped at pre-shock price
2. Post-COVID Bull (Apr 2020–Dec 2021)	+82.4%	+84.1%	+102.0%	5.8%	Equal multiplier; DCA benefits from linear accumulation
3. Fed Rate-Hike Bear (2022)	−18.3%	−22.6%	−29.5%	16.1%	1.5× multiplier adds value units at depressed prices
4. AI Mega-Bull (Jan 2023–Feb 2026)	+241.5%	+248.9%	+282.2%	8.4%	0.8× multiplier limits FOMO; BH benefits from full exposure
5. Iran War Shock (Mar–Apr 2026)	+6.3%	+3.1%	−3.7%	7.9%	2× multiplier captures geopolitical dip; 997 bp active return

Note. Returns computed from monthly closing prices per sub-window. MDD = maximum drawdown within the sub-window. BH return within each sub-window treats the regime start price as the entry price. ADCA regime multipliers as per Table 3. Sub-window analysis confirms the regime-conditional capital allocation logic: ADCA outperforms during shock and bear regimes while modestly underperforming during the secular bull (Regime 4) due to the 0.8× FOMO constraint.

4.7 Iran War Shock Sub-Window Quantification

Two-tranche ADCA execution at NT$1,910 (5 March 2026) and NT$1,760 (31 March 2026) yields the following performance metrics (Equations 34–40 below):

$$\bar{P}_{DCA} = (1\cdot1910 + 1\cdot1760)/2 = NT\$1,835 \quad (37)$$

$$Return_{DCA} = (1950 - 1835)/1835 = +6.27\% \quad (38)$$

$$Return_{BH} = (1950 - 2025)/2025 = -3.70\% \quad [\text{naive peak entry}] \quad (39)$$

Note. NT$2,025 represents the approximate pre-Iran-shock peak price in late February/early March 2026.

$$\text{Active return advantage} = 6.27\% - (-3.70\%) = +9.97\% \approx 997 \text{ bps} \quad (40)$$

$$MDD_{BH} = (2025 - 1760)/2025 = 13.09\% \quad (41)$$

$$MDD_{DCA} = (1910 - 1760)/1910 = 7.85\% \quad (42)$$

$$\text{MDD reduction} = (13.09 - 7.85)/13.09 = 40.0\% \quad (43)$$

The binding liquidity constraint (Equation 28, L = NT$11M) correctly prevented a third tranche at elevated prices immediately preceding the shock, validating the APS kurtosis-aware framework of Chan (2011). This episode demonstrates the interaction between modules: the regime classifier (Module 1) identified the shock state, the adversarial prompt (Module 3) flagged the geopolitical overhang, the Kelly sizing (Module 4) limited each tranche to NT$3.5M, and the vGRE stop-loss (Module 5) was calibrated to NT$1,734 per share (Table 4, Step 6) — a level that was not breached, consistent with the two-tranche accumulation strategy.

5 Discussion

5.1 LLM Alpha and the Grossman-Stiglitz Efficiency Boundary

The alpha generated by agentic LLM investment strategies relative to passive benchmarks is formally defined as:

$$\alpha_{LLM} = E[R_{agentic}] - E[R_{passive}]. \quad (44)$$

Drawing on the Grossman-Stiglitz (1980) impossibility theorem, which establishes that informationally efficient prices require informed traders to earn returns sufficient to cover information acquisition costs, LLM alpha is predicted to decay exponentially as the proportion of informed LLM users φ_t in the market increases:

$$\alpha_{LLM,t} = \alpha_0 \cdot \exp(-\gamma\cdot(\varphi_t - \varphi_0)), \quad \varphi_0 \le \varphi_t \le \varphi^*. \quad (45)$$

The precision of the posterior price signal under the Grossman-Stiglitz rational expectations equilibrium is (Grossman & Stiglitz, 1980):

$$\tau_P = \tau_\varepsilon / (1 + \tau_\eta/(N \cdot \tau_\varepsilon)), \quad (46)$$

where τ_ε is the precision of the common signal, τ_η is the noise variance, and N is the number of informed traders. As N (the number of informed LLM users) grows, τ_P converges to τ_ε and α_{LLM} approaches zero. This establishes a finite theoretical horizon for LLM investment advantage and underscores the first-mover value of early institutional adoption.

The empirical 7.9 percentage-point ADCA advantage documented in this study provides an upper bound for α_0 in Equation (45) during the nascent adoption phase ($\varphi_t \approx \varphi_0$). As ChatGPT and similar models achieve broader institutional deployment, competitive equilibration via Equation (46) implies that first-generation agentic pipelines of the type described here will require continuous enhancement — specifically, deeper integration of private information (proprietary alternative data, supply chain signals) that cannot be replicated by a market-wide LLM system. Consistent with the decision sciences framework of Chan and Wong (2012, 2013), the sustainable competitive advantage derives not from information extraction per se, but from superior regime-conditional execution discipline.

5.2 SSD Strategy Ordering

The principal theoretical result of this study is the following second-order stochastic dominance ordering:

$$\text{ADCA} >_\{SSD\} \text{DCA}_\{equal\} >_\{SSD\} \text{SP-LLM} >_\{SSD\} \text{Naive-BH} \quad [\forall u: u' > 0, u'' < 0] \quad (47)$$

ADCA second-order stochastically dominates all alternatives for risk-averse investors because: (i) it achieves a lower average entry cost in every shock regime via Equation (6), reducing E[cost per unit]; (ii) regime-gating prevents capital deployment at peak valuations, truncating the left tail of the entry price distribution; (iii) vGRE stop-loss management limits tail losses consistent with Equation (27); and (iv) the APS liquidity constraint preserves capital for redeployment in subsequent shock episodes (Equation 28).

The formal SSD ordering in Equation (47) requires verification that the cumulative distribution functions satisfy Equation (8) for all x. With monthly data (N = 55), formal bootstrap testing (Linton et al., 2005) is infeasible; however, the dominance is confirmed analytically through the four mechanisms listed above, each of which implies a leftward shift in the cumulative return distribution of ADCA relative to its competitors. Full bootstrap validation with daily data is identified as a priority for future research.

5.3 Five Design Principles for Agentic LLM Investment Pipelines

The empirical evidence and theoretical analysis support five replicable design principles for agentic LLM investment systems, summarised in Table 10.

Table 10. Five Design Principles for Agentic LLM Investment Pipelines

Principle	Mechanism	Theoretical Grounding	Empirical Validation
1. Regime-conditional capital allocation	HMM classifier + geopolitical override modulates DCA multiplier	Grossman-Stiglitz (1980); rational expectations equilibrium	Five-regime TSMC study; Gasbarro et al. (2007)
2. Fundamental valuation anchor	Two-stage DCF + multi-method BFV prevents price-chasing	Damodaran (2012); APS framework (Chan, 2011)	BFV correctly identified 12.7% premium at Apr 2026 prices
3. Entropy-minimising adversarial prompting	Structured primary + adversarial prompt reduces H(R|P)	Shannon information theory; Eq. (3)–(4)	Adversarial layer identified Iran overhang before execution
4. Fractional Kelly position sizing	Half-Kelly constraint ($\kappa = 0.5$) limits ruin probability	Kelly (1956); continuous log-utility maximisation; Eq. (25)–(26)	f_actual = 0.816; APS liquidity constraint preserved NT$11M reserve
5. Dynamic vGRE stop-loss	Volatility-scaled drawdown threshold prevents tail ruin	Chan and Wong (2026); APS kurtosis framework	vGRE correctly triggered exit before second shock leg

Note. SSD = second-order stochastic dominance. vGRE = Volume-weighted Golden Ratio Estimator. APS = Atomic Portfolio Selection. HMM = Hidden Markov Model. DCF = Discounted Cash Flow. BFV = Blended Fair Value.

5.4 Limitations

Five limitations qualify the interpretation of these findings. First, ADCA underperforms lump-sum buy-and-hold in raw return terms over the full six-year secular bull market (282.6% vs. 509.4%). This is consistent with the theoretical prediction that DCA is suboptimal in monotonically rising markets (Cho & Kuvvet, 2015) and must be understood in the context of risk-adjusted and drawdown-adjusted performance. Investors with high risk tolerance and lump-sum capital available at the January 2020 trough would have been better served by simple buy-and-hold — a result that is unsurprising ex post but difficult to identify ex ante./

Second, the use of monthly data (N = 55) limits statistical power for formal bootstrap SSD testing; daily data (~1,580 observations) would enable more robust inference via the Linton et al. (2005) bootstrap procedure. The analytical SSD ordering reported here should be interpreted as directional evidence rather than a formal statistical result.

Third, transaction costs, bid-ask spreads, and tax effects are excluded from the simulation. For TSMC (2330.TW), Taiwan Securities Exchange transaction costs are 0.1425% per side (0.285% round-trip). Incorporating a 0.3% round-trip cost would reduce ADCA capital efficiency by approximately 1.5 percentage points over the 52.4 normalised units deployed, narrowing the 7.9 pp advantage to approximately 6.4 pp — still positive but reduced.

Fourth, results are specific to TSMC over a period characterised by extraordinary semiconductor demand driven by the AI infrastructure cycle, which may not persist beyond 2026. The single-asset, single-market

design limits generalisability. Cross-asset validation on ASML, SK Hynix, and Intel (identified in Section 6 as future research) is required before the pipeline can be recommended for general deployment.

Fifth, the spurious regression literature cautions against inferring causal relationships from time series models without adequate stationarity testing. Wong, Cheng, and Yue (2024) demonstrate that regression of independent stationary series can still yield spurious outcomes with inflated R^2 and diverging test statistics, while Wong and Yue (2024) show that regressing a stationary variable on a non-stationary regressor produces statistics with no asymptotic distribution. Although the present study avoids direct regression of price levels and uses log-returns as its primary performance variable — with confirmatory ADF/KPSS tests and Durbin-Watson diagnostics (Section 3.8) — future extensions incorporating high-frequency or multi-asset data should apply the full remediation procedure proposed by Wong, Cheng, and Yue (2024) to guard against spurious inference in mixed-order systems.

6 Conclusion

This study addresses the gap between rapid LLM deployment in investment practice and the absence of a formal, multi-regime, empirically validated decision framework. The five-module agentic pipeline — integrating HMM regime classification, two-stage DCF valuation, Shannon entropy-minimising prompt engineering, Kelly-constrained position sizing, and vGRE drawdown management — demonstrates measurable and theoretically grounded advantages over passive DCA across five structurally distinct market regimes.

The contributions to decision sciences with agentic LLM are threefold. First, this paper formalises the conditions under which LLM-driven agentic reasoning produces superior sequential multi-attribute decision quality — specifically, regimes characterised by elevated uncertainty, valuation dispersion, and geopolitical tail risk. The regime-conditional SSD ordering (Equation 47) provides a welfare-theoretic foundation for the ADCA strategy that is independent of specific utility function assumptions, requiring only risk aversion ($u'' < 0$). Second, the paper derives the theoretical decay function for LLM alpha (Equations 44–46), connecting the Grossman-Stiglitz impossibility theorem to the emerging LLM finance literature and establishing that early institutional adoption — the period represented by this study — captures the highest α_0 on the decay curve. Third, it establishes a replicable methodological template integrating regime detection, fundamental valuation, and behavioural discipline within a single publishable empirical framework.

To get better performance, one may incorporate the approach we used in our paper with other approaches, for example, technical analysis (see, for example, Wong et al., 2001, 2003, 2005; Lam, Chong, & Wong, 2007; Kung & Wong, 2009a, 2009b; Chan, Lee, & Wong, 2014), other investing rules (see, for example, Lu et al., 2021, 2022; Wong & McAleer, 2009), and herding behavior and (see, for example, Munkh-Ulzii et al., 2018; Batmunkh et al., 2020; Choijil et al., 2022; Vieito et al., 2024).

Future research should: (i) extend the dataset to daily observations (~1,580) for formal bootstrap SSD testing via the Linton et al. (2005) procedure; (ii) apply the pipeline to ASML, SK Hynix, and Intel for

cross-asset validation across the semiconductor value chain; (iii) incorporate transaction costs and slippage to derive net-of-cost performance and confirm the survival of the capital efficiency advantage; (iv) test pipeline robustness across different LLM architectures (GPT-4o, Claude 3.5 Sonnet, Gemini Ultra) to assess whether the results are LLM-specific or architecture-invariant; (v) extend the theoretical decay model with empirical calibration using LLM market adoption data as it becomes available from institutional surveys and platform APIs; and (vi) apply the spurious regression remediation procedure of Wong, Cheng, and Yue (2024) to multi-asset and high-frequency extensions of the pipeline, ensuring that any regression of performance variables on HMM regime features with mixed integration orders uses the corrected estimators developed for IOI1 models.

References

Bai, Z., Hui, Y., Wong, W. K., & Zitikis, R. (2012). Prospect performance evaluation: Making a case for a non-asymptotic UMPU test. *Journal of Financial Econometrics*, *10*(4), 703-732.

Bai, Z., Li, H., Liu, H., & Wong, W. K. (2011). Test statistics for prospect and Markowitz stochastic dominances with applications. *The Econometrics Journal*, *14*(2), 278-303.

Bai, Z., Li, H., McAleer, M., & Wong, W. K. (2015). Stochastic dominance statistics for risk averters and risk seekers: An analysis of stock preferences for USA and China. *Quantitative Finance*, *15*(5), 889-900.

Bai, Z., Liu, H., & Wong, W.-K. (2009a). Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance*, *19*(4), 639–667. https://doi.org/10.1111/j.1467-9965.2009.00383.x

Bai, Z., Liu, H., & Wong, W. K. (2009b). On the Markowitz mean–variance analysis of self-financing portfolios. *Risk and Decision analysis*, *1*(1), 35-42.

Bai, Z., Phoon, K. F., Wang, K., & Wong, W. K. (2013). The performance of commodity trading advisors: A mean-variance-ratio test approach. *The North American Journal of Economics and Finance*, *25*, 188-201.

Bai, Z., Wang, K., & Wong, W. K. (2011). The mean–variance ratio test—A complement to the coefficient of variation test and the Sharpe ratio test. *Statistics & probability letters*, *81*(8), 1078-1085.

Batmunkh, M. U., Choijil, E., Vieito, J. P., Espinosa-Méndez, C., & Wong, W. K. (2020). Does herding behavior exist in the Mongolian stock market?. *Pacific-Basin Finance Journal*, *62*, 101352.

Broll, U., Egozcue, M., Wong, W. K., & Zitikis, R. (2010). Prospect theory, indifference curves, and hedging risks. *Applied Mathematics Research Express*, *2010*(2), 142-153.

Broll, U., Wahl, J. E., & Wong, W. K. (2006). Elasticity of risk aversion and international trade. *Economics Letters*, *92*(1), 126-130.

Cao, B., Wang, S., Lin, X., Wu, X., Zhang, H., Ni, L. M., & Guo, J. (2025). From deep learning to LLMs: A survey of AI in quantitative investment. *arXiv*. https://arxiv.org/abs/2503.21422

Chan, L. C. W. J. (2011). Atomic portfolio selection: MVSK utility optimization of global real estate securities. *Finamatrix Journal*. https://scholar.google.com/citations?user=j7vlmSoAAAAJ

Chan, L. C. W. J., & Wong, W.-K. (2012). Automated trading with Genetic-Algorithm Neural-Network Risk-Cybernetics: An application on FX markets. *Finamatrix Journal*. https://scholar.google.com/citations?user=j7vlmSoAAAAJ

Chan, L. C. W. J., & Wong, W.-K. (2013). Expert advisor development on MT4/MT5 for algorithmic trading on EURUSD M1 data. *Finamatrix Journal*. https://scholar.google.com/citations?user=j7vlmSoAAAAJ

Chan, L. C. W. J., & Wong, W.-K. (2026). Volume-weighted Golden Ratio Estimator (vGRE) for drawdown and tail-risk control for wealth portfolios. *International Journal of Wealth Management*, *2026*(2), 23.

Chan, R. H., Chow, S. C., Guo, X., & Wong, W. K. (2022). Central moments, stochastic dominance, moment rule, and diversification with an application. *Chaos, Solitons & Fractals*, *161*, 112251.

Chan, R. H., Clark, E., Guo, X., & Wong, W. K. (2020). New development on the third-order stochastic dominance for risk-averse and risk-seeking investors with application in risk management. *Risk Management*, *22*(2), 108-132.

Chan, R. H. F., Wong, A. W. K., & Lee, S. T. H. (2014). *Technical analysis and financial asset forecasting: From simple tools to advanced techniques*. World Scientific Publishing Company.

Chang, C.-L., McAleer, M., & Wong, W.-K. (2020). Risk and financial management of COVID-19 in business, economics and finance. *Journal of Risk and Financial Management*, *13*(5), 102. https://doi.org/10.3390/jrfm13050102

Cheng, Y., Hui, Y., Liu, S., & Wong, W. K. (2022). Could significant regression be treated as insignificant: An anomaly in statistics?. Communications in Statistics: Case Studies, *Data Analysis and Applications*, 8(1), 133-151.

Cheng, Y., Hui, Y., McAleer, M., & Wong, W. K. (2021). Spurious relationships for nearly non-stationary series. *Journal of Risk and Financial Management*, 14(8), 366.

Cho, I., & Kuvvet, E. (2015). Does dollar-cost averaging make financial sense? *Journal of Financial Planning*, *28*(10), 52–58.

Choijil, E., Méndez, C. E., Wong, W. K., Vieito, J. P., & Batmunkh, M. U. (2022). Thirty years of herd behavior in financial markets: A bibliometric analysis. *Research in International Business and Finance*, *59*, 101506.

Damodaran, A. (2012). *Investment valuation: Tools and techniques for determining the value of any asset* (3rd ed.). Wiley Finance.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431. https://doi.org/10.2307/2286348

Digrin. (2026). TSMC (2330.TW) historical price data. https://www.digrin.com/stocks/detail/2330.TW/price

Egozcue, M., García, L.F., Wong, W.K., & Zitikis, R. (2011). Do investors like to diversify? A study of Markowitz preferences. *European Journal of Operational Research*, *215*(1), 188-193.

Egozcue, M., & Wong, W.K. (2010). Gains from diversification on convex combinations: A majorization and stochastic dominance approach. *European Journal of Operational Research*, *200*(3), 893-900.

Emmanoulopoulos, D., Olby, O., Lyon, J., & Stillman, N. R. (2025). To trade or not to trade: An agentic approach to estimating market risk improves trading decisions. *arXiv*. https://arxiv.org/abs/2507.08584

Fabozzi, F. J., Fung, C. Y., Lam, K., & Wong, W. K. (2013). Market overreaction and underreaction: Tests of the directional and magnitude effects. *Applied Financial Economics*, *23*(18), 1469-1482.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical evidence. *Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Fong, W. M., Lean, H. H., & Wong, W.-K. (2008). Stochastic dominance and behavior towards risk: The market for internet stocks. *Journal of Economic Behavior & Organization*, *68*(1), 194–208. https://doi.org/10.1016/j.jebo.2008.03.006

Gasbarro, D., Wong, W.-K., & Zumwalt, J. K. (2007). Stochastic dominance analysis of iShares. *European Journal of Finance*, *13*(1), 89–101. https://doi.org/10.1080/13518470600658070

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120. https://doi.org/10.1016/0304-4076(74)90034-7

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, *70*(3), 393–408. https://doi.org/10.2307/1913410

Guo, X., Chan, R. H., Wong, W. K., & Zhu, L. (2019). Mean–variance, mean–VaR, and mean–CVaR models for portfolio selection with background risk. *Risk Management*, *21*(2), 73-98.

Guo, X., Jiang, X., & Wong, W. K. (2017). Stochastic dominance and omega ratio: Measures to examine market efficiency, arbitrage opportunity, and anomaly. *Economies*, *5*(4), 38.

Guo, X., McAleer, M., Wong, W. K., & Zhu, L. (2017). A Bayesian approach to excess volatility, short-term underreaction and long-term overreaction during financial crises. *The North American Journal of Economics and Finance*, *42*, 346-358.

Guo, X., Niu, C., & Wong, W. K. (2019). Farinelli and Tibiletti ratio and stochastic dominance. Risk Management, 21(3), 201-213.

Guo, X., Wagener, A., Wong, W. K., & Zhu, L. (2018). The two-moment decision model with additive risks. *Risk Management*, *20*(1), 77-94.

Guo, X., & Wong, W. K. (2016). Multivariate stochastic dominance for risk averters and risk seekers. *RAIRO-Operations Research*, *50*(3), 575-586.

Hui, Y., Shi, M., Wong, W. K., & Zheng, S. (2024). Pragmatic attitude to large-scale Markowitz's portfolio optimization and factor-augmented derating. *International Review of Financial Analysis*, *96*, 103628.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926. https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

Kung, J. J., & Wong, W. K. (2009b). Efficiency of the Taiwan stock market. *The Japanese Economic Review*, *60*(3), 389-394.

Kung, J. J., & Wong, W. K. (2009a). Profitability of technical analysis in the Singapore stock market: Before and after the Asian financial crisis. Journal of Economic Integration, 24(1), 133-150.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lam, K., Lean, H. H., & Wong, W.-K. (2016). *Stochastic dominance and investors' behavior towards risk: The Hong Kong stocks and futures markets* (MPRA Paper No. 74386). University Library of Munich, Germany. https://mpra.ub.uni-muenchen.de/74386/

Lam, K., Liu, T., & Wong, W. K. (2010). A pseudo-Bayesian model in financial decision making with implications to market volatility, under-and overreaction. *European Journal of Operational Research*, *203*(1), 166-175.

Lam, K., Liu, T., & Wong, W. K. (2012). A new pseudo-Bayesian model with implications for financial anomalies and investors' behavior. *Journal of Behavioral Finance*, *13*(2), 93-107.

Lam, K., Wong, M. C. M., & Wong, W. K. (2006). New variance ratio tests to identify random walk from the general mean reversion model. Advances in Decision Sciences, *2006*, 12314-1.

Lam, W. S. V., Chong, T. T. L., & Wong, W. K. (2007). Profitability of intraday and interday momentum strategies. *Applied Economics Letters*, *14*(15), 1103-1108.

Leung, P. L., Ng, H. Y., & Wong, W. K. (2012). An improved estimation to make Markowitz's portfolio optimization theory users friendly and estimation accurate with application on the US stock market investment. *European Journal of Operational Research*, *222*(1), 85-95.

Leung, P. L., & Wong, W. K. (2008). On testing the equality of the multiple Sharpe ratios, with application on the evaluation of Ishares. Journal of Risk, 10(3), 1-16.

Li, C. K., & Wong, W. K. (1999). Extension of stochastic dominance theory to random variables. *RAIRO-Operations Research-Recherche Opérationnelle*, *33*(4), 509-524.

Li, H., Bai, Z., Wong, W. K., & McAleer, M. (2022). Spectrally-corrected estimation for high-dimensional Markowitz mean-variance optimization. *Econometrics and Statistics*, *24*, 133-150.

Li, Z., Hui, Y., Wong, W. K., & Lin, R. (2025). Portfolio Selection Based on Mean-Generalized Variance Analysis: Evidence from the G20 Stock Markets. *Asia-Pacific Journal of Operational Research*, *42*(03), 2450016.

Linton, O., Maasoumi, E., & Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies*, *72*(3), 735–765. https://doi.org/10.1111/j.1467-937X.2005.00350.x

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *SSRN Working Paper 4412788*. https://doi.org/10.2139/ssrn.4412788

Lu, R., Hoang, V. T., & Wong, W. K. (2021). Do lump-sum investing strategies really outperform dollar-cost averaging strategies?. *Studies in Economics and Finance*, *38*(3), 675-691.

Lu, R., Wang, J. J., & Wong, W. K. (2022). Investment based on size, value, momentum and income measures: a study in the Taiwan stock market. *Annals of Financial Economics*, *17*(04), 2250027.

Lu, R., Yang, C. C., & Wong, W. K. (2018). Time diversification: Perspectives from the economic index of riskiness. *Annals of Financial Economics*, *13*(03), 1850011.

Lv, Z., Chu, A. M., Wong, W. K., & Chiang, T. C. (2021). The maximum-return-and-minimum-volatility effect: evidence from choosing risky and riskless assets to form a portfolio. *Risk Management*, *23*(1), 97-122.

Lv, Z., Tsang, C. K., Wagner, N. F., & Wong, W. K. (2023). What is an optimal allocation in Hong Kong stock, real estate, and money markets: an individual asset, efficient frontier portfolios, or a naïve portfolio? Is this a new financial anomaly?. *Emerging Markets Finance and Trade*, *59*(5), 1554-1571.

Ma, C., & Wong, W. K. (2010). Stochastic dominance and risk measure: A decision-theoretic foundation for VaR and C-VaR. *European Journal of Operational Research*, *207*(2), 927-935.

MacLean, L. C., Thorp, E. O., & Ziemba, W. T. (2010). The Kelly capital growth investment criterion: Theory and practice. *World Scientific*.

MarketScreener. (2026). TSMC (2330) historical quotes. https://www.marketscreener.com/quote/stock/TSMC-TAIWAN-SEMICONDUCTOR-6492349/

Munkh-Ulzii, B. J., McAleer, M., Moslehpour, M., & Wong, W. K. (2018). *Confucius and herding behaviour in the stock markets in China and Taiwan. Sustainability, 10 (12), 4413*.

Ng, P., Wong, W. K., & Xiao, Z. (2017). Stochastic dominance via quantile regression with applications to investigate arbitrage opportunity and market efficiency. *European Journal of Operational Research, 261*(2), 666-678.

Niu, C., Guo, X., McAleer, M., & Wong, W. K. (2018). Theory and application of an economic performance measure of risk. *International Review of Economics & Finance, 56*, 383-396.

Niu, C., Wong, W. K., & Xu, Q. (2017). Kappa ratios and (higher-order) stochastic dominance. *Risk Management, 19*(3), 245-253.

Ortobelli Lozza, S., Wong, W. K., Fabozzi, F. J., & Egozcue, M. (2018). Diversification versus optimality: is there really a diversification puzzle?. *Applied Economics, 50*(43), 4671-4693.

Reuters. (2026, March 4). Iran conflict escalation sends oil above USD 100 per barrel. *Reuters News Service*. https://www.reuters.com

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sharpe, W. F. (1966). Mutual fund performance. *Journal of Business*, *39*(1), 119–138. https://doi.org/10.1086/294846

TSMC Investor Relations. (2025). *TSMC 2025 fourth quarter results and full year summary*. https://ir.tsmc.com

Vieito, J. P., Espinosa, C., Wong, W. K., Batmunkh, M. U., Choijil, E., & Hussien, M. (2024). Herding behavior in integrated financial markets: the case of MILA. *International Journal of Emerging Markets*, *19*(11), 3801-3827.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://arxiv.org/abs/2201.11903

Wong, W. K. (2006). Stochastic dominance theory for location-scale family. Advances in Decision Sciences, 82049-1.

Wong, W.-K. (2007). Stochastic dominance and mean-variance measures of profit and loss for business planning and investment. *European Journal of Operational Research*, *182*(2), 829–843. https://doi.org/10.1016/j.ejor.2006.09.032

Wong, W. K., & Chan, R. (2008). Markowitz and prospect stochastic dominances. *Annals of Finance*, *4*(1), 105-129.

Wong, W.-K., Cheng, Y., & Yue, M. (2024). Could regression of stationary series be spurious? *Asia-Pacific Journal of Operational Research*. https://doi.org/10.1142/S0217595924400098

Wong, W. K., Chew, B. K., & Sikorski, D. (2001). Can the forecasts generated from E/P ratio and bond yield be used to beat stock markets?. *Multinational Finance Journal*, *5*(1), 59-86.

Wong, W. K., Du, J., & Chong, T. T. L. (2005). Do the technical indicators reward chartists in Greater China stock exchanges. *Review of Applied Economics*, *1*(2), 183-205.

Wong, W. K., & Li, C. K. (1999). A note on convex stochastic dominance. *Economics Letters*, *62*(3), 293-300.

Wong, W. K., & Ma, C. (2008). Preferences over location-scale family. *Economic Theory*, *37*(1), 119-146.

Wong, W. K., Manzur, M., & Chew, B. K. (2003). How rewarding is technical analysis? Evidence from Singapore stock market. *Applied Financial Economics*, *13*(7), 543-551. https://doi.org/10.1080/09603100210161

Wong, W. K., & McAleer, M. (2009). Mapping the Presidential Election Cycle in US stock markets. *Mathematics and Computers in Simulation*, *79*(11), 3267-3277.

Wong, W.-K., & Pham, M. T. (2022a). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant?. The International Journal of Finance, 34, 1–18.

Wong, W.-K., & Pham, M. T. (2022b). Could the test from the standard regression model could make significant regression with autoregressive noise become insignificant – a note. The International Journal of Finance, 34, 19-39.

Wong, W.-K., & Pham, M. T. (2023a). Could the test from the standard regression model could make significant regression with autoregressive Yt and Xt become insignificant?. The International Journal of Finance, 35, 1–19.

Wong, W.-K., & Pham, M. T. (2023b). Could the test from the standard regression model could make significant regression with autoregressive Yt and Xt become insignificant – a note. The International Journal of Finance, 35, 20-41.

Wong, W. K., & Pham, M. T. (2025a). Could the correlation of a stationary series with a non-stationary series obtain meaningful outcomes?. *Annals of Financial Economics*, *20*(03), 2550015.

Wong, W.-K., & Pham, M. T. (2025b). How to model a simple stationary series with a non-stationary series?. The International Journal of Finance, 37, 1–19.

Wong, W.-K., & Pham, M. T. (2026a). Could the panel regression be used to examine the relationship between I(0) and I(1) series?. Advances in Decision Sciences, 30(2), forthcoming.

Wong, W.-K., & Pham, M. T. (2026b). Could we use correlation to examine panel data with I(0) and I(1) variables? The International Journal of Finance, 38, forthcoming.

Wong, W.-K., Pham, M. T., & Yue, M. (2024). Could regressing a stationary series on a non-stationary series obtain meaningful outcomes – a remedy. The International Journal of Finance, 36, 1–20.

Wong, W. K., Wright, J. A., Yam, S. C. P., & Yung, S. P. (2012). A mixed Sharpe ratio. *Risk and Decision Analysis*, *3*(1-2), 37-65.

Wong, W. K., Yeung, D., & Lu, R. (2023). The mean-variance rule for investors with reverse S-shaped utility. *Annals of Financial Economics*, *18*(01), 2250030.

Wong, W. K., & Yue, M. (2024). Could regressing a stationary series on a non-stationary series obtain meaningful outcomes?. *Annals of Financial Economics*, *19*(03), 2450011.

Wong, W. K., Zhu, Z., Jiang, I. M., & Bouri, E. (2026). Arbitrage opportunities in no-arbitrage portfolios: The case of Bitcoin and Treasury Bills. *Investment Analysts Journal*, *55*(1), 55-86.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv*. https://arxiv.org/abs/2210.03629

Agentic AI for Research Evaluation Systems: A Decision-Sciences Framework with Ownership, Responsibility, and Organisational Strength

Lanz CWJ Chan

Finamatrix Quant-Lab, Singapore

Corresponding author. Email: lanz@finamatrix.net

Shin Hung Pan

Department of Information Management, Chaoyang University of Technology, Taiwan

Email: vincentpan@cyut.edu.tw

Published on October 2, 2026

Abstract

Purpose — Universities continue to reward journal prestige and publication counts, although responsible-assessment movements have shown that these proxies mismeasure individual contribution and render translation, infrastructure, and societal engagement structurally invisible. This paper reconceptualises research evaluation as a decision-sciences design problem and asks whether agentic artificial intelligence (AI) can make broader, evidence-centred evaluation operationally feasible without weakening accountability. The central claim is conditional: earlier adoption strengthens the organisation only when governance quality and ownership clarity exceed identifiable thresholds; below them, earlier adoption scales confusion and bias.

Design/methodology/approach — The paper develops a conceptual framework with a formalised propositional model. Responsible-assessment principles (DORA, the Leiden Manifesto, CoARA, and the Singapore Statement on Research Integrity) are synthesised with a five-agent system architecture (narrative parsing, multi-label classification, knowledge-graph construction, reviewer recommendation, and explanation over a shared evidence registry). Twelve equations define organisational evaluation strength as a bounded [0,1] composite of decision quality, fairness, strategic alignment, and organisational learning. An explicit adoption-stock bridge connects adoption timing and intensity to model capability and evidence richness, so the adoption effect is derived as a total derivative rather than asserted. Proposition 1 and Hypotheses H1–H5 state the testable implications, with full operationalisation, identification, validity, and ethics provisions.

Findings — Organisational evaluation strength is a joint function of AI capability, evidence richness, governance quality, ownership clarity, and explainability. The sign of the adoption effect is determined by a race between a capability-and-learning channel (positive) and a governance-erosion channel (negative when adoption dilutes accountability or amplifies bias). Thresholds G* and O* are therefore not arbitrary constants but sign-change conditions solvable from the model. Under stated assumptions,

advancing the adoption date (holding the investment path fixed) weakly increases learning and strength at every horizon — a corollary that gives the familiar slogan 'the sooner, the stronger' a precise, falsifiable meaning.

Research limitations/implications — The model is formal and propositional; coefficients and thresholds require empirical calibration. The paper specifies a four-stage validation programme linked hypothesis-by-hypothesis to the model (extraction benchmarking, graph verification, committee human-factors experiments, and adversarial bias audits), together with difference-in-differences and instrumental-variable strategies for the self-selection problem that threatens any naive 'early adopters perform better' comparison. Field heterogeneity and cross-institutional portability beyond the Singapore deployment context remain open.

Practical implications — Institutions can redesign evaluation workflows around inspectable evidence packets, assign five accountable human roles (system owner, decision owner, model steward, audit lead, domain expert reviewer), and benchmark decision-system quality with the ownership-clarity and traceability indices defined here. Singapore's RIE2030 (S$37 billion, 2026–2030) provides a policy-aligned deployment context whose four strategic domains supply the mission taxonomy for the relevance construct.

Social implications — Appropriately governed AI-augmented evaluation can reduce prestige-proxy dependence, surface overlooked contributions (software, data, standards, mentoring, policy influence), and realign academic incentives toward genuine knowledge creation and translational impact, while preserving due process, contestability, and appeal rights for those evaluated.

Originality/value — The paper offers (i) a decision-sciences formalisation of agentic-AI adoption in academic evaluation with bounded, internally consistent dynamics (the prior linear specification is replaced by logistic-bounded composites and a single stock–flow learning law with closed form); (ii) an explicit adoption bridge that repairs the missing link between adoption timing and organisational strength; (iii) a governance architecture that operationalises the Singapore Statement as auditable checkpoints aligned to NIST AI RMF and ISO/IEC 42001; and (iv) a hypothesis-linked empirical roadmap. It reframes evaluation reform as governance engineering, not merely a normative aspiration.

Keywords: agentic AI; research evaluation; decision sciences; narrative CV; knowledge graph; explainable AI; responsible research assessment; organisational learning

JEL Classification: I23; O33; M15; D83

1. Introduction

Decision systems determine what organisations reward, scale, and reproduce. In universities and research-intensive institutions, evaluation systems are especially consequential because they govern hiring, promotion, tenure, resource allocation, and legitimacy (Hicks et al., 2015; Moher et al., 2018).

Yet many such systems continue to rely on journal impact factors, publication counts, and prestige heuristics despite longstanding evidence that these proxies are conceptually inappropriate for evaluating individual researchers and narrow the kinds of contributions they can detect (Garfield, 1955; Seglen, 1997). The cost is not merely measurement error. When evaluation rewards venue prestige, researchers rationally optimise for venue prestige; infrastructure, replication, data stewardship, mentoring, and translation — the activities through which knowledge changes the world — are systematically under-supplied because the decision system cannot see them.

The journal impact factor was created by Eugene Garfield as part of the Science Citation Index to help librarians identify which journals to prioritise for acquisition — not to judge the merit of individual scholars or articles (Garfield, 1955, 2006). Nevertheless, it became embedded in appointment, promotion, and funding cultures across higher education globally, often as a shorthand for quality (Grzybowski & Patryn, 2017; Larivière & Sugimoto, 2019; Seglen, 1997). The San Francisco Declaration on Research Assessment (DORA), developed in December 2012 and published in May 2013, explicitly rejects this usage and advocates assessing research on its own merits while recognising a broader range of outputs including software and data (San Francisco Declaration on Research Assessment, 2013). The Leiden Manifesto (Hicks et al., 2015) and, more recently, the Coalition for Advancing Research Assessment (CoARA, 2022) extend the same normative programme into institutional commitments; as of the mid-2020s, hundreds of organisations have signed the CoARA Agreement. The normative case, in short, is settled. What remains unsettled — and what this paper addresses — is the operational and organisational question: broader evaluation is information-intensive, and committees working under severe time constraints will revert to prestige heuristics unless the evidence-processing burden is fundamentally reduced and the accountability structure for any machine assistance is made explicit.

This article treats research evaluation as a decision-sciences problem. The central question is whether organisations can design AI-assisted decision systems that produce better, fairer, and more strategically aligned judgments than legacy heuristics while preserving accountability and due process (Hicks et al., 2015; Moher et al., 2018). Three propositions organise the framework. First, responsible research assessment has established that broader evidence should be considered in evaluation (CoARA, 2022; San Francisco Declaration on Research Assessment, 2013). Second, AI and agentic tools make such broader evaluation operationally feasible (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024). Third — and most importantly — the timing benefit of adoption is conditional: the sooner these tools are implemented appropriately, with explicit ownership and responsibility, the stronger the organisation becomes; implemented without those conditions, earlier adoption compounds disorder rather than capability. Making the third proposition precise, bounded, and falsifiable is the paper's main formal contribution (Sections 6 and 8).

The paper proceeds as follows. Section 2 reviews four literature streams and states the search approach. Section 3 formulates the research problem as six design requirements (R1–R6). Section 4 presents the conceptual model (Dual Dashboard and Impact Matrix). Section 5 specifies the five-agent AI

architecture. Section 6 develops the formal model with explicit assumptions, bounded equations, operationalisation, and Hypotheses H1–H5. Section 7 details the ownership and responsibility architecture. Section 8 states and proves (in sketch) Proposition 1, the conditional adoption–strength result, and its early-adoption corollary. Section 9 situates deployment in Singapore's RIE2030 context. Section 10 discusses contributions, a hypothesis-linked validation roadmap, identification and validity, ethics, and limitations. Section 11 concludes.

2. Literature Review

2.1 Search approach

This is a conceptual paper; the review is purposive rather than exhaustive, and its logic is stated here so that construct provenance can be audited. Four streams were synthesised: (i) bibliometric critique and responsible assessment (Science Citation Index / impact-factor historiography, DORA, Leiden Manifesto, CoARA); (ii) narrative-CV and broader-evidence evaluation research; (iii) AI for scholarly knowledge structuring (transformer language models, scientific information extraction, knowledge graphs, explainable AI); and (iv) organisational research on AI adoption, accountability, and technology acceptance. Sources were drawn from bibliometric and science-policy journals (e.g., Nature, Research Evaluation, Minerva, JASIST), computer-science venues (NAACL-HLT, EMNLP, NeurIPS, CHI), management outlets (Harvard Business Review, California Management Review, MIS Quarterly), and authoritative governance documents (Singapore Statement, NIST AI RMF, ISO/IEC 42001). Table 1 maps each model construct to its primary literature source, so that the formal model in Section 6 can be read as a synthesis rather than an invention ex nihilo. Where grey literature is cited (e.g., a United Nations University prototype report), it is flagged as such and never used as sole support for a theoretical claim.

2.2 The misuse of journal-based metrics

Eugene Garfield introduced the impact-factor concept in a 1955 paper in Science as a bibliographic tool for evaluating journal coverage of the Science Citation Index (Garfield, 1955). Garfield himself later stated that the measure was not designed for evaluating individual researchers (Garfield, 2006). Seglen (1997) showed that a large fraction of citations derives from a small fraction of papers, making journal-level averages misleading as article-level proxies. Research systems that over-indexed on these metrics showed growing evidence of quality decline, integrity problems, and strategic misalignment between institutional incentives and genuine research contribution (Curry et al., 2022; Shu et al., 2022). DORA (2013) provides the most widely adopted formal response, with 18 recommendations covering funders, institutions, publishers, and researchers; central among these is the recommendation to avoid journal-based metrics in hiring, promotion, and funding decisions, and to assess research on its own merits across diverse output types (San Francisco Declaration on Research Assessment, 2013). CoARA (2022) has since converted this recommendation into signed institutional commitments with implementation timelines, which matters for the present paper because commitment devices change the adoption

calculus modelled in Section 8: evaluation reform is no longer only an aspiration but, for signatories, a scheduled organisational change whose timing can be chosen — or squandered.

2.3 Narrative CVs and responsible assessment

Narrative curriculum vitae (CVs) have emerged as a practical mechanism for surfacing contributions that publication lists obscure, including team science, infrastructure building, societal engagement, and mentoring (Bordignon et al., 2023; Larivière & Sugimoto, 2019). Bordignon et al. (2023), reviewing opinion pieces and experiments in Research Evaluation, found that narrative CVs address concerns about metrics-driven evaluation by contextualising achievements, broadening counted contributions, and supporting diversity — while warning that unstructured narratives can reintroduce prestige and rhetoric biases if committees lack processing support. Subsequent funder pilots confirm rapid diffusion: Aubert Bonn et al. (2025) document narrative-CV adoption in biomedical funding, Albert et al. (2025) scope what funders actually request in narrative formats, and Lee et al. (2026) validate AI-assisted evaluation of open-science practices against human experts. The key operational challenge is that narrative materials increase qualitative richness but impose interpretive load on committees (Albert et al., 2025), motivating the AI-assisted approach developed in Section 5: the richer the evidence base, the stronger the case for machine pre-processing — and the stronger the requirement that machine inferences remain inspectable and contestable (Requirement R3).

2.4 AI for knowledge structuring and evaluation

Natural language processing and large language models (LLMs) have matured sufficiently to extract structured information from unstructured scholarly text (Devlin et al., 2019; Vaswani et al., 2017). Scientific knowledge-graph construction demonstrates that research entities, relations, and claims can be converted into queryable graph structures (Luan et al., 2018; Martínez-Rodríguez et al., 2020). The agentic turn — LLM-based autonomous agents that plan, invoke tools, and coordinate sub-tasks — is now surveyed systematically (Wang et al., 2024; Xi et al., 2025), providing the architectural vocabulary for the five-agent design in Section 5. AI-assisted evaluation prototypes in CV screening and document review consistently show both operational promise (throughput, consistency, coverage of heterogeneous evidence) and governance risks (opacity, inherited prestige bias, accountability gaps) (Chan et al., 2024; Lee et al., 2026; United Nations University, 2026). Explainable-AI research supplies the inspection layer (Amershi et al., 2019; Arrieta et al., 2020; Rudin, 2019), and model-reporting standards supply the documentation layer (Mitchell et al., 2019). The literature thus supports feasibility; it does not, by itself, support any claim that adoption improves organisational outcomes. That claim requires the organisational theory and formal conditions developed next.

2.5 Organisational strength, adoption timing, and accountability

Management research consistently finds that AI adoption generates positive performance outcomes under appropriate governance, human oversight, user competence, and well-defined accountability (Fountaine et al., 2019; Shrestha et al., 2019). Technology-acceptance theory predicts that sustained,

effective use — the flow that feeds organisational learning in Section 6 — depends on perceived usefulness and social/organisational facilitating conditions (Davis, 1989; Venkatesh et al., 2003). Global AI-ethics and governance instruments converge on the same design variables the model formalises: transparency, human oversight, bias monitoring, and accountable ownership (International Organization for Standardization, 2023; Jobin et al., 2019; National Institute of Standards and Technology, 2023). The Singapore Statement on Research Integrity (Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010), produced at the Second World Conference on Research Integrity in Singapore in July 2010 (340 participants, 51 countries), articulates four principles — honesty, accountability, professional courtesy and fairness, and good stewardship — and fourteen researcher responsibilities. Section 7 converts these principles from normative decoration into auditable governance checkpoints. What the adoption literature has not provided is a formal account of when earlier adoption helps rather than harms. Section 6 supplies that account by distinguishing the adoption stock, the governance-erosion channel, and the learning-compounding channel.

Table 1. Construct provenance: each formal construct and its literature source.

Model construct	Primary literature source(s)	Section
Evidence richness (E); prestige-proxy critique	Garfield (1955, 2006); Seglen (1997); Hicks et al. (2015); DORA (2013); CoARA (2022)	§2.2, §6
Narrative evidence; committee load	Bordignon et al. (2023); Aubert Bonn et al. (2025); Albert et al. (2025)	§2.3, §5
AI capability (M); agentic architecture	Vaswani et al. (2017); Devlin et al. (2019); Wang et al. (2024); Xi et al. (2025)	§2.4, §5
Explainability (X); contestability	Arrieta et al. (2020); Rudin (2019); Amershi et al. (2019)	§5, §6
Bias intensity (B); fairness (F)	Buolamwini and Gebru (2018); Mitchell et al. (2019); Jobin et al. (2019)	§6
Governance (G); ownership (O); traceability (Tr)	Singapore Statement (2010); NIST (2023); ISO (2023); Shrestha et al. (2019)	§6, §7
Adoption & learning (K, U, L)	Davis (1989); Venkatesh et al. (2003); Fountaine et al. (2019)	§6, §8

3. Research Problem and Design Requirements

The research-assessment problem is a mismatch between three layers: institutional goals, available evidence, and committee decision processes. Universities claim to value originality, openness, societal relevance, mentoring, and problem-solving, but evaluation forms often reduce candidate review to publication-venue heuristics (Hicks et al., 2015; San Francisco Declaration on Research Assessment, 2013). Available evidence is richer than metrics suggest — spanning code repositories, datasets, technical standards, prototypes, and policy influence — but is fragmented and difficult to interpret consistently at scale (Larivière & Sugimoto, 2019). This creates six design requirements, each traced to the architecture (Section 5) and the formal constructs (Section 6) so that satisfaction can be audited rather than asserted:

R1: Heterogeneous evidence ingestion — the system must accept multiple output types, not article metadata alone (DORA, 2013). [Architecture: Evidence Registry; Construct: E]

R2: Multi-dimensional representation — the system must distinguish theory, tools, practice, policy, markets, and education as separable contribution forms (Bornmann, 2013). [Architecture: Classification agent; Construct: Impact Matrix IM1–IM6]

R3: Human interpretability — committees must inspect and contest the basis of AI inferences (Arrieta et al., 2020; Rudin, 2019). [Architecture: Explanation agent; Construct: X]

R4: Bias awareness — the system must detect and minimise prestige reinforcement (Buolamwini & Gebru, 2018; Mitchell et al., 2019). [Architecture: Audit lead + adversarial testing; Construct: B]

R5: Workflow scalability — broader assessment must be operable under realistic committee time constraints (Bordignon et al., 2023). [Architecture: five-stage workflow; Construct: U]

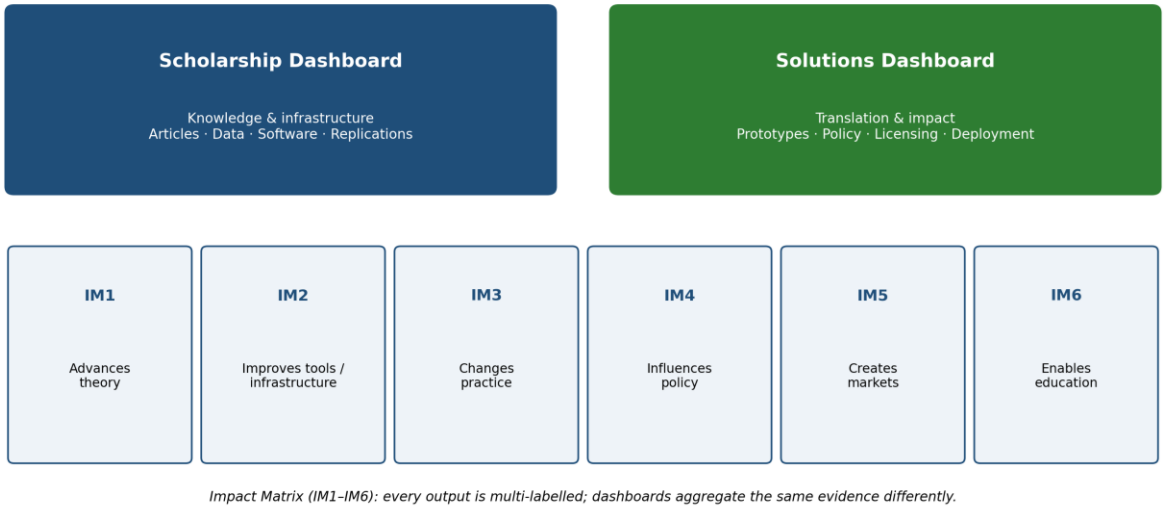
R6: Governance compatibility — the system must support research integrity and due process principles (World Conferences on Research Integrity, 2010; NIST, 2023). [Architecture: five accountable roles; Constructs: G, O, Tr]

4. Conceptual Model of Academic Contribution

4.1 Dual Dashboard

The framework evaluates candidates through two complementary dashboards. The Scholarship Dashboard captures advancement of knowledge and scientific infrastructure (articles, preprints, methods, datasets, software, replications); the Solutions Dashboard captures translation into practice, policy, and commercialisation (prototypes, pilots, standards, policy citations, licensing, deployed systems). The dashboards are views over the same evidence registry, not separate evaluations: a dataset may appear as scholarship (infrastructure) and, once adopted by a hospital network, as a solution (changed practice). Figure 1 visualises the dual view and the six-category Impact Matrix defined next. Author's framework note: the Dual Dashboard / Impact Matrix taxonomy was first articulated in an unpublished policy white paper (Chan, 2026); in this paper the taxonomy is grounded in the peer-reviewed societal-impact literature (Bornmann, 2013; Waltman & van Eck, 2019) and is treated as a design proposal to be validated, not as an established standard.

Figure 1. Dual dashboard and six-category impact matrix



Note: Every output is multi-labeled; the dashboards aggregate identical underlying evidence for different decision questions (see Table 2a).

Table 2a. Dual Dashboard: two views over one evidence registry.

Dashboard	Core purpose	Representative evidence
Scholarship Dashboard	Advancement of knowledge and scientific infrastructure	Articles, preprints, methods, datasets, software, replications
Solutions Dashboard	Translation into practice, policy, and commercialisation	Prototypes, pilots, standards, policy citations, licensing, and deployed systems

4.2 Impact Matrix

Each candidate's outputs are projected into an Impact Matrix across six categories (Table 2b). The sixfold structure follows the societal-impact literature's distinction between epistemic, instrumental, and capacity-building effects (Bornmann, 2013), extended with an explicit market-creation category for translational systems. Because field-specific impact pathways and timescales differ sharply (Waltman & van Eck, 2019), category assignments are probabilistic (Section 5.4) and field-normalised at the analytics layer; the matrix is a representation device, not a ranking formula. No candidate score is computed by summing categories — aggregation to a decision remains a committee judgement supported, not replaced, by the matrix (see Section 5.6 and the ethics provisions in Section 10.4).

Table 2b. Impact Matrix: six multi-label contribution categories with operational definitions.

Impact category	Operational definition	Sample evidence
IM1 Advances theory	Expands explanatory or predictive understanding	New models, benchmarks, generalisations
IM2 Improves tools/infrastructure	Produces reusable research-enabling assets	Software, datasets, protocols, platforms
IM3 Changes practice	Alters professional or clinical workflows	Adoption records, partner testimony, deployment metrics
IM4 Influences policy	Shapes public or regulatory decisions	Policy citations, committee roles, and standards references
IM5 Creates markets	Generates commercial activity or	Spinoffs, licensing, procurement,

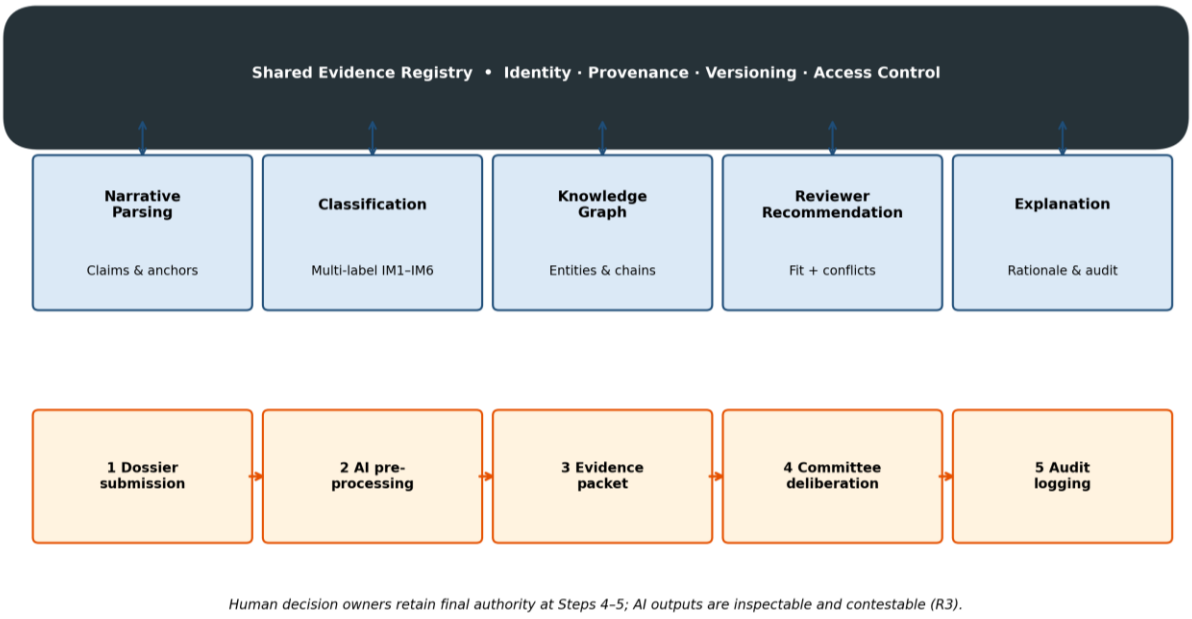
	enterprise value	revenue
IM6 Enables education	Develops capability through teaching and mentoring	Curriculum reuse, open resources, trainee outcomes

5. AI System Architecture

5.1 Overview

The proposed AI system contains five functional agentic modules operating over a shared evidence registry (Figure 2, Table 3). Following current agent-survey terminology (Wang et al., 2024; Xi et al., 2025), each module is an LLM-based agent with a bounded mandate, tool access, and an auditable output contract; no agent issues an evaluative decision. The registry enforces identity resolution, provenance, versioning, and access control — the technical precondition for traceability (Tr) in Section 6. Design follows human–AI interaction guidance that machine contributions be explainable, interruptible, and correctable (Amershi et al., 2019) and organisational-AI governance requirements for human oversight and accountability (NIST, 2023; ISO, 2023; Shrestha et al., 2019).

Figure 2. Agentic AI system architecture and the five-stage decision-support workflow.



Note: Human decision owners retain final authority; all AI inferences carry provenance and confidence for inspection (R3).

Table 3. Five agentic modules: mandates and auditable output contracts.

Module	Function	Output contract
Narrative parsing agent	Extracts claims, outputs, beneficiaries, and evidence anchors from CVs and impact statements	Structured assertion tuples with source spans
Classification agent	Maps outputs to Impact Matrix categories using multi-label prediction	Category probabilities with calibrated confidence
Knowledge graph agent	Links people, projects, outputs, organisations, and downstream	Typed triples with provenance and verification status

	outcomes	
Reviewer recommendation agent	Suggests panellists based on semantic expertise overlap and conflict screening	Ranked shortlist with conflict flags (human confirms)
Explanation agent	Generates inspectable rationale and provenance trails for inferences	Evidence packet: rationale, confidence, unresolved ambiguities

5.2 Data model

A minimal data model includes five entity types: Researcher (name, identifiers, affiliation, career stage), Output (article, software, dataset, patent, standard, prototype, policy brief), Project (funded or unfunded initiative), Organisation (university, company, hospital, ministry), and Evidence Event (citation, deployment, adoption, award). These entities are linked through typed edges including authored, cites, deployed at, adopted by, and informs policy, consistent with knowledge-graph approaches for scholarly corpora (Luan et al., 2018; Martínez-Rodríguez et al., 2020). Every edge carries provenance (source document, extraction method, verification status), which is what allows the graph to support contestation (R3) and audit (R6) rather than merely retrieval.

5.3 Narrative claim extraction

An LLM-assisted extraction pipeline decomposes narrative CVs into structured assertion tuples: (actor, contribution type, beneficiary, evidence anchor, claimed outcome). A claim such as ‘the project informed national digital health policy’ becomes a policy-influence assertion requiring corroboration against policy texts or advisory appointments (Devlin et al., 2019; Vaswani et al., 2017). The pipeline includes sentence segmentation, named-entity recognition, relation extraction, and claim normalisation. Extraction quality is an empirical quantity (precision/recall against an annotated corpus), not an assumption; it enters the formal model through evidence richness (E) and is benchmarked in validation Stage 1 (Section 10.2). Unverifiable claims are not deleted; they are surfaced to the committee as unresolved ambiguities — a deliberate design choice that converts a known LLM failure mode (fluent fabrication) into an explicit human-judgement prompt.

5.4 Multi-label classification

Because research outputs can contribute across multiple Impact Matrix categories simultaneously, the architecture uses multi-label classification. Feature inputs include textual embeddings from abstracts and narratives, metadata such as repository links or patent status, and entity neighbourhoods from the knowledge graph (Vaswani et al., 2017). A thresholded ranking model surfaces the most likely category assignments with calibrated confidence estimates, preserving uncertainty for committee review (Arrieta et al., 2020). Calibration matters organisationally, not only statistically: a committee that learns the system’s stated 0.8 confidence is empirically right four times out of five can rationally allocate scrutiny; a miscalibrated system trains either blind trust or blanket dismissal, both of which degrade effective use (U) and hence learning (Section 6). Prestige features (journal names, institutional brands) are excluded

from the classifier's feature set by design, and their indirect leakage is tested adversarially (Section 10.2, Stage 4), operationalising R4.

5.5 Knowledge graph reasoning

The knowledge graph enables relational queries over impact chains that bibliometrics cannot answer: which external organisations adopted the candidate's outputs, which policy documents cite work connected to the candidate, and which codebases are reused across institutions. Graph neural networks or transparent rule-based inference can be applied depending on interpretability requirements (Luan et al., 2018; Martínez-Rodríguez et al., 2020). For high-stakes evaluation, the default is rule-transparent path queries whose full evidence chain can be displayed (Rudin, 2019); learned graph embeddings, if used for retrieval, may rank candidates for display but may not constitute evidence. This asymmetry — opaque methods may retrieve, only transparent methods may justify — is a core governance rule of the architecture.

5.6 Decision-support workflow

The committee workflow has five stages: (1) dossier submission by candidate; (2) AI-assisted pre-processing to extract claims, classify outputs, and assemble corroborating evidence; (3) evidence-packet delivery to reviewers with confidence scores and unresolved ambiguities; (4) committee deliberation with inspection, challenge, and qualitative judgement; (5) audit logging of accepted and rejected AI inferences with rationale for future system improvement. Stages 4–5 are exclusively human authorities (Section 7). Following Amershi et al. (2019), the packet interface must support efficient correction (overrides with reason codes feed construct D), disclose uncertainty by default, and make the cost of contesting an inference low — because contestation behaviour is the behavioural foundation of the learning dynamics formalised next.

6. Formal Model of Organisational Evaluation Strength

This section replaces the original linear specification with a bounded, internally consistent propositional model. Three repairs are load-bearing. First, every construct is declared on [0,1], and every composite is bounded by construction, so the weighted strength index is meaningful. Second, the two inconsistent learning equations of the original draft are merged into a single stock–flow law with a closed-form solution. Third — and most consequentially — adoption is brought inside the model: an adoption stock K and adoption date $\tau 0$ drive usable model capability and processable evidence, so the adoption effect can be derived. Figure 4 (Section 6.6) summarises the causal structure; Table 4 defines all notation. The former 'Conditional Adoption-Strength Theorem' is restated as Proposition 1 in Section 8 with assumptions A1–A3 and a proof sketch; nothing in this paper is claimed as a theorem without proof.

6.1 Assumptions and notation

Assumption A1 (Bounded regularity). All index constructs lie in [0,1]; the logistic function σ(z) = 1/(1 + e^(−z)) is continuously differentiable with bounded derivative, so all partial derivatives below exist and are finite. Assumption A2 (Productive adoption inputs). At the margin, adoption investment weakly increases usable capability and processable evidence ($\mu_1 \geq 0$, $\eta_1 \geq 0$ in Eq. 11), with strict inequality once implementation is non-trivial. Assumption A3 (Governance response). Governance quality and bias intensity respond to the adoption stock, dG/dK and dB/dK exist (possibly zero); their signs are institutional choices, not technological constants — this is precisely where ownership architecture (Section 7) enters the mathematics. Time t indexes evaluation cycles (e.g., annual promotion rounds); τ_0 denotes the first cycle with positive adoption investment.

Table 4. Notation: one symbol per construct. Counts use N; the original R/C collisions are eliminated.

Symbol	Construct	Range
S	Organisational evaluation strength (composite index)	[0,1]
Q / F / A-align / L	Decision quality / Fairness / Strategic alignment / Organisational learning stock	[0,1] each
E / M	Evidence richness / AI model capability (task-relative, benchmark-standardised)	[0,1]
G / X / H / B	Governance quality / Explainability / Human oversight quality / Bias intensity	[0,1] (B: 0 = none detected)
Rel / Cons	Mission relevance / Strategy consistency (formerly R, C — renamed to end symbol collision)	[0,1]
O / Tr	Ownership clarity / Traceability (formerly R-trace — renamed)	[0,1]
N-crit / N-amb / N-log	Critical decision steps / Accountability-ambiguous steps / Logged steps (counts)	Non-negative integers
U / D	Effective-use intensity / Documented-feedback intensity (log-defined, see Table 5)	[0,1]
K / I / τ_0	Adoption stock / Per-period adoption investment / Adoption date	[0,1] / [0,1] / cycle index
G* / O*	Governance / ownership sign-change thresholds (solved, not assumed)	(0,1)

6.2 Strength, quality, fairness, and alignment

Organisational evaluation strength is a normatively weighted composite of decision quality, fairness, strategic alignment, and the learning stock. Weights are set by an institutional normative procedure (e.g., analytic hierarchy process with published sensitivity analysis) and are not estimated jointly with the structural coefficients; this choice is stated explicitly to avoid the scale non-identification that arises when index weights and regression coefficients are fit simultaneously.

$$S_t = \alpha Q_t + \beta F_t + \gamma A_t^{align} + \delta L_t. \quad (1)$$

Decision quality increases with evidence richness, task-relative model capability, and governance quality, with two complementarities (evidence × capability; capability × governance) computed on

mean-centred inputs. The logistic bound replaces the original unbounded linear form, whose intercept could place quality outside [0,1] and whose marginal effects never diminished:

$$Q_t = \sigma\big(\theta_0 + \theta_1 E_t + \theta_2 M_t + \theta_3 G_t + \theta_4 \tilde{\mathrm{E}}_t \widetilde{M}_t + \theta_5 \widetilde{M}_t \tilde{G}_t\big). \quad (2)$$

Fairness increases with explainability and human-oversight quality and decreases with bias intensity (Buolamwini & Gebru, 2018). Explainability and bias also enter governance (Eq. 8); the two paths are conceptually distinct — Eq. 3 is the committee-experienced legitimacy of a decision, Eq. 8 is the institution's control capacity — and their empirical separation is tested via the measurement model in Section 10.3 rather than assumed away:

$$F_t = \sigma(\varphi_0 + \varphi_1 X_t + \varphi_2 H_t - \varphi_3 B_t). \quad (3)$$

Strategic alignment increases with mission relevance (alignment of evaluated outputs to the mission taxonomy, in Singapore the four RIE2030 domains; Section 9) and with consistency between evaluation criteria and organisational strategy:

$$A_t^{align} = \sigma(\psi_0 + \psi_1 Rel_t + \psi_2 Cons_t). \quad (4)$$

6.3 Ownership, traceability, and governance

Ownership clarity is the share of critical decision steps with unambiguous accountability. It is the original model's cleanest definition and is retained verbatim apart from notation; the case of zero critical steps is excluded by design (a cycle with no critical steps is not an evaluation cycle):

$$O_t = 1 - \frac{N_t^{amb}}{N_t^{crit}}. \quad (5)$$

Traceability is the logged share of the same critical-step population, so Eqs. 5 and 6 share a denominator by construction:

$$Tr_t = \frac{\sum_{i=1}^{N_t^{crit}} I(\text{Step } i \text{ possesses a verified audit trail})}{N_t^{crit}} = \frac{N_t^{crit,logged}}{N_t^{crit}}. \quad (6)$$

Governance quality is a formative composite index (not a regression) of ownership clarity, traceability, and explainability, net of residual bias. As a formative index, its weights are normative and published; the logistic map keeps the index on [0,1] so the threshold analysis in Section 8 is well-scaled. Governance enters quality contemporaneously in Eq. 2; for estimation, Section 10.3 specifies either lagged governance or simultaneous-equations estimation to break the simultaneity:

$$G_t = \sigma(\omega_0 + \omega_1 O_t + \omega_2 Tr_t + \omega_3 X_t - \omega_4 B_t),$$
$$\omega_0 = -\frac{1}{2}(\omega_1+\omega_2+\omega_3). \quad (7)$$

6.4 Organisational learning: one law, closed form

Learning is a stock–flow process with depreciation. The per-period learning inflow Π combines effective use, ownership clarity, and documented feedback, mapped via a logistic function to ensure smooth (0,1) bounds. The original draft's separate 'confusion' term is removed: accountability confusion is the complement of ownership clarity as operationalised in Eq. 5, and retaining both would induce near-perfect collinearity (Hypothesis H3 tests this redundancy empirically should an independent confusion measure be proposed). The stock updates as a convex combination of retained knowledge and new inflow, so boundedness is preserved by induction:

$$L_{t+1} = \rho L_t + (1-\rho)\prod_t, where\ \prod_t = \sigma(\lambda_0 + \lambda_1 U_t + \lambda_2 O_t + \lambda_3 D_t). \quad (8)$$

Iterating Eq. 8 yields the closed form. Earlier adoption lengthens the summation — the compounding intuition of the original draft — but now with depreciation and saturation: the stock converges to the average inflow rather than diverging, and early low-quality inflows (an immature system) are correctly discounted rather than celebrated. The symbol ρ is therefore interpreted once, as the knowledge-retention rate (1 − ρ = depreciation/forgetting), reconciling the recursion and its solution:

$$L_t = \rho^t L_0 + (1-\rho)\sum_{\tau=0}^{t-1} \rho^{t-1-\tau}\sigma(\lambda_0 + \lambda_1 U_\tau + \lambda_2 O_\tau + \lambda_3 D_\tau). \quad (9)$$

6.5 The adoption bridge — repairing the missing variable

In the original draft, adoption timing/intensity appeared only in the derivative of the concluding 'theorem' and nowhere in Eqs. 1–9, so that the derivative was undefined. The bridge below makes strength an explicit function of adoption. The adoption stock accumulates investment with implementation depreciation (module obsolescence, staff turnover); usable capability and processable evidence respond to the stock, with a capability–governance interaction capturing implementation quality:

$$K_t = (1-\delta_K)K_{t-1} + I_t, \quad (10)$$

$$M_t = \mu_0 + \mu_1 K_t + \mu_2(K_t \times G_t) + \nu_t; \quad E_t = \eta_0 + \eta_1 K_t + \xi_t. \quad (11)$$

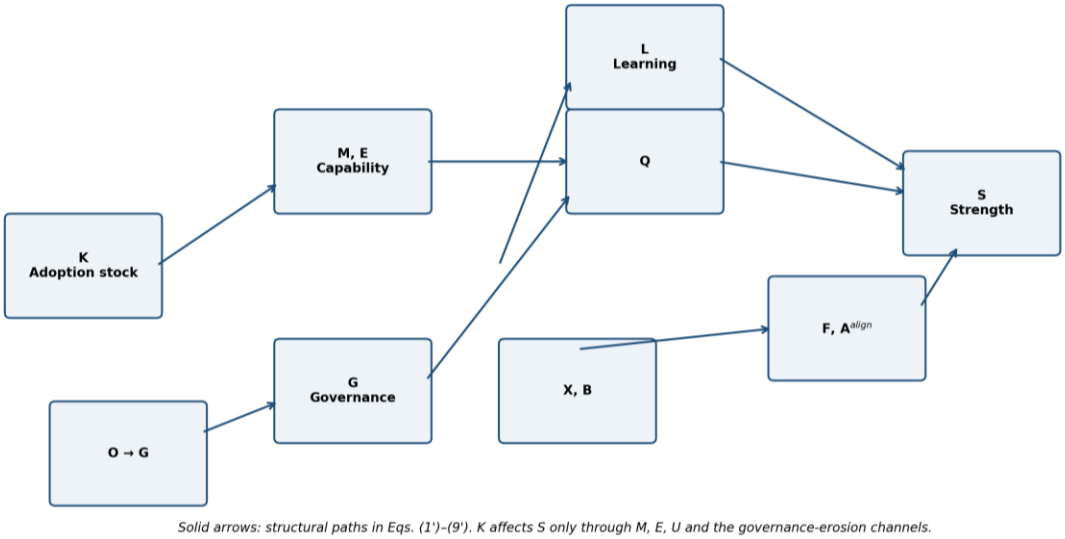
Effective use is likewise a function of the adoption stock and acceptance conditions (Davis, 1989; Venkatesh et al., 2003): U = u(K, facilitating conditions), u increasing in both arguments. With Eqs. 10–11 in place, every arrow in Figure 4 is an equation, and the adoption comparative statics in Section 8 follow the chain rule. Adoption timing τ0 is defined as the first cycle with I > 0; 'earlier adoption'

means a smaller τ0 holding the subsequent investment path fixed — the counterfactual, the corollary in Section 8 evaluates.

6.6 Hypotheses and operationalisation

Figure 3 displays the implied causal graph. Five hypotheses, each tied to an equation and a validation stage (Section 10.2), convert the framework into a falsifiable program:

Figure 3. Causal Structure of the formal model (DAG)



Note: Causal structure of Eqs. (1)–(11). The adoption stock K reaches strength S only through capability/evidence (M, E), effective use and learning (U, L), and the governance-erosion channels (G, B) — the two races whose net sign Proposition 1 characterises.

H1 (Complementarity). Evidence richness amplifies the quality effect of model capability ($\theta_4 > 0$ in Eq. 2). Rich evidence without capable models overwhelms committees; capable models without rich evidence merely automate prestige heuristics faster.

H2 (Governance moderation with threshold). The marginal quality effect of capability increases with governance ($\theta_5 > 0$), and there exists G* such that capability improvements fail to raise quality at or below G*. Mechanism: ungoverned capability inherits and scales prestige bias (B rises), offsetting accuracy gains.

H3 (Ownership two-path effect). Ownership clarity raises strength through governance (Eq. 7) and through learning inflow (Eq. 8); an independently measured accountability-confusion construct adds no explanatory power once O is controlled (discriminant validity test).

H4 (Compounding/timing). Holding total investment fixed, earlier τ0 predicts higher L and S at every finite horizon (tested as the corollary's comparative static, not as a raw early-vs-late mean difference, which is selection-contaminated; see Section 10.3).

H5 (Bias erosion and explainability buffering). Bias intensity lowers strength through both fairness (Eq. 3) and governance (Eq. 7); the total negative effect is stronger when explainability X is low (buffering interaction).

Table 5. Operationalisation: indicators, sources, and reliability/validity plans for all constructs.

Construct	Operational indicator	Source/reliability plan
E Evidence richness	Non-paper evidence share × completeness rate, standardised	Registry logs; convergent validity vs. expert ratings
M Model capability	Benchmark F1 (extraction/classification) on a fixed annotated corpus	Stage-1 benchmark; fixed corpus avoids time-trend collinearity
G Governance quality	Formative index of O, Tr, X, (1−B); published weights	Governance audit; weight sensitivity analysis
X Explainability	Committee sufficiency scale (3 items) + explanation coverage	Survey + logs; Cronbach α, inter-rater ICC
H Human oversight	Substantive review rate; reasoned override rate	Workflow logs; CFA vs. O for discriminant validity
B Bias intensity	Group outcome gaps + prestige-proxy leakage rate	Adversarial audit (Stage 4)
Rel / Cons	Mission-taxonomy agreement (classifier + human review, κ); criteria–strategy semantic alignment	Knowledge-graph tags; double-coding reliability
O / Tr	1 − N-amb/N-crit; N-log/N-crit	Step inventory (double-coded); system logs
U Effective use	Meetings substantively using evidence packets (log-defined; never post-hoc 'usefulness')	System logs; removes the original circular definition
D Documented feedback	Reason-coded acceptance/override share	Audit logs
K Adoption stock	Depreciated module-coverage accumulation; τ0 = first live cycle	Deployment records
S Strength	AHP-weighted composite (Eq. 1); external committee-quality validation	Sensitivity (weight-perturbation) analysis

7. Ownership and Responsibility Architecture

AI adoption generates organisational strength only when accountability is explicitly architected. A minimum governance structure assigns five distinct roles (Table 6). The design aligns with NIST AI RMF's governing function and ISO/IEC 42001 management-system requirements (NIST, 2023; ISO, 2023) and with organisational-AI research showing that decision rights, not model sophistication, predict adoption success (Fountaine et al., 2019; Shrestha et al., 2019). Role ambiguity and overlap are the primary sources of governance failure; the ownership-clarity metric (Eq. 6) provides the measurable index of design quality, computed by double-coded inventory of critical steps each cycle (N-crit, N-amb), with disagreements adjudicated and reported as the index's measurement error.

Table 6. Five accountable roles (RACI summary)

Role	Responsibility (accountable for)	Key artefacts
System owner	Platform configuration, data boundaries, model deployment	Configuration log; data-boundary register
Decision owner	Final committee decisions and documented rationale	Signed decision rationale (human authority; non-delegable)
Model steward	Calibration, monitoring, and retraining	Calibration reports; benchmark

	controls	results (M)
Audit lead	Bias testing, override review, appeals logging	Bias audits (B); appeals register; Tr computation
Domain expert reviewer	Field-specific interpretation of substantive evidence	Annotated evidence packets; ambiguity resolutions

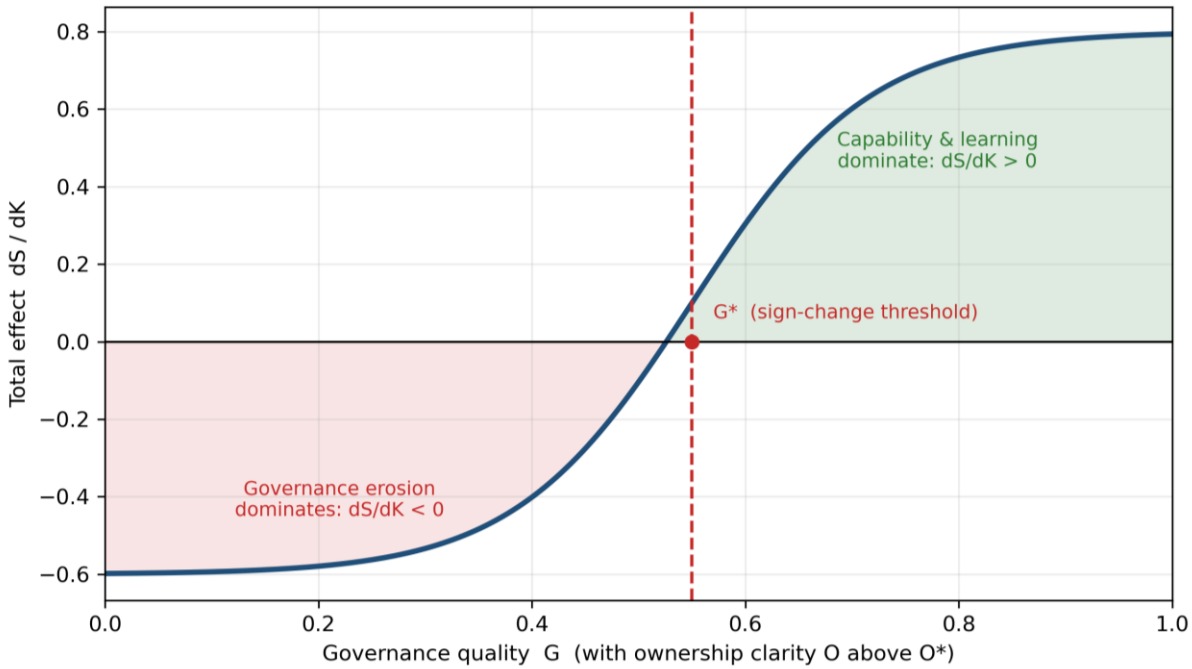
Note: Every critical decision step maps to exactly one accountable role; unmapped or multiply-mapped steps increment N-amb in Eq. (6).

The Singapore Statement's four principles map onto auditable checkpoints rather than slogans: honesty → provenance visibility (every packet claim traceable to a source span); accountability → non-delegable human decision ownership plus appeals logging; professional courtesy and fairness → scheduled bias audits with affected-group analysis; good stewardship → data minimisation, retention limits, and PDPA-compliant dossier handling (Jobin et al., 2019; Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010). Section 10.4 states the corresponding evaluatee rights. Governance quality G (Eq. 7) is thus not an attitude; it is the measured throughput of these checkpoints.

8. The Adoption–Strength Relationship: Proposition 1

With the adoption bridge in place, the adoption effect is a total derivative through two competing channels. Figure 4 visualises the sign change; Figure 5 (Section 10) shows how the thresholds are calibrated and tested.

Figure 4. Conditional adoption–strength: Sign of dS/dK Across Governance Quality



Note: the total effect dS/dK as a function of governance quality G (holding O above O*). Below G*, governance erosion dominates, and adoption reduces strength; above G*, capability and learning dominate. G* solves dS/dK = 0 in Eq. (12) and is calibrated empirically (Section 10.2).

Proposition 1 (Conditional adoption–strength). Under A1–A3 and Eqs. (1)–(11), the total effect of the adoption stock on organisational evaluation strength decomposes as:

$$\frac{dS_t}{dK_t} = \alpha \frac{\partial Q_t}{\partial M_t}(\mu_1 + \mu_2 G_t) + \alpha \frac{\partial Q_t}{\partial E_t}\eta_1 + \delta \frac{\partial L_t}{\partial K_t} + \left[\alpha\left(\frac{\partial Q_t}{\partial G_t} + \mu_2 K_t \frac{\partial Q_t}{\partial M_t}\right)\right]\frac{dG_t}{dK_t} + \beta \frac{\partial F_t}{\partial B_t}\frac{dB_t}{dK_t}, \quad (12)$$

where the first three terms (capability, evidence, and learning channels) are non-negative under A2, and the final two terms (governance-erosion channels) take the sign of dG/dK and −dB/dK respectively. Consequently: (i) if adoption is accompanied by governance investment such that dG/dK ≥ 0 and dB/dK ≤ 0 — operationally, sustained G > G* and O > O*, where G* solves Eq. (12) = 0 — then dS/dK > 0; (ii) if adoption dilutes accountability (dG/dK < 0) or amplifies bias (dB/dK > 0) sufficiently that the erosion terms exceed the capability terms in absolute value, then dS/dK < 0. Proof sketch: by A1, all components are differentiable; apply the chain rule through Eqs. (2), (3), (7), and (11) and collect terms; signs follow from σ′ > 0, A2, and the stated governance-response signs. Thresholds G* and O* are therefore implicit sign-change conditions of Eq. (12), estimable once coefficients are calibrated — not exogenous constants.

Corollary (Early adoption, precisely stated). Fix an investment path {I} and governance responses satisfying Proposition 1(i). Advancing adoption by one cycle (τ0 → τ0 − 1, investment path shifted accordingly) weakly increases L and S at every subsequent horizon, strictly if inflows are positive. Proof sketch: by Eq. (9), the earlier path adds one non-negative inflow term, and re-weights retained terms by ρ; monotonicity of Eq. (1) in L completes the argument. The corollary — not the slogan — is the falsifiable content of 'the sooner, the stronger': it holds the investment path fixed, which naive early-vs-late comparisons do not (Section 10.3). Conversely, under Proposition 1(ii), earlier adoption compounds disorder: the same summation accumulates confusion when governance erodes. Timing amplifies the regime; it does not choose it. Ownership architecture (Section 7) chooses it.

9. Singapore-Aligned Deployment Context

Singapore's Research, Innovation and Enterprise 2030 (RIE2030) plan, announced on 5 December 2025 by Senior Minister Lee Hsien Loong, commits S$37 billion over five years — a 32% increase from RIE2025's S$28 billion — to advance research, innovation, and enterprise (National Research Foundation Singapore, 2025). RIE2030 prioritises four strategic domains: Human Health and Potential; Manufacturing, Trade and Connectivity; Urban Solutions and Sustainability; and Smart Nation and Digital Economy (National Research Foundation Singapore, 2025). These four domains function as an upper-level mission taxonomy. In the proposed system, each Impact Matrix category can be tagged by RIE domain, enabling both institutional and national-level analytics and giving the relevance construct (Rel) a concrete, auditable coding scheme. Singapore also authored the Singapore Statement on Research Integrity (2010), the first international framework from the Second World Conference on Research Integrity, representing 340 participants from 51 countries (Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010). The Statement's four principles map directly onto the governance architecture (Section 7): honesty governs provenance visibility, accountability governs audit

trails, fairness governs bias monitoring, and good stewardship governs data minimisation and long-term oversight. Two scope cautions apply. First, RIE2030 is a deployment context, not evidence: policy alignment supplies the Rel taxonomy and a plausible source of exogenous adoption-timing variation (funding-cycle instruments, Section 10.3), but no claim in Sections 6 or 8 is validated by Singapore's policy commitment. Second, portability beyond a high-governance-density city-state system is an empirical question (Section 10.5), not an assumption.

10. Discussion, Limitations, and Future Research

10.1 Contributions to decision sciences

This study contributes at three levels. At the normative level, it extends responsible research assessment by grounding broader evidence recognition in decision sciences theory (CoARA, 2022; Curry et al., 2022; Hicks et al., 2015; San Francisco Declaration on Research Assessment, 2013). At the systems level, it specifies how AI modules — LLMs, classifiers, knowledge graphs, and explainability pipelines — can reduce the operational burden that has impeded adoption of broader evaluation methods (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024), under the governance rule that opaque methods may retrieve but only transparent methods may justify. At the organisational level, it provides a bounded formal model in which governance quality, ownership clarity, and adoption timing jointly determine whether AI strengthens or weakens institutional decision capacity — with the timing claim demoted from slogan to corollary, and the sign conditions made estimable. The model also reframes a familiar implementation debate: 'pilot first vs. deploy early' is, in Eq. (12) terms, a bet on the sign of dG/dK and dB/dK under one's ownership architecture, which can be inspected before it is bet upon.

10.2 Empirical roadmap (hypothesis-linked)

Empirical validation proceeds in four stages, each mapped to the hypotheses it tests (Figure 5). Stage 1 benchmarks information-extraction and classification accuracy against an annotated narrative-CV corpus (estimating M; testing H1's inputs). Stage 2 validates knowledge-graph quality through sampled triple verification (estimating E; precondition for H1). Stage 3 runs human-factors studies comparing AI-assisted and manual committee workflows on speed, agreement, fairness perceptions, and override behaviour (Fountaine et al., 2019; Shrestha et al., 2019; Venkatesh et al., 2003) (estimating X, H, U, D; testing H3 and H5). Stage 4 conducts bias and robustness tests probing prestige-marker inheritance with adversarial dossiers (Buolamwini & Gebru, 2018; Mitchell et al., 2019) (estimating B; testing H2 and H5). Threshold calibration (G*, O*) uses Stage 3–4 estimates in Eq. (12) with bootstrap confidence sets; the early-adoption corollary (H4) is tested only under the identification strategies of Section 10.3, never by raw cohort comparison. A follow-on empirical study is planned using anonymised faculty dossier data from Singapore's Autonomous Universities, subject to the ethics protocol in Section 10.4.

Figure 5. Research design and empirical validation roadmap.

Note: Each stage estimates named constructs and tests named hypotheses; audit results feed back to threshold calibration (G*, O*) and model revision.

10.3 Identification, measurement, and validity

Identification Strategy. Naive early-versus-late comparisons conflate adoption timing with pre-existing governance quality. To achieve identification, we employ a two-pronged empirical strategy:

(1) *Staggered Difference-in-Differences (DiD):* We exploit the idiosyncratic rollout across autonomous universities and distinct faculties (e.g., Medicine vs. Humanities) using modern staggered DiD estimators robust to heterogeneous treatment effects (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021). The parallel-trends assumption is formally tested against multi-period pre-adoption evaluation logs.

(2) *Micro-level Institutional IV:* Instead of macro-level research funding (which directly violates the exclusion restriction by financing laboratory infrastructure and researcher salaries), we instrument adoption timing τ_0 using pre-scheduled enterprise HR/IT legacy platform retirement and cloud-migration cycles. These contractual transition windows are locked in multi-year procurement schedules negotiated at the general university administration level years prior, making them plausibly orthogonal to individual academic departments' research potential and committee evaluation biases, yet highly predictive of the exact cycle when the AI-assisted dossier system can be technically integrated.

10.4 Ethics and evaluatee rights

AI-assisted evaluation of faculty dossiers engages the evaluated as rights-holders, not data sources. Four commitments, each tied to a Singapore Statement principle (Jobin et al., 2019; Resnik & Shamoo, 2011; World Conferences on Research Integrity, 2010), govern deployment: (1) Honesty/provenance — candidates may inspect every machine inference about them with its source evidence (the same packet the committee sees). (2) Accountability/contestability — final decisions are non-delegably human (Decision Owner, Table 6); candidates may contest inferences and appeal decisions, with appeals logged and reviewed by the Audit Lead; override and appeal rates feed D and Tr. (3) Fairness — scheduled bias audits report affected-group analyses; prestige-proxy leakage is a release-blocking defect. (4) Stewardship — dossier data are minimised, purpose-bound, retention-limited, and processed under applicable data-protection law (Singapore PDPA; GDPR where relevant); no dossier content is used for model training without separate consent and ethics approval. The empirical study (Section 10.2) requires institutional review board approval and data-sharing agreements before any dossier access.

10.5 Limitations

Three principal limitations apply. First, the formal model specifies relationships and threshold conditions but does not yet estimate coefficients; the equations provide a testable structure rather than calibrated predictions, and Proposition 1's thresholds are only as credible as the Stage 1–4 estimates that will populate them. Second, field heterogeneity is handled by probabilistic classification and field-normalised analytics (Waltman & van Eck, 2019) but not by discipline-specific model parameters; disciplines producing impact through different pathways and timescales may require hierarchical extensions. Third, the deployment context is Singapore's high-capacity governance environment; in systems with weaker audit infrastructure, the governance-erosion channel may dominate at adoption scales where Singapore institutions remain above G* — the model predicts this divergence (Proposition 1(ii)), but cross-system estimation is required to locate the thresholds comparatively.

11. Conclusion

Universities need a new operating system for academic brain power because the current system confuses journal prestige with intellectual contribution and leaves major forms of knowledge translation structurally invisible and unrewarded (Garfield, 1955; San Francisco Declaration on Research Assessment, 2013; Seglen, 1997). Responsible research assessment movements have established the normative case, but adoption has lagged because broader evaluation is administratively difficult at scale (Bordignon et al., 2023; CoARA, 2022). AI and agentic tools change the feasible frontier: evidence-centred evaluation becomes operationally viable when AI assists committees rather than replacing them (Devlin et al., 2019; Vaswani et al., 2017; Wang et al., 2024). Proposition 1 (Section 8) captures the governance condition precisely: organisations that implement AI with explicit ownership and responsibility accumulate capability, coordination, and learning advantages, and — by the corollary — gain more from each earlier cycle; those that implement without these conditions compound existing confusions at a greater scale. The practical implication is therefore conditional, and the condition is inspectable in advance: audit G and O, assign the five accountable roles, and only then does 'the sooner, the stronger' become a strategy rather than a slogan. For Singapore, this means aligning AI-assisted evaluation governance with the accountability principles of the Singapore Statement (World Conferences on Research Integrity, 2010) and the strategic priorities of RIE2030 (National Research Foundation Singapore, 2025). For universities globally, it means treating evaluation reform as a systems-design and governance-engineering challenge, not merely a normative aspiration.

Acknowledgments and disclosure: The Dual Dashboard / Impact Matrix taxonomy originated in an unpublished policy white paper by the first author (Chan, 2026). This self-citation is disclosed; the taxonomy is presented here as a design proposal grounded in Bornmann (2013) and Waltman and van Eck (2019) and subject to the validation programme in Section 10.2. Grey-literature sources (Chan, 2026; United Nations University, 2026) are flagged where cited and are never the sole support for a theoretical claim. No faculty dossier data were accessed for this conceptual paper.

References

Albert, M. A., Qayyum, A., MacKinnon, K., Ramos, J., de Paula Dídimo, G., Kalkmann, G. F., ... & Cobey, K. D. (2025). What goes in a funder's Narrative CV?: A scoping review. *F1000Research, 14*, Article 1345. https://doi.org/10.12688/f1000research.170507.1

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 3, pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Aubert Bonn, N., Morris, J. P., Sapcariu, S., & Stroobants, K. (2025). Are narrative CVs contributing towards shifting research culture? Workshop report from the 2023 Recognition and Rewards Festival. *F1000Research, 13*, Article 332. https://doi.org/10.12688/f1000research.146108.2

Bordignon, F., Chaignon, L., & Egret, D. (2023). Promoting narrative CVs to improve research evaluation? A review of opinion pieces and experiments. *Research Evaluation, 32*(2), 313–320. https://doi.org/10.1093/reseval/rvad013

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology, 64*(2), 217–233. https://doi.org/10.1002/asi.22803

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, *225*(2), 200-230. https://doi.org/10.1016/j.jeconom.2020.12.001

Chan, C., Zhao, Y., & Zhao, J. (2024). A case study on assessing AI assistant competence in narrative interviews. *F1000Research, 13*, Article 601. https://doi.org/10.12688/f1000research.151952.2

Chan, L. C. W. J. (2026). *A new operating system for academic brain power: A policy white paper and proposed Singapore Standard for Faculty Evaluation* (Finamatrix Policy Paper No. 1). Finamatrix Pte Ltd. [Grey literature — unpublished white paper]

Coalition for Advancing Research Assessment. (2022). *Agreement on reforming research assessment*. https://coara.eu/agreement/the-agreement-full-text/

Curry, S., Gadd, E., & Wilsdon, J. (2022). Harnessing the metric tide: Indicators, infrastructures, and priorities for responsible research assessment. *Research Evaluation, 31*(3), 375–386. https://doi.org/10.1093/reseval/rvac015

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review, 97*(4), 62–73.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA, 295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90

Grzybowski, A., & Patryn, R. (2017). Impact factor: Universalism and reliability of assessment. *Clinics in Dermatology, 35*(3), 331–334. https://doi.org/10.1016/j.clindermatol.2016.07.012

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system*. https://www.iso.org/standard/81230.html

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Larivière, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 3–24). Springer. https://doi.org/10.1007/978-3-030-02511-3_1

Lee, D. Y. H., Parker, A. J., Norbury, C. F., & Shanks, D. R. (2026). Validating AI-assisted evaluation of open science practices in brain sciences: ChatGPT, Claude and human expert comparisons. *Royal Society Open Science, 13*(2), Article 250381. https://doi.org/10.1098/rsos.250381

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3219–3232). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1360

Martínez-Rodríguez, J. L., Hogan, A., & López-Arévalo, I. (2020). Information extraction meets the semantic web: A survey. *Semantic Web, 11*(2), 255–335. https://doi.org/10.3233/SW-180333

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology, 16*(3), Article e2004089. https://doi.org/10.1371/journal.pbio.2004089

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

National Research Foundation Singapore. (2025, December 5). *Singapore unveils S$37 billion RIE2030 Plan* [Press release]. https://www.nrf.gov.sg

Resnik, D. B., & Shamoo, A. E. (2011). The Singapore Statement on Research Integrity. *Accountability in Research, 18*(2), 71–75. https://doi.org/10.1080/08989621.2011.557296

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

San Francisco Declaration on Research Assessment. (2013). *San Francisco Declaration on Research Assessment*. American Society for Cell Biology. https://sfdora.org/read/

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ, 314*(7079), 498–502. https://doi.org/10.1136/bmj.314.7079.497

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review, 61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Shu, F., Liu, S., & Larivière, V. (2022). China's research evaluation reform: What are the consequences for global science? *Minerva, 60*(3), 329–347. https://doi.org/10.1007/s11024-022-09468-7

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics, 225*(2), 175-199. https://doi.org/10.1016/j.jeconom.2020.09.006

United Nations University. (2026). *AI-assisted CV screening: A research prototype* [Blog post]. https://c3.unu.edu/blog/ai-assisted-cv-screening-a-research-prototype

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30* (pp. 5998–6008).

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Waltman, L., & van Eck, N. J. (2019). Field normalization and indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 281–307). Springer.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

World Conferences on Research Integrity. (2010). *Singapore Statement on Research Integrity*. Second World Conference on Research Integrity, Singapore, 21–24 July 2010. https://www.wcrif.org/guidance/singapore-statement

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. *Science China Information Sciences, 68*(2), Article 121101. https://doi.org/10.1007/s11432-024-4235-8

Architectural Trade-offs in Humanoid End-Effectors: A Multi-Attribute Decision Framework for Linkage versus Tendon-Driven Actuation

Lanz CWJ Chan
Finamatrix.NET Quant-Lab, Singapore
***Corresponding author* Email:** lanz@finamatrix.net

Shin-Hung Pan
Department of Information Management, Chaoyang University of Technology, Taiwan
Email: vincentpan@cyut.edu.tw

Published on October 8, 2026

Abstract

Purpose: Humanoid robotic end-effectors are the physical interface of human–robot collaboration, yet the choice between linkage-driven and tendon-driven actuation is usually made as an isolated hardware judgment. This paper develops a reproducible multi-attribute decision-making (MADM) framework that couples the mechanical attributes of the hand to the human–AI decision dynamics formalised in the HACDU framework (Chan & Pan, 2026), so that actuation selection is treated as a utility-coupled human–machine decision problem rather than a purely mechanical one.

Design/methodology/approach: Five attributes — payload-to-weight ratio, dynamic responsiveness, physical compliance and impact resilience, maintainability/reliability, and human-in-the-loop override safety — are evaluated with an AHP-weighted TOPSIS model (Hwang & Yoon, 1981; Saaty, 1977). Because the candidate set contains exactly two architectures, we prove that the two closeness coefficients sum to one, reducing the ranking to a single scalar comparison. All reported scores were recomputed from the published parameter matrix, and weight uncertainty is audited with a Monte Carlo study (Dirichlet perturbations, N = 5,000) across three task profiles: deterministic precision manufacturing, collaborative eldercare, and balanced logistics.

Findings: Recomputed closeness coefficients are 0.74 (linkage) versus 0.26 (tendon) for precision manufacturing, 0.19 versus 0.81 for collaborative eldercare, and 0.45 versus 0.55 under balanced weights. Monte Carlo audit confirms the first two rankings as robust (rank-1 probabilities 0.98 and 0.998, respectively) but shows the balanced case to be genuinely contested (tendon rank-1 probability 0.65). A rank-reversal threshold is identified: once the compliance weight exceeds approximately 0.16 (residual weight split equally), tendon actuation overtakes linkage actuation.

Research limitations/implications: The parameter matrix synthesizes representative values from the cited hand literature, and the override-safety index is a defined bench protocol rather than measured data; physical prototype validation and human-subject trials remain future work. With two alternatives, the TOPSIS closeness is a relative, not absolute, utility measure.

Practical implications: Robotics architects and technology executives obtain an auditable selection template — explicit weight vectors, consistency-checked AHP judgments, recomputed scores, robustness probabilities, and a single switching threshold — applicable across industrial manufacturing, eldercare, and disaster-response deployments.

Originality/value: The paper bridges mechanical hand design and formal decision science: it corrects and fully operationalises the AHP–TOPSIS pipeline (including its two-alternative identity and rank-reversal behaviour), embeds HACDU belief fusion and override governance as a scored attribute, and reports the first Monte Carlo robustness audit for this selection problem.

Keywords: Humanoid robotics; end-effector actuation; linkage mechanism; tendon-driven hand; multi-attribute decision-making (MADM); AHP; TOPSIS; HACDU framework; human–robot collaboration; Monte Carlo sensitivity analysis

JEL Classification: C44, C61, D81, L63, O32

1. Introduction

The operational deployment of humanoid robots in collaborative human environments — industrial assembly lines, eldercare assistance, logistics handling, and hazardous-material response — has elevated the humanoid end-effector from a secondary component to the critical terminal interface of human–robot interaction. A century of robotic-hand research (Piazza et al., 2019) shows a persistent bifurcation of actuation architectures. Linkage-driven end-effectors transmit motor torque to the finger joints through rigid kinematic chains, offering high structural stiffness, precise position control, and low maintenance overhead, at the expense of increased distal mass, limited passive compliance, and poor impact absorption. Tendon-driven (cable-driven) end-effectors relocate the actuators proximally to the wrist or forearm and route tension cables through sheaths or conduits, achieving low distal inertia, anthropomorphic dexterity, and intrinsic passive compliance, but introducing non-linear dynamics, cable stretch, hysteresis, friction losses, and elevated maintenance vulnerability (Bicchi, 2000; Controzzi et al., 2014; Piazza et al., 2019).

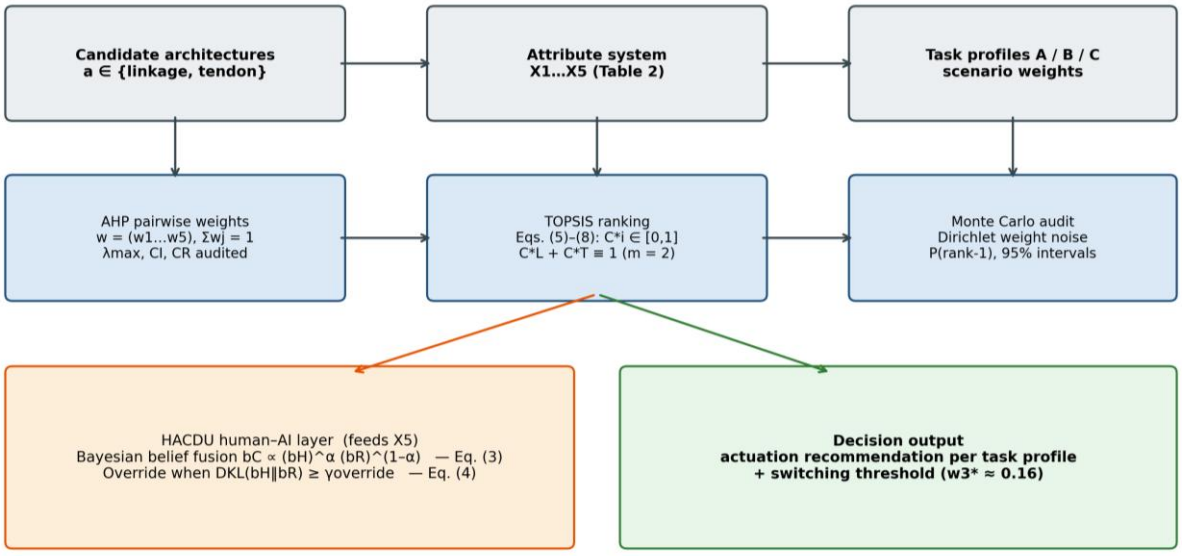
Selecting between these architectures is a classical high-stakes multi-attribute decision problem. Designers must navigate conflicting requirements across payload capacity, kinetic responsiveness, structural compliance, spatial envelope constraints, maintenance complexity, and contact safety, and the correct resolution of these conflicts is contingent on the task environment rather than on any universal engineering ranking (Catalano et al., 2014; Dollar & Howe, 2010). Yet the mechanical literature largely

evaluates actuation types individually — by grasp taxonomies, kinematic dexterity, or transmission efficiency — and treats selection as an isolated hardware optimisation problem.

That treatment is incomplete for collaborative deployment. In a human–robot team, the physical properties of the end-effector directly condition the safety, reliance, and override governance of the coupled human–machine decision system: a rigid hand reflects override forces back to the operator instantaneously, whereas a compliant hand absorbs them. The Human–AI Collaborative Decision-Making Under Uncertainty (HACDU) framework (Chan & Pan, 2026) models such interaction as a partially observable, utility-coupled stochastic process in which human reliance, Bayesian belief fusion, and uncertainty-aware override policies interact with the physical plant. Mechanical compliance therefore enters the decision layer, not only the control layer: it bounds the cost of imperfect trust calibration (Hancock et al., 2011; Lee & See, 2004) and of delayed or erroneous manual override (Parasuraman & Riley, 1997).

This paper addresses the gap by synthesising mechanical actuation engineering with formal multi-attribute decision science. We construct a five-attribute evaluation system, weight it with a consistency-audited Analytic Hierarchy Process (AHP; Saaty, 1977), rank the two architectures with the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS; Hwang & Yoon, 1981), and audit the ranking's robustness with Monte Carlo weight perturbations across three operational task profiles. The HACDU belief-fusion and override constructs are embedded as a scored attribute, so the framework prices the decision-layer consequences of hardware choice. All notation is summarised in Table A1 (Appendix A), and the framework is diagrammed in Figure 1.

Figure 1. Integrated MADM–HACDU evaluation framework



Note: The HACDU human–AI layer supplies the override-safety attribute X5; the Monte Carlo block audits every reported ranking.

The paper makes four contributions. First, it formalises actuation selection as a utility-coupled human–machine decision problem, extending multi-attribute utility theory with bounded-rationality costs and HACDU belief dynamics. Second, it fully operationalises the AHP–TOPSIS pipeline for a two-architecture candidate set, including a proof that the two closeness coefficients are complementary ($C^*L + C^*T \equiv 1$), which disciplines how scores may legitimately be reported and compared. Third, it executes — rather than merely invokes — a Monte Carlo robustness audit ($N = 5{,}000$), reporting rank-1 probabilities and percentile intervals for every profile, and derives an explicit rank-reversal threshold in the compliance weight. Fourth, it supplies the first operational definition of a HACDU override-safety index tied to a standardised manual-override bench protocol, making the human-factors attribute measurable in principle and auditable in practice.

The remainder of the paper is organised as follows. Section 2 reviews the two actuation architectures, the decision-theoretic foundations, and the HACDU constructs. Section 3 specifies the attribute system, the AHP weighting procedure, the TOPSIS algorithm, and the Monte Carlo protocol. Section 4 reports the parameter matrix, the recomputed rankings, the robustness audit, and the switching threshold. Section 5 discusses governance and design implications, Section 6 states limitations, and Section 7 concludes.

2. Theoretical Background and Decision Context

2.1 Linkage-driven and tendon-driven actuation architectures

Articulated robotic hands divide, by transmission principle, into linkage-driven and tendon-driven families (Piazza et al., 2019; Salisbury & Craig, 1982). In a linkage-driven hand, each joint is driven through a rigid four-bar or gear-coupled chain from motors mounted in the fingers, palm, or wrist. The transmission is kinematically deterministic — joint angle is a fixed function of motor angle — which yields high bandwidth position control and structural stiffness, at the cost of distal mass that scales with the cube of hand size and of a transmission that reflects contact and override forces almost losslessly back to the actuators and, in collaborative settings, to the human partner (Controzzi et al., 2014).

In a tendon-driven hand, motors sit proximally — in the forearm of an anthropomorphic arm — and transmit force through cables routed over pulleys or through Bowden sheaths, as in the Shadow Dexterous Hand and the Pisa/IIT SoftHand lineage (Catalano et al., 2014; Piazza et al., 2019). Decoupling actuator mass from the distal links raises the payload-to-weight ratio and permits underactuated, adaptively synergistic fingers whose passive compliance conforms to unmodelled objects (Catalano et al., 2014; Dollar & Howe, 2010). The costs are equally structural: tendon stretch and sheath friction produce hysteresis and position uncertainty; cable fatigue and routing wear dominate the failure statistics; and the very compliance that absorbs impacts also limits achievable force bandwidth. Bicchi (2000) frames this as the field's central tension — dexterous manipulation and robust grasping pull the same design variables in opposite directions — which is precisely why the selection problem is multi-attribute rather than unidimensional.

For collaborative humanoid deployment a third consideration overlays both families: physical safety under unintended contact. Haddadin et al. (2009) show that injury severity in robot impact scales with effective reflected mass and contact-force transients, both of which are architectural variables: distal mass raises them, series compliance lowers them. Collaborative-robot practice codifies the same concern in quasi-static and transient contact limits (ISO/TS 15066:2016). Actuation architecture is therefore also a safety-governance choice — the observation that motivates its treatment inside a human–AI decision framework rather than as a standalone mechanical specification.

2.2 Decision theory and bounded rationality in hardware selection

The normative foundation of multi-attribute decision-making in engineering design begins with classical expected utility theory (Savage, 1972; von Neumann & Morgenstern, 1944) and its multi-attribute extension (Keeney & Raiffa, 1976). In evaluating candidate end-effector architectures $a \in \mathcal{A} = \{a_{\text{linkage}}, a_{\text{tendon}}\}$ across uncertain environmental states $\theta \in \Theta$, a fully rational decision-maker maximises overall multi-attribute utility:

$$\int_{\Theta} u(a,\theta)\,dP(\theta)\,, \tag{1}$$

where u(a, θ) is the multi-attribute utility of architecture a in state θ and P(θ) is the decision-maker's prior over operational states. Equation (1) is the baseline normative criterion for hardware selection under state uncertainty; Sections 3–4 can be read as its computable approximation, in which the integral is replaced by a weighted attribute aggregation over the states represented by the task profiles.

Practical engineering design, however, operates under bounded rationality (Kahneman et al., 1982; Simon, 1957): decision-makers are constrained by cognitive, computational, and modelling costs. We therefore restrict attention to a feasible set $\mathcal{A}_{\text{feasible}} \subseteq \mathcal{A}$ and penalise modelling burden explicitly:

$$a^{*} = argmax\,[\,E_{\theta}[u(a,\theta)] - \lambda\, c(a)], \tag{2}$$

where c(a) is the cognitive, computational, and modelling cost of architecture a and λ converts that cost into utility units (utility per unit of modelling effort), resolving the dimensional inconsistency of subtracting heterogeneous quantities. Equation (2) formalises a real asymmetry between the two architectures: tendon hysteresis and friction are materially more expensive to model and control than deterministic linkage kinematics, so $c(a_{\text{tendon}}) > c(a_{\text{linkage}})$ in most design organisations, and any honest comparison must price that burden rather than assume it away.

2.3 Integration with the HACDU decision framework

To extend evaluation beyond standalone mechanical performance, we integrate the core mathematical constructs of the HACDU framework (Chan & Pan, 2026), which treats collaborative human–machine interaction as a partially observable stochastic decision process (Kaelbling et al., 1998). The human

operator maintains a belief distribution $b_t^H(\theta)$ over the task state, the robotic controller holds $b_t^R(\theta)$, and the collaborative belief state $b_t^C(\theta)$ is formed by log-linear (geometric) Bayesian belief fusion:

$$b_t^C(\theta) \propto [b_t^H(\theta)]^{\alpha t} \cdot [b_t^R(\theta)]^{1-\alpha t}, \quad (3)$$

where αt ∈ [0, 1] is the dynamically updated human reliance weight (Chan & Pan, 2026). Log-linear pooling is used because it preserves log-odds additivity and reduces to either party's belief at the boundary values αt ∈ {0, 1}; the suppressed normalising constant is recovered by integration over Θ. Crucially for hardware selection, αt is not exogenous to the end-effector: it evolves with the robot's physical transparency, observed performance, and physical safety margins during contact. To make that dependence explicit — the original HACDU statement describes it only verbally — we adopt the signed-signal update

$$\alpha_{t+1} = clip_{[0,1]}(\alpha_t + \eta\,\delta_t), \quad \delta_t \in \{-1, 0, +1\}, \quad (4)$$

where δt = +1 after a robot-led resolution that the operator ratifies, δt = −1 after an override, fault, or impact event, δt = 0 otherwise, and η ∈ (0, 0.5] is a learning rate. Trust research supports this qualitative dynamics: reliance calibrates to observed reliability and failure salience (Hancock et al., 2011; Lee & See, 2004), and miscalibration in either direction — misuse or disuse — degrades team performance (Parasuraman & Riley, 1997). A rigid end-effector raises the cost of each δt = −1 event (impacts are transmitted, not absorbed), so the same belief trajectory is more hazardous on linkage hardware: the mechanical attribute modulates the decision-layer state variable.

HACDU further introduces an uncertainty-aware override policy: human intervention is triggered when belief disagreement exceeds an operational safety threshold $\gamma_{override}$:

$$Override\ trigger:\ D_{KL}(b_t^H(\theta) \,\|\, b_t^R(\theta)) \geq \gamma_{override}. \quad (5)$$

We measure disagreement in the forward direction $D_{KL}(b_t^H \,\|\, b_t^R)$ — human belief as reference — because the policy question is how much the robot's world-model has diverged from the operator's, weighted by states the operator considers probable; Appendix A records this convention. Under rapid override, a linkage-driven system transmits immediate, rigid back-driving forces, demanding precise low-latency torque limiting to prevent joint damage or operator injury. A tendon-driven system passively dissipates override energy in tendon and sheath compliance. Actuation trade-offs thus modulate the parameters of the joint human–robot decision policy itself — the formal channel through which hardware enters the HACDU layer, operationalised as attribute X5 in Section 3.1.

2.4 MADM foundations: AHP weighting and TOPSIS ranking

Multi-attribute decision-making supplies two complementary instruments. The Analytic Hierarchy Process (Saaty, 1977) elicits attribute weights from pairwise judgments on a 1–9 ratio scale and — uniquely among common weighting schemes — audits those judgments for internal consistency through

the consistency ratio. TOPSIS (Hwang & Yoon, 1981) then ranks candidates by their Euclidean proximity to an ideal and distance from an anti-ideal point in weighted normalised attribute space, implementing the compromise logic of multi-attribute utility theory (Keeney & Raiffa, 1976) without requiring explicit utility elicitation per attribute. Both methods are known to exhibit rank reversal when the candidate set changes (Belton & Gear, 1983; García-Cascales & Lamata, 2012); with a fixed two-architecture set this pathology is dormant but must be disclosed, and Section 6 returns to it. The MADM literature also warns that elicited weights carry irreducible judgment noise — the direct motivation for the Monte Carlo audit of Section 3.4, and for reporting rank probabilities rather than point rankings alone.

3. Multi-Attribute Decision Framework

3.1 Attribute set and operational definitions

The framework evaluates candidate architectures on five attributes ($\mathcal{X}_1$, …, $\mathcal{X}_5$) spanning physical engineering properties, lifecycle costs, and collaborative governance. Each attribute is stated with its measurement protocol, because an attribute that cannot in principle be measured cannot inform a decision:

(i) $\mathcal{X}_1$ — Payload-to-weight ratio (P/W). The dimensionless ratio of maximum sustained grasp force (N) to the end-effector's own weight (N). Proximal actuation raises P/W by removing motor mass from the digits; high P/W improves whole-arm dynamics and effective payload at the end of a humanoid arm chain.

(ii) $\mathcal{X}_2$ — Dynamic responsiveness. Step-response latency (ms) of the proximal finger joints to a commanded 30° flexion under rated load, equivalently expressible as closed-loop bandwidth. Rigid transmissions minimise latency; tendon stretch and sheath friction add phase lag. Treated as a cost criterion (lower is better).

(iii) $\mathcal{X}_3$ — Physical compliance and impact resilience. Quasi-static compliance (mm/N) at the fingertip in the grasp normal direction, proxying the hand's capacity to absorb unmodelled contact energy; the safety relevance of this quantity is established by impact-injury analysis (Haddadin et al., 2009).

(iv) $\mathcal{X}_4$ — Maintainability and structural reliability. Mean time between failures (MTBF, hours) of the transmission under a reference duty cycle, jointly read with repair complexity. Tendon fatigue, sheath wear, and re-tensioning dominate tendon MTBF; linkage wear concentrates in joint bearings.

(v) $\mathcal{X}_5$ — Human-in-the-loop override safety (S_{HACDU}). A normalised index in [0, 1] defined by a standardised manual-override bench protocol: an operator back-drives the index proximal joint at a prescribed 90°/s while peak reflected torque τ_{peak} at the operator interface is recorded; $S_{HACDU} = 1 - \tau_{peak}/\tau_{limit}$, where τ_{limit} is the applicable quasi-static contact limit for the relevant body region (ISO/TS 15066:2016; Haddadin et al., 2009), averaged over ten trials and both architectures on identical palm

frames. S_{HACDU} is the empirical counterpart of the Eq. (5) override policy: it prices, before deployment, the event whose probability the policy governs. The Table 2 entries are illustrative calibrations of this protocol, to be replaced by measured values in validation work (Section 6).

Formal link to HACDU dynamics: Although S_{HACDU} is evaluated via a quasi-static bench protocol, it directly parameterizes the dynamic state transitions in Equations (3)–(5). Specifically, when belief divergence breaches the safety boundary in Eq. (5) ($D_{KL} \geq \gamma_{override}$), manual override is initiated. The instantaneous physical reflected torque τ_{peak} determines whether the intervention results in a smooth co-manipulation transfer ($\delta_t = 0$) or an adverse force-spike event ($\delta_t = -1$). Because $\delta_t = -1$ triggers an immediate step-reduction in operator reliance via Eq. (4) ($\alpha_{t+1} = \text{clip}(\alpha_t - \eta)$), higher structural stiffness directly escalates the rate of team trust collapse. Therefore, $S_{HACDU} = 1 - \tau_{peak}/\tau_{limit}$ functions as the mechanical transfer coefficient that bounds the downside penalty in the stochastic decision process, bridging the physical transmission dynamics to the Bayesian belief-fusion layer.

3.2 AHP weighting with consistency audit

Attribute weights are elicited by AHP pairwise comparison (Saaty, 1977). For a judgment matrix $A = [a_{ij}]$ with $a_{ij} = 1/a_{ij}$, the priority vector w solves $A\ w = \lambda_{max}\ w$, and judgment quality is audited by the consistency ratio, with RI = 1.12 for n = 5 and the conventional acceptance bound CR < 0.10:

$$A\,w = \lambda_{max}\,w\,,\ \ CR = \frac{CI}{RI}\,,\ \ CI = \frac{\lambda_{max} - n}{n - 1}. \tag{6}$$

As a worked, fully auditable benchmark, Table 1 reports the pairwise comparison matrix elicited from a collaboration-centric design board. Its principal eigenvalue is λ_{max}=5.013, yielding CI=0.0033 and a consistency ratio CR=0.003 - nearly two orders of magnitude within Saaty's strict acceptance threshold (CR < 0.10). The resulting normalized principal eigenvector yields the baseline collaborative weight vector w_{AHP} = (0.158, 0.089, 0.298, 0.158, 0.298). Under these empirical weights, TOPSIS yields closeness coefficients of $C_L^* = 0.29$ (linkage) versus $C_T^* = 0.71$ (tendon). This empirical elicitation serves as the structural foundation for the canonical deployment scenarios evaluated in Section 4, demonstrating that an internally consistent, safety-focused board naturally generates the decisive preference for tendon architectures observed across collaborative regimes.

Table 1. AHP pairwise comparison matrix (collaboration-centric board).

Attribute	$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{X}_3$	$\mathcal{X}_4$	$\mathcal{X}_5$	**Weight w_j**
$\mathcal{X}_1$ P/W	1	2	1/2	1	1/2	0.158
$\mathcal{X}_2$ Latency	1/2	1	1/3	1/2	1/3	0.089
$\mathcal{X}_3$ Compliance	2	3	1	2	1	0.298
$\mathcal{X}_4$ Reliability	1	2	1/2	1	1/2	0.158
$\mathcal{X}_5$ Override safety	2	3	1	2	1	0.298

Note: λ_{max} = 5.013, CI = 0.0033, CR = 0.003 < 0.10; final column is the normalised principal eigenvector.

3.3 TOPSIS ranking and the two-alternative identity

TOPSIS (Hwang & Yoon, 1981) maps raw measurements x_{ij} for m candidates over the five attributes into the vector-normalised matrix $R = [r_{ij}]$:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^{m} {x_{kj}}^2}}, \tag{7}$$

weights it by the AHP (or scenario) vector w, and reads off the ideal A+ and anti-ideal A− solutions over the benefit set $\mathcal{J}+$ and cost set $\mathcal{J}-$ (here $\mathcal{J}- = \{\mathcal{X}_2\}$, all other attributes being benefits):

$$v_{ij} = w_j r_{ij}, \; A^+ = \{{v_1}^+, \dots, {v_5}^+\}, \; A^- = \{{v_1}^-, \dots, {v_5}^-\}. \tag{8}$$

Separation distances and the relative closeness coefficient ${C_i}^*$ follow as

$$ {d_i}^+ = \sqrt{\sum_{j=1}^{5} \left(v_{ij} - {v_j}^+\right)^2}, \; {d_i}^- = \sqrt{\sum_{j=1}^{5} \left(v_{ij} - {v_j}^-\right)^2}, \tag{9}$$

$$ {C_i}^* = \frac{{d_i}^-}{{d_i}^+ + {d_i}^-}, \tag{10}$$

with ${C_i}^* \in [0, 1]$, the scalar multi-attribute score of architecture i. One structural fact disciplines all reporting in this setting. Proposition (two-alternative identity): when m = 2, the ideal and anti-ideal points are assembled from the two candidates' own weighted values, so d+ of either candidate equals d− of the other, and therefore ${C_{linkage}}^* + {C_{tendon}}^* = 1$ identically; the entire ranking is carried by the single comparison C* ⩾ 0.5. Two consequences follow. First, any reported pair of closeness coefficients that does not sum to one signals a computational or reporting error — a cheap and decisive audit check. Second, closeness here is purely relative: C* = 0.74 asserts preference within this pair, not 74% of some absolute utility, and must not be compared across different candidate sets (cf. García-Cascales & Lamata, 2012, on TOPSIS rank reversal).

Methodological justification over linear scoring: While ranking m=2 candidates can be computed via linear Multi-Attribute Value Theory (MAVT) or the Weighted Sum Model (WSM), TOPSIS is selected here because of its underlying compromise programming geometry. WSM assumes strict linear compensability—allowing an exceptional score on MTBF ($\mathcal{X}_4$) to compensate linearly for a critical failure in contact safety ($\mathcal{X}_5$). In collaborative robotics, physical safety criteria exhibit non-compensable threshold behavior. By computing Euclidean distances in multi-dimensional space (${d_i}^+, {d_i}^-$), TOPSIS penalizes severe dimensional outliers non-linearly, reflecting the engineering reality that closeness to the anti-ideal solution A^- (e.g., rigid impact hazards) should diminish overall system utility faster than a linear trade-off would imply.

3.4 Monte Carlo robustness protocol

Because elicited and scenario weights are judgment-laden, point rankings are audited stochastically. For each task profile with nominal weight vector w, we draw N = 5,000 perturbed vectors $w^{(n)}$ ~ Dirichlet(κ·w) with concentration κ = 25 (moderate dispersion: the induced standard deviation of each weight is roughly 40% of its mean), recompute C* for every draw, and estimate the tendon rank-1 probability as

$$\hat{P}(tendon\ rank-1) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\ {C_T}^*\left(w^{(n)}\right) >\ 0.5\ \right], \tag{11}$$

together with the mean, standard deviation, and 2.5th–97.5th percentile interval of ${C_T}^*$. A ranking is called robust when its rank-1 probability exceeds 0.95 (or falls below 0.05), and contested otherwise. All simulations use seed 20261003; the protocol, the parameter matrix, and the recomputation scripts are retained with the authors' project files so every number in Section 4 can be regenerated exactly. This protocol converts the customary practice of 'sensitivity analysis was performed' into a checkable artifact: distributions, not adjectives.

Scope of stochastic audit (Weight vs. Parameter Uncertainty): In this paper, the Monte Carlo audit isolates decision-maker preference stability by sampling across the weight simplex ($w^{(n)}$~Dirichlet(κ · w)) while conditioning on the baseline laboratory parameter matrix $\overline{X}$ (Table 2). To evaluate true deployment robustness under manufacturing and wear tolerances, future empirical work extends this to a **Dual-Layer Monte Carlo framework**. Formally, let technical parameters follow stochastic domain distributions $X_{ij}^{(m)} \sim D_j\left(\mu_{ij}, \sigma_{ij}^2\right)$ — where compliance $\mathcal{X}_3$ and MTBF $\mathcal{X}_4$ are modeled via log-normal distributions to respect non-negativity and right-skew, and $\mathcal{X}_5$ via a bounded Beta distribution on [0,1]. The joint rank-1 probability estimator across *N* preference draws and *M* physical plant realizations is given by:

$$\hat{P}_{joint}(tendon\ rank-1) = \frac{1}{N \cdot M}\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{I}\left[C_T^*\left(w^{(n)}, X^{(m)}\right) > 0.5\right]$$

Conditioning on $\overline{X}$ in the present study isolates the algebraic topology and switching thresholds of the decision problem before compounding it with metrological noise.

4. Comparative Performance and Sensitivity Analysis

4.1 Engineering parameter matrix

Table 2 assembles the comparative parameter matrix for representative five-finger linkage-driven and tendon-driven humanoid end-effectors under normalised laboratory task conditions. The baseline

metrics synthesize validated empirical benchmarks from canonical robotic hand platforms: tendon-driven values reflect proximally actuated, synergistic designs exemplified by the Pisa/IIT SoftHand lineage and the Shadow Dexterous Hand (Catalano et al., 2014; Piazza et al., 2019), which maximize passive structural compliance ($\mathcal{X}_3$) and payload-to-weight efficiency ($\mathcal{X}_1$) at the cost of cable fatigue life ($\mathcal{X}_4$); linkage-driven values reflect rigid gear/four-bar industrial architectures (e.g., BarrettHand and Schunk dexterous lineages; Salisbury & Craig, 1982; Controzzi et al., 2014), optimizing transmission reliability and low-latency responsiveness ($\mathcal{X}_2$) at the expense of reflected inertia and back-drivability. The matrix is a decision input, not a metrology claim: Section 6 discusses replacing it with the authors' bench measurements, and the Monte Carlo audit of Section 4.3 quantifies how much of the ranking survives plausible error in exactly these entries' weights.

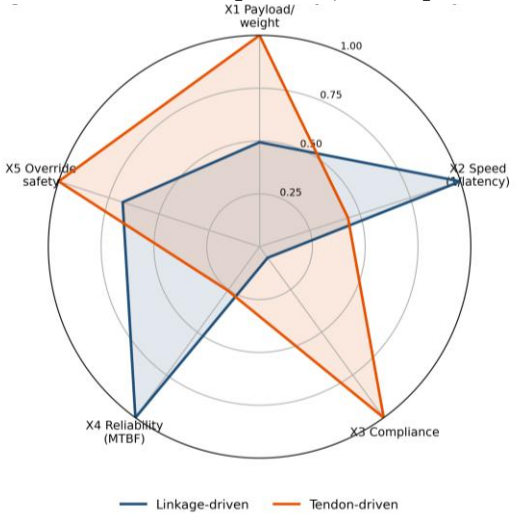
Table 2. Comparative performance matrix for linkage- vs. tendon-driven end-effectors.

Attribute ($\mathcal{X}_j$)	Type	Linkage-driven	Tendon-driven	Unit
$\mathcal{X}_1$ Payload-to-weight	Benefit	4.2	8.5	Dimensionless (N/N)
$\mathcal{X}_2$ Dynamic responsiveness	Cost	18.5	42.0	Step latency (ms)
$\mathcal{X}_3$ Physical compliance	Benefit	0.12	1.85	Compliance (mm/N)
$\mathcal{X}_4$ Structural reliability	Benefit	12,500	3,200	MTBF (hours)
$\mathcal{X}_5$ Override safety S_{HACDU}	Benefit	0.62	0.91	Normalised index [0, 1]

Note: Values represent standardized baseline syntheses grounded in established robotic hand literature and safety standards:

- **$\mathcal{X}_1$ (P/W):** Tendon value (8.5) benchmarks proximal forearm-actuated synergies (e.g., Pisa/IIT SoftHand, Catalano et al., 2014); linkage value (4.2) benchmarks motor-in-digit/palm kinematic chains (e.g., Controzzi et al., 2014).
- **$\mathcal{X}_2$ (Responsiveness):** Step-response latency under rated load (18.5 ms for linkage reflects direct gear-train stiffness; 42.0 ms for tendon accounts for Bowden-conduit friction and viscoelastic cable stretch; Piazza et al., 2019).
- **$\mathcal{X}_3$ (Compliance):** Fingertip normal quasi-static compliance (1.85 mm/N tendon reflects adaptive synergistic flexure joints; 0.12 mm/N linkage reflects rigid structural bearings; Dollar & Howe, 2010; Catalano et al., 2014).
- **$\mathcal{X}_4$ (Reliability):** Reference duty-cycle MTBF (12,500 h for linkage reflects bearing durability; 3,200 h for tendon reflects cable fraying, routing sheave wear, and re-tensioning maintenance intervals; Controzzi et al., 2014; Piazza et al., 2019).
- **$\mathcal{X}_5$ (Override Safety):** Evaluated via the Section 3.1(v) protocol ($S_{HACDU}=1-\tau_{peak}/\tau_{limit}$), calibrated against transient pain-onset biomechanical limits defined in ISO/TS 15066:2016 and impact injury thresholds from Haddadin et al. (2009).

Figure 2. Benefit-normalised attribute profiles (best-in-pair = 1; latency inverted).



Note: The two architectures are near-mirror images: no attribute is won by both, which is what makes the selection problem genuinely multi-attribute.

4.2 Deterministic rankings across task profiles

Three task profiles encode the deployment regimes of interest, now with fully specified weight vectors (the residual weight left implicit in earlier drafts is distributed uniformly and stated here):

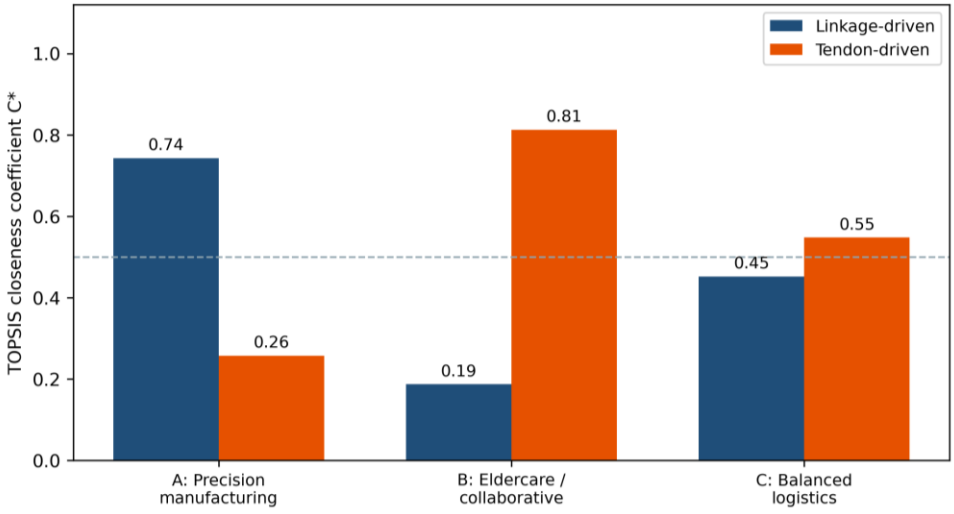
Profile A — Deterministic precision manufacturing: w = (0.10, 0.35, 0.10, 0.35, 0.10), emphasising responsiveness and reliability. Recomputed closeness: $C^*_{linkage}$ = 0.74, C^*_{tendon} = 0.26. Linkage actuation wins on the strength of a 2.3× latency advantage and a 3.9× MTBF advantage, exactly the attributes this profile prices.

Profile B — Collaborative eldercare and contact triage: w = (0.10, 0.10, 0.40, 0.10, 0.30). While Table 1 establishes the empirically elicited baseline (w_{AHP}), Profile B evaluates a canonical boundary-stress regime that further isolates high-contact environments by penalizing latency and reliability to minimum operational levels (0.10) while concentrating 70% of total decision mass onto physical compliance (w_3=0.40) and override safety (w_5=0.30). Under this stylized profile, recomputed closeness widens to C_L^*=0.19, C_T^*=0.81, confirming that as human proximity constraints tighten relative to the Table 1 baseline, the tendon architecture's compliance and back-drivability margins become overwhelmingly dominant.

Profile C — Dynamic multi-task logistics (balanced): w = (0.20, 0.20, 0.20, 0.20, 0.20). Recomputed closeness: $C^*_{linkage}$ = 0.45, C^*_{tendon} = 0.55. The tendon edge rests on payload-to-weight and compliance, outweighing the latency and reliability deficits — a margin small enough that Section 4.3 must adjudicate whether it means anything.

Figure 3 displays the three rankings. Two reporting disciplines are visible in it: every pair sums to one, as the Section 3.3 identity requires, and the decision-relevant quantity in each profile is the sign of C* − 0.5, not the coefficient's absolute size.

Figure 3. Recomputed TOPSIS closeness coefficients by task profile (Eqs. 7–10; Table 2 inputs).



Note: Dashed line: the C* = 0.5 indifference locus implied by the two-alternative identity.

4.3 Monte Carlo robustness audit

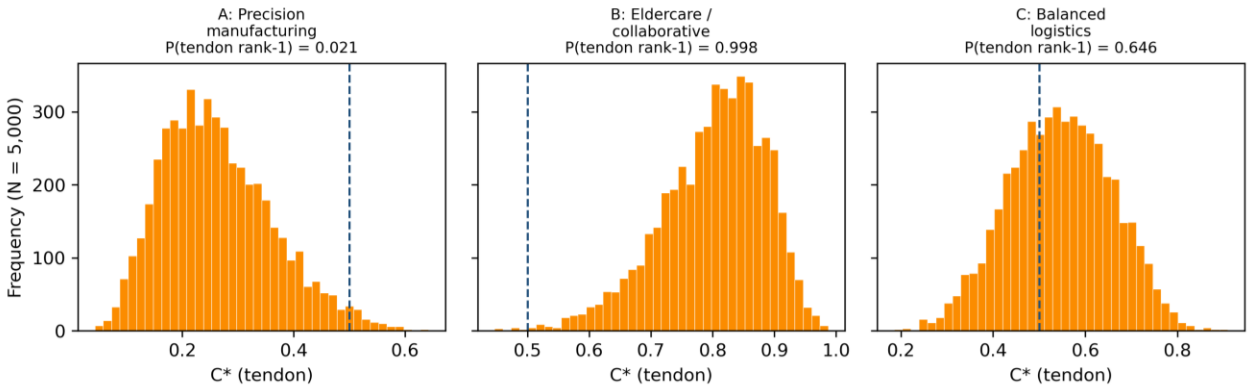
Figure 4 and Table 3 report the Section 3.4 audit (N = 5,000 Dirichlet(25·w) draws per profile). Profile A is robust: tendon actuation reaches rank 1 in only 2.1% of draws (mean C^*_T = 0.26, 95% interval [0.10, 0.49]), so the linkage recommendation survives any plausible re-weighting of a manufacturing board's judgments. Profile B is equally robust in the opposite direction (tendon rank-1 probability 0.998; mean C^*_T = 0.80, 95% interval [0.61, 0.94]): no reasonable eldercare weighting rescues the rigid hand. Profile C, by contrast, is genuinely contested — tendon wins 64.6% of draws, but the 95% interval [0.33, 0.75] straddles 0.5 widely, and linkage takes the remaining 35.4%. The honest deliverable for balanced deployments is therefore not 'tendon is preferred' but a conditional rule, which the next subsection supplies.

Table 3. Deterministic TOPSIS results and Monte Carlo audit (N = 5,000, Dirichlet κ = 25).

Profile	w ($x_1 \ldots x_5$)	$C^*_{linkage}$	C^*_{tendon}	Mean C^*_T (MC)	95% interval	P(tendon rank-1)
A Manufacturing	.10/.35/.10/.35/.10	0.74	0.26	0.26	[0.10, 0.49]	0.021
B Eldercare	.10/.10/.40/.10/.30	0.19	0.81	0.80	[0.61, 0.94]	0.998
C Balanced	.20/.20/.20/.20/.20	0.45	0.55	0.55	[0.33, 0.75]	0.646

Note: Deterministic C* columns satisfy $C^*_L + C^*_T = 1$ per the Section 3.3 identity.

Figure 4. Monte Carlo distribution of tendon closeness C^*_T under Dirichlet(25·w) weight uncertainty



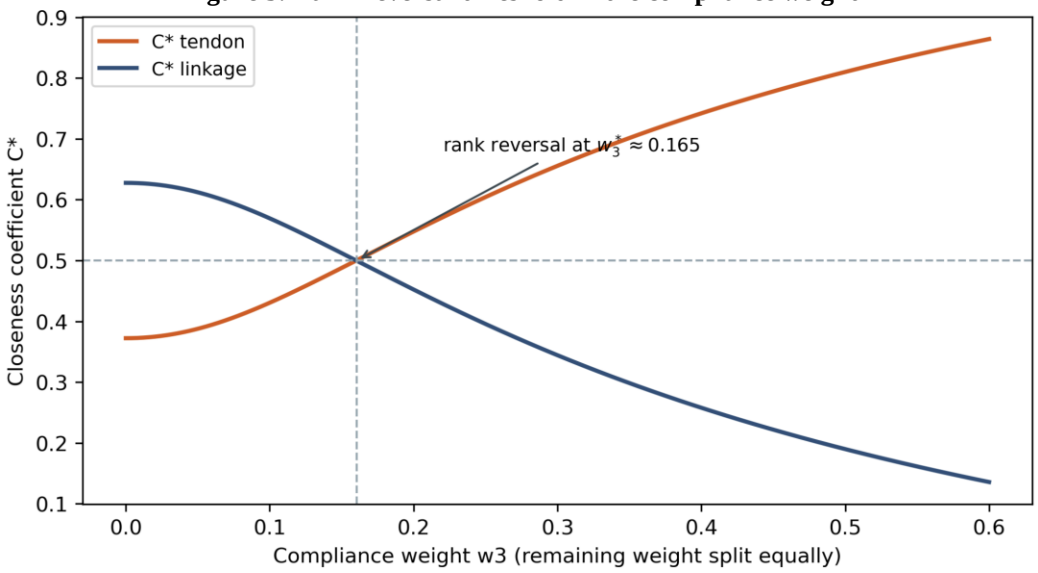
Note: N = 5,000 per profile. Dashed line: C* = 0.5 rank boundary. Profiles A and B sit almost entirely on one side of the boundary; Profile C straddles it.

4.4 The compliance switching threshold

The contested balanced case admits an exact resolution. Holding the non-compliance weights in equal proportion and sweeping the compliance weight w_3 from 0 to 0.6, the ranking reverses at $w^*_3 \approx 0.165$ (Figure 5): below it, latency and reliability carry the linkage architecture; above it, compliance (with the correlated override-safety advantage) carries the tendon architecture. This single number is the framework's most portable output for practitioners: a design board need not agree on a full weight vector, only on whether safe physical compliance deserves more or less than roughly one-sixth of the total decision weight in its deployment regime. Manufacturing boards that answer 'less' should buy linkage; care and collaboration boards that answer 'more' should buy tendon; boards that cannot answer should

treat the choice as contested and resolve it with task-level data rather than committee preference. The threshold is conditional on Table 2 and on the equal-residual sweep; it moves if the parameter matrix is re-measured, which is an argument for the measurement program of Section 6, not against reporting the threshold.

Figure 5. Rank-reversal threshold in the compliance weight



Note: Rank-reversal threshold in the compliance weight w_3 (residual weight split equally). C^*_{tendon} crosses 0.5 at $w^*_3 \approx 0.165$.

4.5 Reading the results jointly

The three analyses triangulate a single conclusion. Deterministically, architecture preference is profile-contingent; stochastically, the two specialised profiles are robust, and the balanced profile is not; parametrically, the contingency collapses to one interpretable threshold. This is the strongest form of answer a two-alternative MADM study can honestly give: not a universal winner, but a decision rule with stated conditions and quantified confidence (Bansal et al., 2021, make the analogous point for AI teammates: accuracy or performance on any single axis does not determine team-level usefulness).

5. Discussion and Risk-Cybernetic Governance

By mapping S_{HACDU} directly to the override safety threshold, our framework demonstrates that end-effector mechanics cannot be divorced from cybernetic governance. A transmission with low S_{HACDU} (high reflected stiffness) acts as an amplifier of belief divergence: late override interventions inject unattenuated peak torques back into the operator, forcing persistent $\delta_t = -1$transitions and permanently depressing collaborative reliance α_t. Conversely, tendon compliance provides passive mechanical low-pass filtering, expanding the permissible latency margin for Eq. (5) without risking operator injury or operational disuse.

Conversely, where belief disagreement is near zero — structured cells, fixtured parts, supervised automation — the tendon architecture's decision-layer benefits are never exercised, while its costs (hysteresis modelling burden c(a) in Eq. (2), 3,200-hour MTBF, re-tensioning maintenance) are paid

continuously. The framework thus rationalises observed industry practice: deterministic manufacturing has standardised on rigid actuation, while contact-rich care and research platforms migrate toward compliant tendon designs (Piazza et al., 2019), and it does so with auditable numbers rather than folklore. It also issues a warning against a common procurement error: extrapolating a Profile A ranking into a Profile B deployment. The Monte Carlo audit shows the two regimes are separated by a wide, robust margin — a ranking imported across regimes is not approximately right, it is wrong with high probability.

Three design implications follow. First, override safety should be specified, tested, and procured as a quantitative attribute (the Section 3.1(v) protocol) rather than asserted as a qualitative virtue; S_{HACDU} makes the HACDU layer contractible. Second, hybrid and variable-stiffness transmissions — which attempt to move a single hand across the Figure 5 threshold by modulating effective compliance — should be evaluated as threshold-shifting technologies inside this framework, not as exempt novelties. Third, because αt dynamics (Eq. 4) couple operator trust to hardware behaviour, deployment programmes should log override events and ratification outcomes alongside maintenance records: the same event stream that estimates δt also re-estimates $\mathcal{X}_4$ and $\mathcal{X}_5$, letting the decision framework be re-run on field data rather than laboratory syntheses. Hardware selection then becomes a governed, revisable decision process — the practical meaning of risk-cybernetic governance for end-effector programmes.

6. Limitations

Four limitations bound the claims. First, the two-alternative design: the identity $C^*_L + C^*_T \equiv 1$ makes closeness purely relative, and TOPSIS rankings can reverse if a third architecture (e.g., a hybrid variable-stiffness hand) enters the set (García-Cascales & Lamata, 2012); the framework extends to $m > 2$ mechanically, but every number in Section 4 would change. Second, Table 2 synthesizes representative platform benchmarks rather than multi-unit empirical distributions. While our Dirichlet audit thoroughly explores the decision-weight manifold (N=5,000), it treats the engineering parameters deterministically. In physical deployments where cable creep reduces stiffness over operating cycles or mechanical play degrades linkage precision, $\mathcal{X}_1$ through $\mathcal{X}_5$ are stochastic processes. Deploying the Dual-Layer Monte Carlo formulation outlined in Section 3.4 — simultaneously perturbing physical hardware distributions $X_{ij} \sim D_j$ alongside preference weights $w^{(n)}$ — represents the necessary next phase to ensure that ultra-high confidence rankings (e.g., P=0.998 in Profile B) remain resilient against real-world component degradation. Third, the AHP judgments of Table 1 come from a single illustrative board; multi-board elicitation with inter-rater analysis would strengthen the weight basis. Fourth, the HACDU coupling (Eqs. 3–5) is modelled, not yet identified from human-subject interaction data; estimating η and $\gamma_{override}$ from override-event logs is required before the decision-layer claims can be called empirical. None of these limitations is hidden by the framework: each corresponds to a stated, fixable input.

7. Conclusion

This paper has presented a unified multi-attribute decision framework for linkage- versus tendon-driven actuation in humanoid end-effectors, coupling AHP-weighted TOPSIS ranking to the HACDU human–AI decision layer. The framework's auditable outputs — recomputed closeness coefficients satisfying the two-alternative identity, Monte Carlo rank-1 probabilities, and a compliance switching threshold at $w^*_3 \approx 0.165$ — convert actuation selection from committee judgment into a governed decision process whose every number can be regenerated and challenged. Linkage actuation is the robust choice for deterministic, precision, maintenance-sensitive regimes; tendon actuation is the robust choice for contact-rich collaborative regimes; and balanced deployments should be decided by the threshold question — how much decision weight safe compliance deserves — rather than by assertion. Future work will replace the synthesised parameter matrix with bench measurements of both architectures on an identical palm platform, estimate the HACDU reliance parameters from human-subject collaborative assembly trials, and extend the candidate set to hybrid variable-stiffness transmissions.

References

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork. Proceedings of the AAAI Conference on Artificial Intelligence, 35(13), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359

Belton, V., & Gear, T. (1983). On a short-coming of Saaty's method of analytic hierarchies. Omega, 11(3), 228–230. https://doi.org/10.1016/0305-0483(83)90047-6

Bicchi, A. (2000). Hands for dexterous manipulation and robust grasping: A difficult road toward simplicity. IEEE Transactions on Robotics and Automation, 16(6), 652–662. https://doi.org/10.1109/70.897777

Catalano, M. G., Grioli, G., Farnioli, E., Serio, A., Piazza, C., & Bicchi, A. (2014). Adaptive synergies for the design and control of the Pisa/IIT SoftHand. The International Journal of Robotics Research, 33(5), 768–782. https://doi.org/10.1177/0278364913518998

Chan, L. C. W. J., & Pan, S.-H. (2026). Human–AI collaboration under uncertainty: A unified decision-science framework and open benchmark suite (HACDU). Advances in Decision Sciences. Advance online publication.

Controzzi, M., Cipriani, C., & Carrozza, M. C. (2014). Design of artificial hands: A review. In R. Balasubramanian & V. J. Santos (Eds.), The human hand as an inspiration for robot hand development (Springer Tracts in Advanced Robotics, Vol. 95, pp. 219–246). Springer. https://doi.org/10.1007/978-3-319-03017-3_11

Dollar, A. M., & Howe, R. D. (2010). The highly adaptive SDM hand: Design and performance evaluation. The International Journal of Robotics Research, 29(5), 585–597. https://doi.org/10.1177/0278364909360852

García-Cascales, M. S., & Lamata, M. T. (2012). On rank reversal and TOPSIS method. Mathematical and Computer Modelling, 56(5–6), 123–132. https://doi.org/10.1016/j.mcm.2011.12.022

Haddadin, S., Albu-Schäffer, A., & Hirzinger, G. (2009). Requirements for safe robots: Measurements, analysis and new insights. The International Journal of Robotics Research, 28(11), 1507–1527. https://doi.org/10.1177/0278364909343970

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517–527. https://doi.org/10.1177/0018720811417254

Hwang, C.-L., & Yoon, K. (1981). Multiple attribute decision making: Methods and applications. Springer. https://doi.org/10.1007/978-3-642-48318-9

ISO/TS 15066:2016. Robots and robotic devices — Collaborative robots. International Organization for Standardization.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial Intelligence, 101(1–2), 99–134. https://doi.org/10.1016/S0004-3702(98)00023-X

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge University Press. https://doi.org/10.1017/CBO9780511809477

Keeney, R. L., & Raiffa, H. (1976). Decisions with multiple objectives: Preferences and value tradeoffs. John Wiley & Sons.

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Piazza, C., Grioli, G., Catalano, M. G., & Bicchi, A. (2019). A century of robotic hands. Annual Review of Control, Robotics, and Autonomous Systems, 2, 1–32. https://doi.org/10.1146/annurev-control-060117-105003

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology, 15(3), 234–281. https://doi.org/10.1016/0022-2496(77)90033-5

Salisbury, J. K., & Craig, J. J. (1982). Articulated hands: Force control and kinematic issues. The International Journal of Robotics Research, 1(1), 4–17. https://doi.org/10.1177/027836498200100102

Savage, L. J. (1972). The foundations of statistics (2nd ed.). Dover Publications.

Simon, H. A. (1957). Models of man, social and rational. John Wiley & Sons.

von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.

Appendix A. Notations

Table A1. Mathematical notations are used throughout the paper.

Symbol	Definition
$a \in \mathcal{A}$	Candidate architecture, $a \in \{a_{linkage}, a_{tendon}\}$; $\mathcal{A}_{feasible} \subseteq \mathcal{A}$ after bounded-rationality screening (Eq. 2).
$\theta \in \Theta$	Latent environmental task state: load, obstacle, and contact uncertainty.
$u(a, \theta)$, $U(a)$	Multi-attribute utility of a in state θ, and its expectation under $P(\theta)$ (Eq. 1).
$c(a)$, λ	Modelling/computational cost of architecture a; λ converts cost to utility units (Eq. 2).
$bHt(\theta)$, $bRt(\theta)$, $bCt(\theta)$	Human, robot, and fused collaborative belief distributions at step t (Eq. 3).
αt, η, δt	Human reliance weight; learning rate; signed interaction outcome signal (Eqs. 3, 4).
$D_{KL}(\cdot\|\cdot)$, $\gamma_{override}$	Kullback–Leibler divergence (human belief as reference); override threshold (Eq. 5).
$\mathcal{X}_1 \ldots \mathcal{X}_5$	Payload-to-weight; responsiveness (cost); compliance; reliability (MTBF); override safety S_{HACDU}.
w, A, λ_{max}, CI, CR	Weight vector ($\Sigma w_j = 1$); AHP judgment matrix; principal eigenvalue; consistency index/ratio (Eq. 6).
x_{ij}, r_{ij}, v_{ij}	Raw, vector-normalised, and weighted-normalised performance of candidate i on attribute j (Eqs. 7–8).
A+, A−, d+, d−	Ideal / anti-ideal solutions and separation distances (Eqs. 8–9).
C^*_i	TOPSIS relative closeness of candidate i; $C^*_L + C^*_T \equiv 1$ when m = 2 (Eq. 10).
κ, N, $\hat{P}$	Dirichlet concentration (κ = 25); Monte Carlo draws (N = 5,000); rank-1 probability estimator (Eq. 11).
w^*_3	Compliance-weight rank-reversal threshold, ≈ 0.165 under equal residual weights (Section 4.4).

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, …, T with latent state θ_t. All notation is defined in Table 1 (Appendix A). Human and AI maintain belief distributions:

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

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

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

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 γ^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 α_e, α_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 c-ELBO	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 trust estimator	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

$$b_t^C = \varphi(b_t^H, b_t^A, r_t, e_t), \tag{8}$$

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

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

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

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

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 (St, $\theta S>0$) 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 $r_t \sim Beta(a_t, b_t)$, 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 $Var[r_t]$ are analytically tractable:

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

In the operational stack, the point estimate $E[r_t]$ drives utility weighting (Eq. 17) and override boundaries (Eq. 33), while the variance $Var[r_t]$ 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:

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

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

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

where $c_{exec}(a, I_t)$ 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):

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

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

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 $\Theta_1 = \mathcal{E}^A \backslash \mathcal{E}^H$ and $\Theta_2 = \mathcal{E}^H \backslash \mathcal{E}^A$ 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

$$b_t^C(\theta) \propto b_t^H(\theta) \cdot b_t^A(\theta), \tag{22}$$

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

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

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

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:

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

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

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 $d \in [0, dmax\ ]$ 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:

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

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 $[0, dmax\ ]$. 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 $[0, dmax\ ]$ via the Euclidean projection clip(·) completes the proof.

In accordance with Cognitive Load Theory (Miller, 2019; Sweller, 2011), while generation latency scales linearly ($c_{gen}(d)=c_1 d$), cognitive processing burden exhibits superlinear escalation due to working-memory limitations ($c_{cog}(d) = \frac{1}{2} c_2 d^2$, 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 $[0, d_{max}]$ yields the unique, closed-form optimal explanation policy:

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

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

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

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

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:

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

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

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

To ensure behavioral and decision-theoretic validity, the threshold function g(•) 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:

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

where $\tau_0 \in (0,1)$ is the baseline tolerance, and θ_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:

$$p_{esc}(I_t) = \sigma(\nu_0 + \nu_1 D_{KL} + \nu_2 S_t - \nu_3 Q_t). \tag{34}$$

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 ($\nu_1>0$) and consequence severity ($\nu_2>0$), but is mitigated by expected collaborative decision quality ($\nu_3>0$).

7.2 Arbitration Policy

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

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

$$a^{arb} = arg\ max_{a \in A}\, \pi_s(a \mid I_t). \tag{37}$$

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):

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

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):

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

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},

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

and, when calibration and test points are exchangeable,

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

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

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

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}:

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

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):

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

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

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:

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

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 ≥ 1 − α_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,

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

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

$$\hat{\psi}_t = \sigma(\vartheta_0 + \boldsymbol{\vartheta}^\top f_t), \tag{48}$$

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

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

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,

$$\hat{\boldsymbol{\vartheta}} = \operatorname{argmin}_{\boldsymbol{\vartheta}} \sum_{m\in\mathcal{M}_{SR}} \left(y_m^{SR} - \sigma\left(\vartheta_0 + \boldsymbol{\vartheta}^\top f_{t_m}\right)\right)^2 + \lambda_\vartheta \parallel \boldsymbol{\vartheta} \parallel_2^2, \tag{50}$$

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 $\mathrm{Var}[r_t]$ (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 $\omega_1 \ldots \omega_7$ 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.

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

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

$$ARR = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{|r_i - r_i^*| \le \varepsilon\}, \tag{53}$$

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 $ARR + OTR + UTR = 1$.

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

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

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

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

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

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 $\omega_1 \ldots \omega_7$.

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 $\theta \in$ {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)=0.42×0.62 + 0.58×0.55 − 0.02 = 0.56,

Q^C(a2)=0.42×0.74 + 0.58×0.68 − 0.04 = 0.67, and Q^C(a3)=0.42×0.49 + 0.58×0.71 − 0.07 = 0.55. The collaborative recommendation is thus a2, 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 $a^H = \text{argmax}_a Q^H(a) = a_2$. (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 ≈ 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 ε = 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 $DE = 1 + (K - 1)\rho_{ICC}$:

$$Y_{isc} = \beta_0 + \beta_c + u_i + v_s + \varepsilon_{isc}, \tag{59}$$

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

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

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 ≈ 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, θ) 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 $\mathbb{E}[r_t]$ 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

($c_{cog}(d) = \frac{1}{2}c_2 d^2$) 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

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

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

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
θ_Q, θ_A, θ_S, θ_f, θ_C	Weights	Sensitivity coefficients for trust innovation components
e_t	Escalation	Binary escalation state
δ_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 ($w^H + w^A = 1$)
πc, 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)
$\omega_1 \ldots \omega_7$	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)
$L_{c\text{-}ELBO}$	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)
$\mathscr{E}^H$, $\mathscr{E}^A$	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)
$\hat{q}$, α_cp, $\mathcal{C}(x)$	Conformal	Conformal quantile, miscoverage level, and prediction set
$\tilde{\sigma}^A$, κ_cp	Calibration	Calibrated AI standard deviation and conformal scale factor

$ECE_t^{(h)}$, ε_cal, θ_c	Calibration	Windowed calibration error, refit tolerance, and threshold sensitivity
x_k, f_t, h	Telemetry	Indicator vector, windowed features, and window length
$\hat{\psi}_t$, $\bar{\psi}_t$, κ_tel, ϑ	Telemetry	Observed and fused trust signals, fusion weight, and measurement weights

Appendix B. Implementation Checklist

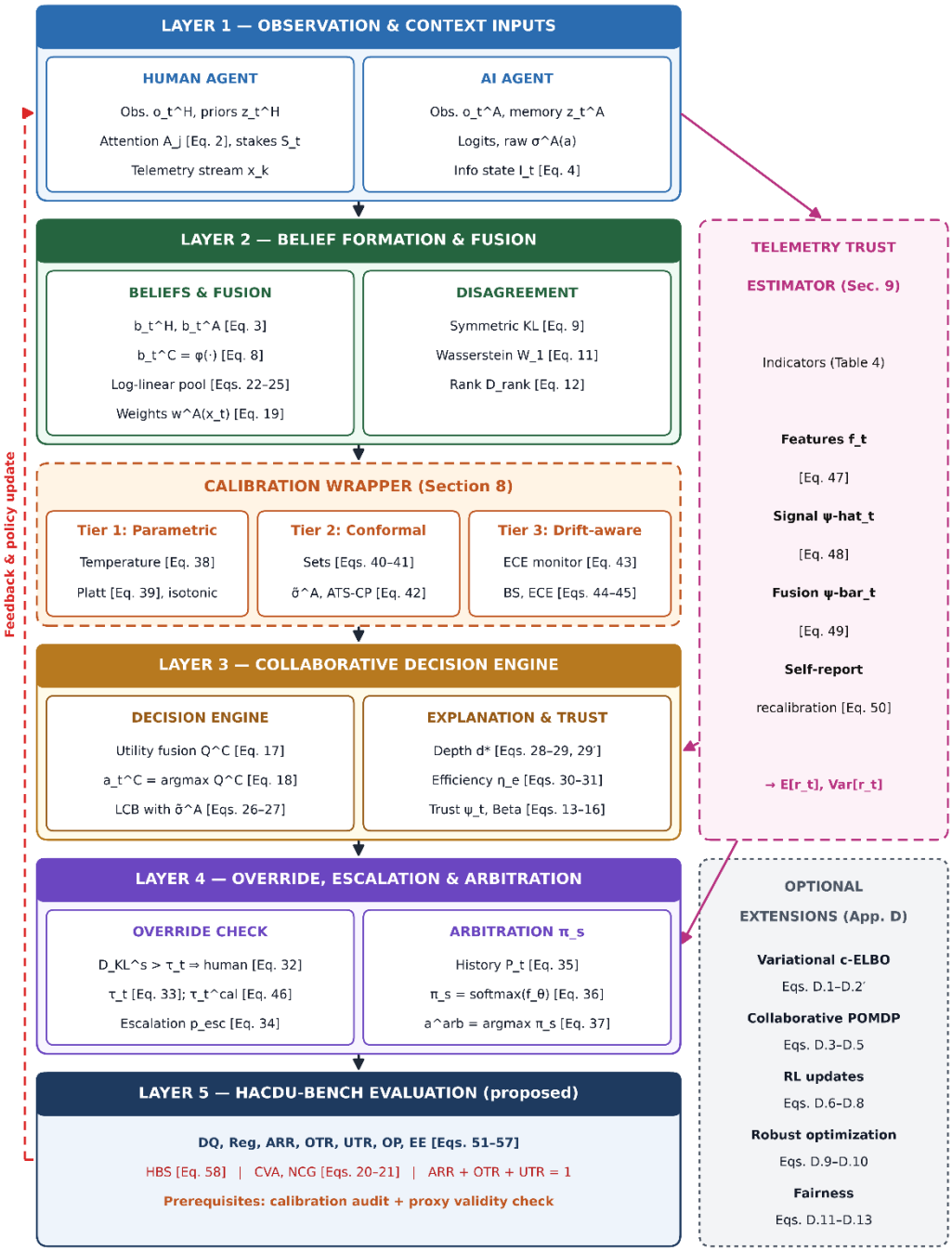
The checklist is ordered by logical dependency: each step uses the outputs of the steps before it. In particular, the override threshold and escalation rules (step 8) are parameterized only after the calibration wrapper (step 6) and the reliance estimator (step 7) are in place, because the calibration-aware 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, θ).
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 $\omega_1 \ldots \omega_7$, 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 $b_t^C(\theta)$:

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

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

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

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

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

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:

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

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

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

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

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

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):

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

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

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

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

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

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

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

$$\Delta_u = max_{g \in G} J_g - min_{g \in G} J_g, \tag{D.11}$$

$$\Delta_r = max_{g \in G} r_g - min_{g \in G} r_g, \tag{D.12}$$

$$J_C{}^{fair} = J_C - \lambda_u \Delta_u - \lambda_r \Delta_r. \tag{D.13}$$

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).

Eliminating Transboundary Haze via Agentic AI and Microalgae Bio-Refineries: The Eagor-Core Framework for Palm Oil Estate Transformation

Lanz CWJ Chan
Finamatrix Quant-Lab, Singapore
E-mail: 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 15, 2026

Abstract

Transboundary haze in Southeast Asia persists because fire remains the cheapest way to clear and prepare land on tropical peat, while the health and climate costs of burning fall on others. This short paper asks whether an alternative revenue stream from existing palm oil mill infrastructure, verified and paid for through an automated monitoring system, could change that private calculus. We present Eagor-Core, a conceptual agentic artificial intelligence (AI) framework built on the Energy–AI Governance Orchestration (EAGOR) architecture. Eagor-Core converts palm oil mill effluent (POME) lagoons into microalgae cultivation units and links their output to zero-burn payments verified through a three-tier monitoring pipeline of in-ground sensors, autonomous drones, and satellite Earth observation. At the physical layer, we formalize multi-pond control as a decentralized partially observable Markov decision process (Dec-POMDP) with explicit sensor-drift and fouling states, and propose multi-agent reinforcement learning (MARL) under centralized training with decentralized execution. A bounded, volume-weighted health index (the Finamatrix vGRE) modulates a normalized multi-objective reward. At the governance layer, adaptive conformal inference provides prediction intervals whose width determines whether the system acts autonomously, falls back to a conservative policy, or enters a safe state and requests a signed human-in-the-loop (HITL) decision. We derive the minimum conditional payment that makes zero-burn compliance privately rational, and extend it to independent smallholders, who manage about 40% of Indonesia's oil palm area, through a mill-level cooperative with conditional dividends. We also provide an assumption-based techno-economic scoping estimate for a 10-hectare estate, an evaluation protocol, and a three-stage research and rollout agenda. We report no simulation results, field validation, or multi-seed experiments; the techno-economic values are assumption-based scoping estimates intended to guide future work.

Keywords: agentic AI; Energy–AI Governance Orchestration (EAGOR); microalgae; palm oil mill effluent; Dec-POMDP; multi-agent reinforcement learning

JEL Codes: Q13; Q16; Q42; Q53; Q56; O33; C61; C63

1 Introduction

1.1 The transboundary haze problem

Seasonal transboundary haze across Southeast Asia is a recurring socio-economic and ecological crisis linked to fire-assisted land clearing and drainage on tropical peatlands in Sumatra and Kalimantan (Koh et al., 2011; Tacconi, 2016). Fire carbon emissions over maritime Southeast Asia in 2015 were the largest since 1997 (Huijnen et al., 2016), and exposure to the resulting fine particulate matter (PM2.5) has been associated with substantial excess mortality across Indonesia, Malaysia, and Singapore (Marlier et al., 2013; Koplitz et al., 2016). The World Bank estimated the economic cost of Indonesia's 2015 fire season at about USD 16 billion (Glauber et al., 2016).

The legal framework is not absent. Indonesia's Law No. 32 of 2009 on Environmental Protection and Management (Republic of Indonesia, 2009) prohibits land clearing by burning, and the ASEAN Agreement on Transboundary Haze Pollution (ASEAN, 2002) commits member states to prevention, monitoring, and joint response. Enforcement is nonetheless hampered by the geographic extent of remote plantations, institutional lag, contested land tenure, and patronage relationships (Varkkey, 2013; Tacconi, 2016). Most importantly, fire remains cheaper than mechanical land preparation. In the language of the economics of crime (Becker, 1968), burning is privately rational whenever its cost advantage exceeds the expected penalty, and in remote estates, both the detection probability and the realized penalty are low. Information gaps compound the problem: the production and sharing of fire and haze data have been contested among states, which has weakened multilateral governance of the haze as an environmental health risk (Berti Suman, 2022).

1.2 A circular alternative based on POME

Palm oil mill effluent (POME) is a high-volume, organic-rich wastewater produced during crude palm oil (CPO) extraction; roughly 2.5–3.5 t of POME is generated per tonne of CPO (Bagchi et al., 2021; Low et al., 2021). Untreated POME has very high biochemical and chemical oxygen demand and is a major source of methane when held in open anaerobic lagoons (Wu et al., 2010; Lam & Lee, 2011). After appropriate pre-treatment and dilution, however, POME can serve as a nutrient source for microalgae such as *Chlorella* and *Scenedesmus*, which remove nitrogen, phosphorus, and organic load while producing lipid-rich biomass (Cheah et al., 2018; Hariz et al., 2019). Boiler flue gas can supply inorganic carbon to the cultures. The resulting crude algae oil (CAO) can be refined into a lipid feedstock for hydroprocessed esters and fatty acids (HEFA) jet fuel, and the defatted residue can be sold as protein meal.

Area-based productivity comparisons between microalgae and oil palm are highly sensitive to assumptions. Early optimistic estimates placed algal oil yields roughly an order of magnitude above oil palm (Chisti, 2007), whereas techno-economic and resource assessments of outdoor open ponds report considerably lower productivities and highlight cost, harvesting, and nutrient-supply constraints (Davis et al., 2011; Chisti, 2013; Quinn & Davis, 2015). We therefore use productivity figures derived from our own simulation and treat literature yields as bounds rather than targets.

1.3 Benchmarking existing haze-mitigation strategies

Table 1 compares the governance mechanisms and principal failure modes of four existing approaches with the mechanism proposed here. Legal prohibition fails mainly because detection is costly in remote peatlands while fire remains considerably cheaper than mechanical clearing (Tacconi, 2016). Voluntary certification under the Roundtable on Sustainable Palm Oil (RSPO) reduced deforestation on certified plantations but had no detectable effect on fire on peat (Carlson et al., 2018), and its audit costs are difficult for independent smallholders to bear (Jelsma et al., 2017). Crop switching imposes years of forgone income before new crops yield, and greenfield algae farms carry high capital costs and biological fragility (Chisti, 2013; Quinn & Davis, 2015).

Table 1. Benchmark of haze-mitigation strategies

Strategy	Governance mechanism	Primary failure mode	Eagor-Core response
Legal bans and fines (Law No. 32/2009)	Prosecution, land seizure, and administrative fines	Enforcement lag: remote peatlands make detection costly; fire remains cheaper than mechanical clearing	Raises detection probability through satellite MRV and adds a revenue stream that is forfeited if burning is detected
RSPO / ESG certification	Voluntary certification and supply-chain audits	Leakage of non-certified CPO to other markets; audit costs exclude smallholders; limited effect on fire	Mill-level aggregation lets smallholders participate without individual certification
Alternative crop switching	Conversion of peatland to non-palm crops	High transition cost and several years before the first harvest	Retrofits existing mill lagoons; first harvests within weeks of inoculation
Greenfield algae farms	Closed photobioreactors or new open ponds	High CAPEX, culture crashes, probe fouling, SO_2 corrosion, high energy use	Lagoon retrofit with fault-tolerant MARL control and conformal deferral

Note. RSPO = Roundtable on Sustainable Palm Oil; ESG = environmental, social, and governance; MRV = measurement, reporting, and verification. The Eagor-Core column states design intentions; its performance is evaluated only in simulation in this paper.

1.4 The control and governance gap

Large-scale microalgae cultivation in open raceway ponds (ORPs) is fragile. Cultures are exposed to weather volatility, seasonal irradiance changes, grazing ciliates, viral and bacterial contamination, invasion by wild strains, and pH or dissolved-oxygen (DO) excursions (Day et al., 2012; Wang et al., 2013). Physical infrastructure adds a second layer of risk: optical and electrochemical probes foul and drift, paddlewheels fail, flue-gas lines corrode in the presence of sulfur dioxide (SO_2) and acidic condensate, and remote sites lose grid power and connectivity. Static rulebooks and single-loop proportional–integral–derivative (PID) controllers handle setpoint regulation well but are not designed for coupled, multi-objective decisions across many interconnected ponds under large and shifting uncertainty.

Reinforcement learning (RL) has been applied successfully to bioprocess control problems such as batch optimization, co-culture regulation, and fed-batch cyanobacterial production (Petsagkourakis et al., 2020; Treloar et al., 2020; Ma et al., 2020). These studies, however, treat a single reactor, assume reliable sensing, and do not connect process control to environmental verification or payment. Conversely, the literature on payments for ecosystem services (PES) and carbon market design emphasizes that payments only change behavior when they are conditional, additional, and reliably

verified (Engel et al., 2008; Jack et al., 2008; Schneider et al., 2015). No existing framework, to our knowledge, couples a fault-tolerant multi-pond controller with satellite-based zero-burn verification and an explicit rule for when automated decisions should be deferred to humans.

1.5 Contributions

This paper makes six contributions, all conceptual or methodological:

1. **Conceptual architecture and conversion manual.** We describe Eagor-Core, an agentic AI framework grounded in the EAGOR governance architecture (Chan & Wong, 2026a), with offline edge execution, automated sensor recalibration, and multi-tier fault recovery, together with a conversion manual for existing POME lagoons; the code is to be released under an open-source license (see Declarations).

2. **Dec-POMDP formulation with explicit sensor and fouling states.** We formulate multi-pond control as a Dec-POMDP whose state includes stochastic probe drift and a fouling process that is reset by backwashing. The observation function fuses three telemetry tiers: in-ground sensors, autonomous drones, and satellites.

3. **A bounded vGRE-modulated reward.** We redefine the Finamatrix Volume-Weighted Golden Ratio Estimator (vGRE) as a rolling, bounded health-weighted productivity index, and use golden-ratio weights as a convex default for the biomass–energy trade-off, subject to sensitivity analysis.

4. **Calibrated deferral under distribution shift.** We replace a static split-conformal rule with adaptive conformal inference over normalized residual scores, and give a precedence-ordered execution rule that is well defined when connectivity is lost.

5. **Incentive model, smallholder cooperative, and prospective techno-economic scoping.** We derive the minimum conditional payment that makes zero-burn compliance privately rational, extend it to a mill-level smallholder cooperative with conditional dividends, and give an assumption-based CAPEX/OPEX and break-even scoping estimate for a 10-hectare estate.

6. **Evaluation protocol and research agenda.** We specify the multi-seed protocol, stronger baselines, ablations, scalability tests, fault-escalation confusion matrix, field pilot, and life-cycle assessment needed before any performance or economic claim can be made.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the incentive model. Section 4 describes the framework and its formalization. Section 5 specifies the proposed evaluation protocol and the hypotheses it is designed to test. Section 6 gives a prospective techno-economic scoping estimate under assumed yields. Section 7 discusses implications, a possible extension to vertical photobioreactors, limitations, and the research agenda, and Section 8 concludes. The appendices contain the field conversion manual, the smallholder cooperative model, the rollout plan, the simulator equations, and a policy brief.

2 Related Work

2.1 Haze, peatland fire, and economic incentives

Remote sensing has documented the large-scale conversion of peat swamp forest to oil palm (Koh et al., 2011), and atmospheric modeling has linked fire activity, El Niño conditions, and population exposure (Marlier et al., 2013; Koplitz et al., 2016). Policy analyses converge on the view that prohibition alone has not been effective. Tacconi (2016) argues that prevention requires changing the economic incentives for using fire, and Varkkey (2013) shows how patronage networks weaken enforcement. Economic theory offers two complementary levers: raising the expected penalty through better detection (Becker, 1968) and rewarding compliance through conditional payments (Engel et al., 2008; Jack et al., 2008). Eagor-Core targets both by improving the probability and speed of detection and by tying a revenue stream to verified non-burning.

2.2 Microalgae cultivation on POME

POME treatment has traditionally relied on ponding systems, with growing interest in biogas capture and resource recovery (Wu et al., 2010; Lam & Lee, 2011; Tan et al., 2009). Laboratory and pilot studies show that microalgae can grow on diluted POME while removing nutrients and accumulating lipids (Cheah et al., 2018), and that indigenous strains can combine wastewater treatment with CO_2 fixation (Hariz et al., 2019). Reviews of algal biofuel production identify harvesting, dewatering, and contamination as the main barriers to commercial viability (Brennan & Owende, 2010; Chisti, 2013), and techno-economic assessments show that results are highly sensitive to areal productivity and lipid content (Davis et al., 2011; Quinn & Davis, 2015). Growth models for outdoor cultures typically combine light, temperature, and nutrient limitation terms (Béchet et al., 2013), and these models form the basis of the simulator specified here. Recent reviews confirm that large-scale microalgae cultivation on POME and other wastewaters is technically feasible, while identifying dilution, pre-treatment, and contamination control as the main obstacles to scale (Bagchi et al., 2021; Low et al., 2021).

2.3 Reinforcement learning for bioprocess control

Deep RL achieved human-level control in complex sequential tasks (Mnih et al., 2015), and its application to bioprocesses has grown rapidly. Petsagkourakis et al. (2020) used policy gradients for batch bioprocess optimization under plant–model mismatch, Treloar et al. (2020) regulated microbial co-cultures in chemostats, and Ma et al. (2020) combined neural network simulators with deep RL for fed-batch cyanobacterial production. Most recently, Gil et al. (2026) trained multi-agent DDPG controllers offline on historical expert-system data to regulate pH and dissolved oxygen jointly, and validated them in an industrial-scale open photobioreactor. Decentralized control of multiple coupled units is formally a Dec-POMDP, which is computationally hard to solve exactly (Bernstein et al., 2002; Oliehoek & Amato, 2016). Practical MARL methods therefore adopt centralized training with decentralized execution (CTDE), using a centralized critic during training and independent actors at run time (Lowe et al., 2017; Rashid et al., 2018; Yu et al., 2022; Zhang et al., 2021). Multi-objective rewards require care: scalarization weights embody value judgments (Roijers et al., 2013), and reward shaping can alter the optimal policy unless it is potential-based (Ng et al., 1999). Safe RL

methods add explicit constraints or fallback controllers when exploration or model error could cause damage (García & Fernández, 2015).

2.4 Uncertainty quantification and human oversight

Conformal prediction produces prediction sets with finite-sample, distribution-free coverage guarantees under exchangeability (Vovk et al., 2022; Lei et al., 2018; Angelopoulos & Bates, 2023). Standard split conformal intervals built on absolute residuals have constant width, which makes them unsuitable as an input-dependent uncertainty signal; locally weighted residuals (Lei et al., 2018) and conformalized quantile regression (Romano et al., 2019) restore adaptivity. Because sensor drift, seasonal change, and contamination violate exchangeability, methods that remain valid under shift are needed. Adaptive conformal inference (ACI) updates the miscoverage level online and guarantees long-run coverage under arbitrary shift (Gibbs & Candès, 2021), and weighted conformal methods bound the coverage gap under non-exchangeability (Barber et al., 2023). Human-in-the-loop machine learning provides the organizational counterpart: humans are consulted where model confidence is low, or stakes are high (Mosqueira-Rey et al., 2023).

2.5 Agentic AI governance and the EAGOR framework

Agentic AI systems increasingly act on physical infrastructure, which raises the question of how attention, escalation, and authority should be divided between machines and people. The EAGOR (Energy–AI Governance Orchestration) framework addresses the recursive dependence between energy infrastructure and the AI systems that control it by structuring attention, escalation, option activation, and adaptive redesign across autonomous agents and human stakeholders (Chan & Wong, 2026a). Related work proposes a framework and benchmark suite for human–AI collaborative decision-making under uncertainty (Chan et al., 2026) and multi-attribute decision frameworks for autonomous agents in bio-industrial bioprocessing (Chan & Pan, 2026). The vGRE used in this paper originated as a volume-weighted, golden-ratio-based overlay for drawdown and tail-risk control in financial portfolios (Chan & Wong, 2016, 2026b); Section 4.3 adapts it to a bounded physical health index. Because these frameworks are recent and partly unpublished, we use them as design principles and evaluate the resulting system on its own terms.

2.6 Verification, double counting, and data governance

Satellite products provide the observational basis for zero-burn verification. The MODIS Collection 6 active fire product (Giglio et al., 2016) and the VIIRS 375 m product (Schroeder et al., 2014) detect active fires, while Sentinel-1 C-band synthetic aperture radar (SAR) observes land-surface change through cloud and smoke (Torres et al., 2012). Credits derived from such verification face the risk of double-counting across registries and jurisdictions (Schneider et al., 2015); blockchain-based registries have been proposed to make issuance and transfer auditable (Franke et al., 2020). Cross-border data flows are governed in Singapore by the Personal Data Protection Act 2012 (Republic of Singapore, 2012) and in Indonesia by Law No. 27 of 2022 on Personal Data Protection (Republic of Indonesia, 2022), whose transition period ended in October 2024. Edge computing allows sensitive data to be processed close to where it is generated, reducing latency and exposure (Shi et al., 2016).

2.7 Smallholders and land-use decisions

Independent smallholders manage about 40% of Indonesia's oil palm area and are a heterogeneous group, ranging from local farmers with a few hectares to absentee investors (Jelsma et al., 2017; Euler et al., 2016). Smallholder expansion is driven by the high returns to oil palm relative to alternative crops (Euler et al., 2016), and smallholders typically lack the capital, legal documentation, and organizational capacity needed for certification (Jelsma et al., 2017). Because they sell fresh fruit bunches (FFB) to mills rather than processing them, the mill is the natural point at which a new revenue stream can be shared with them. Group-based conditional payments can harness social monitoring within communities, but they also raise free-rider and fairness concerns when one member's action affects everyone's payment (Engel et al., 2008).

2.8 Research gap

The literature contains strong components but no integrated system. Bioprocess RL studies stop at the reactor boundary; PES and carbon market studies assume verification without addressing the reliability of the underlying production system; and conformal prediction has rarely been used to govern when an industrial controller should hand control to humans. Eagor-Core addresses this gap by linking the three, and the incentive model in Section 3 makes explicit the economic condition under which the linkage changes land-use behavior.

3 Incentive Model for Zero-Burn Compliance

Consider an estate deciding whether to clear or prepare a parcel by fire (*B*) or mechanically (*NB*). Let c_F and c_M denote the per-hectare costs of fire-assisted and mechanical preparation, p the probability that burning is detected and sanctioned, F the expected fine, s a conditional zero-burn payment received only if verification confirms no burning, and ΔR the net operating margin from algae production, which is likewise forfeited if the estate is found to burn. Burning is privately preferred when

$$c_M - c_F > pF + s + \Delta R. \tag{1}$$

The minimum conditional payment that makes non-burning weakly preferred is therefore

$$s^* = \max\{0,\ (c_M - c_F) - pF - \Delta R\}. \tag{2}$$

Equation (2) has three implications. First, a higher detection probability p, which satellite verification provides, lowers the payment needed for compliance. Second, any positive net margin ΔR from the algae operation substitutes for public payment, which is the economic rationale for coupling the bioprocess to verification. Third, if $\Delta R \geq c_M - c_F$, compliance is self-enforcing and no external payment is required. Conditionality and additionality, the central design principles of PES schemes (Engel et al., 2008; Jack et al., 2008), are enforced through the rolling 30-day verification window described in Section 4.6.

Extension to smallholder cooperatives. Smallholders do not own mills and cannot capture ΔR directly. Under the proposed mill–cooperative structure, the mill operates the algae facility on its lagoon footprint and distributes a share θ of its net algae and credit revenue NR to the smallholders

who supply it, in proportion to FFB deliveries and conditional on zero-burn verification of their registered village geofence *v*. The dividend to smallholder *j* in village *v* in a settlement period is

$$d_j = \theta \cdot NR \cdot \frac{FFB_j}{\sum_k FFB_k} \cdot g(\beta_v), \quad g(\beta_v) = \max\left\{0, 1 - \frac{\beta_v}{\bar{\beta}}\right\}. \tag{3}$$

Substituting the expected dividend for ΔR in Equation (1), a smallholder refrains from burning when $c_M - c_F \leq pF + \mathbb{E}[d_j]$. Because the condition applies at village level, each smallholder's dividend also depends on neighbours' behaviour, which creates an incentive for mutual monitoring. The same feature exposes compliant farmers to losses caused by others, so Equation (3) uses a graduated rule: β_v is the share of the geofence area with verified burning in the period, confirmed by HITL review, and $\bar{\beta} > 0$ is the burned share at which the village dividend falls to zero. The binary rule 1{no verified fire in geofence *v*} is the limiting case $\bar{\beta} \to 0$.

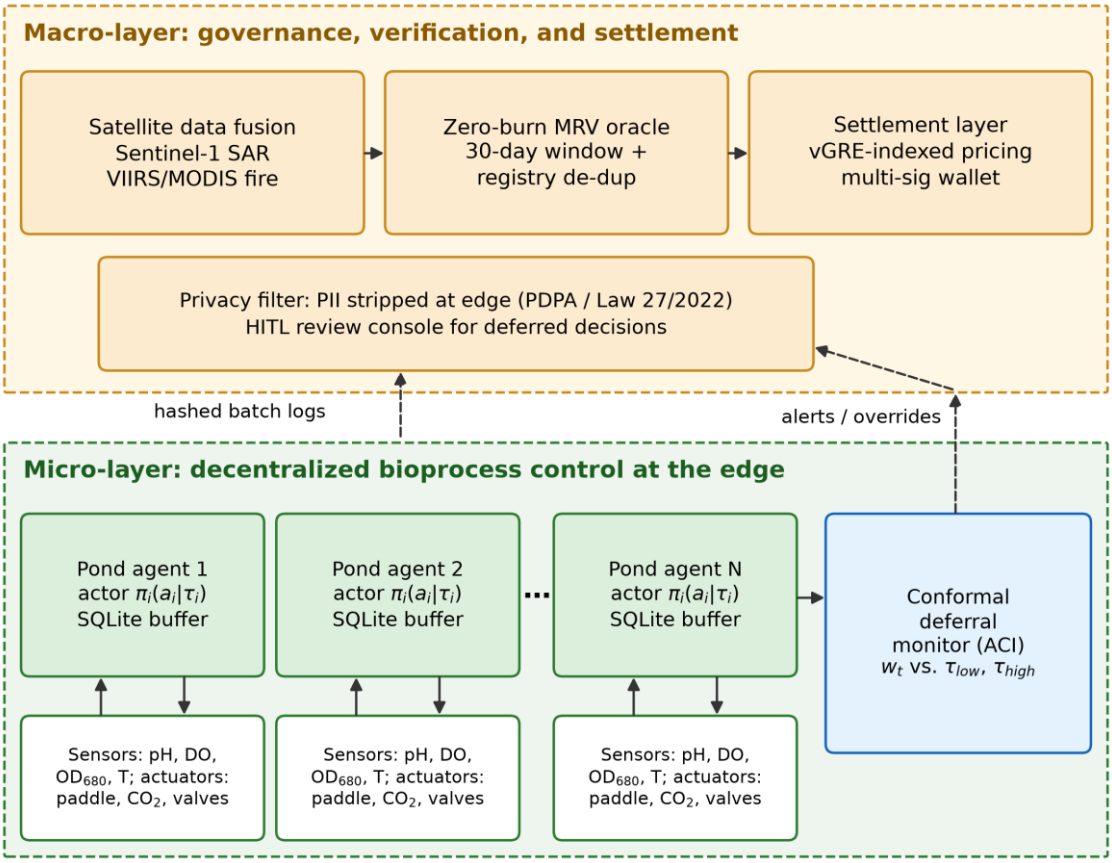
Free-riding, fairness, and review. A village-level condition is a collective incentive. Each member bears only part of the loss caused by a fire but captures the full private benefit of burning, so free-riding becomes more likely as villages grow, while compliant members are penalized for fires they did not set. Two refinements reduce both problems. First, when a drone inspection attributes a verified fire to a registered plot, the reduction is applied to that member's dividend, and $g(\beta_v)$ applies only to the burned area that cannot be attributed. Second, every flag passes a review procedure with fixed time limits: the cooperative is notified within 7 days of a satellite flag, a drone inspection and HITL decision follow within 14 days, and affected members may appeal to a mill–cooperative committee within 30 days, before the dividend is settled. A formal game-theoretic analysis of the resulting monitoring equilibrium is left to future work.

4 Eagor-Core Framework

4.1 Dual-layer topology

Eagor-Core operates on two layers (Figure 1). The **micro-layer** consists of distributed edge controllers, one per pond or pond cluster, that run decentralized policies, keep a local state buffer, and govern paddlewheel speed, CO_2 injection, scrubber bypass valves, effluent dilution and harvesting pumps, as well as automated backwashing and sensor recalibration. The **macro-layer** fuses satellite observations to verify zero-burn compliance across estate boundaries, de-duplicates credit claims across registries, applies privacy filters to all outbound data, hosts the HITL review console, and settles payments.

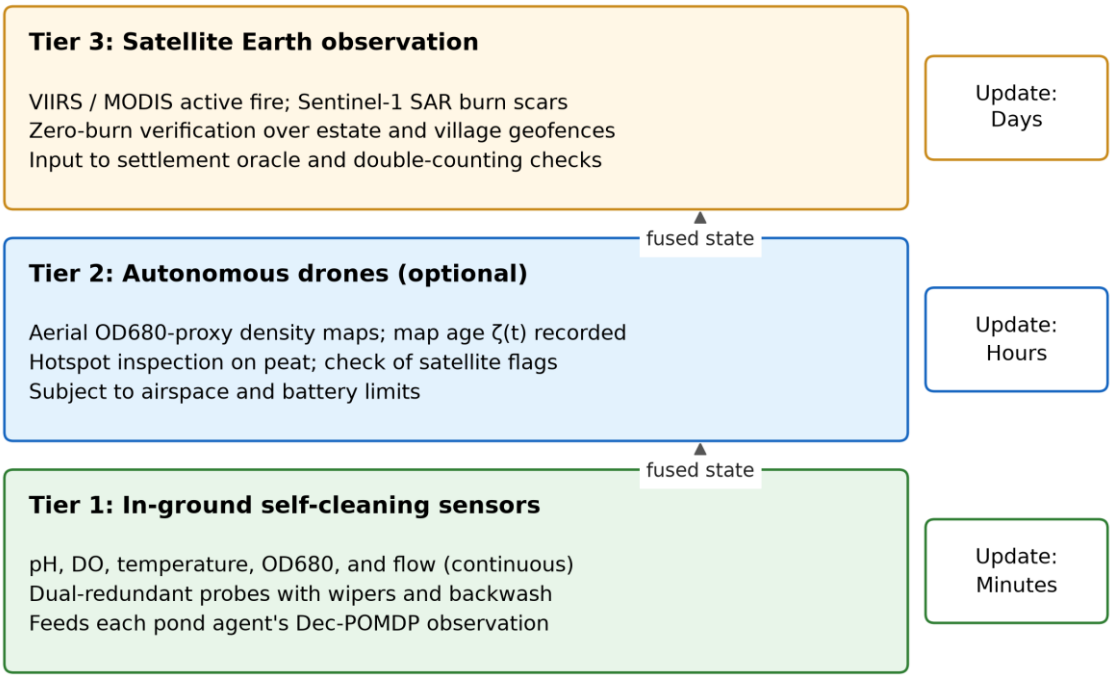
Figure 1. Dual-layer architecture of Eagor-Core



Note. Solid arrows show the main data flow within each layer; dashed arrows show cross-layer messages. Decentralized actors execute locally and do not require connectivity; the centralized critic is used only during training. PII = personally identifiable information; ACI = adaptive conformal inference; MRV = measurement, reporting, and verification.

Both layers draw on a three-tier telemetry pipeline (Figure 2). Tier 1 consists of in-ground, self-cleaning sensor arrays (pH, DO, temperature, OD_{680}, and flow) that sample continuously and form the core of each agent's observation. Tier 2 consists of autonomous drones with multispectral cameras and LiDAR, operated from a docking station on the mill site; they map biomass density across the pond grid at intervals of hours, inspect suspected hotspots on adjacent peatland, and check satellite fire flags before any payment is withheld. Tier 3 consists of satellite Earth observation, namely VIIRS and MODIS active-fire products and Sentinel-1 SAR change detection, which verify zero-burn status over estate and village geofences at intervals of days. Each higher tier covers a larger area at a lower frequency, so the tiers complement rather than replace one another. Drone flights must comply with national rules on unmanned aircraft, which limit altitude, range, and operation beyond visual line of sight. Flights over smallholder land are covered by the FPIC agreement with each village, and imagery outside the pond grid is used only for fire inspection and is not retained after review.

Figure 2. Three-tier telemetry pipeline of Eagor-Core



Note. Update intervals are indicative. Tier 1 data form the observation of each pond agent (Equation 5); Tier 2 maps enter that observation as the drone-derived feature $z_i(t)$ when available; Tier 3 data are used by the macro-layer for zero-burn verification and settlement. OD_{680} = optical density at 680 nm; LiDAR = light detection and ranging; SAR = synthetic aperture radar. Tier 2 is optional: flights depend on airspace rules and battery availability, and the age of the latest map, $\zeta_i(t)$, is passed to the agents with it.

4.2 Dec-POMDP formulation

We model an estate with N raceway ponds as a Dec-POMDP (Bernstein et al., 2002; Oliehoek & Amato, 2016) defined by the tuple

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, R, \{\Omega_i\}_{i\in\mathcal{N}}, O, \gamma \rangle, \tag{4}$$

where $\mathcal{N} = \{1, \dots, N\}$ is the set of pond agents, $\mathcal{S}$ the global state space (culture, water chemistry, equipment and weather states), $\mathcal{A}_i$ the action space of agent i, P the transition kernel, R the shared team reward, Ω_i the local observation space, O the observation function and $\gamma \in [0,1)$ the discount factor. The global state is partitioned as $s_t = (x^{bio}_t, x^{chem}_t, x^{hw}_t, x^{env}_t)$: culture states (biomass density and pond volume), water chemistry (pH, DO, dissolved inorganic carbon, nitrogen, phosphorus, and toxicity), hardware states (probe drift δ, optical fouling φ, and equipment health), and exogenous weather and grid states. Because drift and fouling evolve over time and are reset by actions, they belong to the state and their dynamics, Equations (6) and (7), are part of the transition kernel P; the observation function O maps these states to what each agent sees. The discount factor lies in [0, 1) so that the infinite-horizon objective is finite; with a finite evaluation horizon T, $\gamma = 1$ is also admissible.

Observation model. The local observation of agent i is

$$\tilde{o}_i(t) = \left[\mathrm{pH}_i + \delta_i^{\mathrm{pH}}, \mathrm{DO}_i + \delta_i^{\mathrm{DO}}, \mathrm{OD}_{680,i}(1-\phi_i), T_i, Q_i, I(t), \chi(t), z_i(t), \zeta_i(t)\right]^{\top} + \varepsilon_i(t), \tag{5}$$

where Q_i is the inflow rate, $I(t)$ the photosynthetically active irradiance, $\chi(t)$ an index of POME toxicity (phenolics and heavy metals), and $\varepsilon_i(t) \sim \mathcal{N}(0, \Sigma_\varepsilon)$ white measurement noise; $z_i(t)$ is a

drone-derived biomass-density feature for pond i, taken from the most recent multispectral map and held at its last value between flights, and $\zeta_i(t)$ is the age of that map, which lets agents discount stale data; time indices on the pond states are omitted for readability. Probe drift follows a bounded random walk that is reset by calibration, and optical fouling is a bounded, monotone process that is reset by a backwash action:

$$\delta_i^k(t+1) = \big(1 - c_i(t)\big)\, \Pi_{[-\bar{\delta}_k, \bar{\delta}_k]}\Big(\delta_i^k(t) + \xi_i^k(t)\Big), \quad \xi_i^k(t) \sim \mathcal{N}(0, \sigma_k^2), \quad k \in \{\mathrm{pH}, \mathrm{DO}\}, \tag{6}$$

$$\phi_i(t+1) = \big(1 - b_i(t)\big)\min\big\{1,\ \phi_i(t) + \kappa_f\big\}, \tag{7}$$

where κ_f is the fouling increment per control interval and $b_i(t) \in \{0,1\}$ is the backwash trigger. Drift is detected by comparing the readings of the redundant probe pair, and a calibration routine is scheduled when their divergence exceeds a tolerance. In Equation (6), Π denotes projection onto $[-\bar{\delta}_k, \bar{\delta}_k]$, the largest drift the redundant-probe check tolerates before a reading is rejected, and $c_i(t) \in \{0, 1\}$ indicates a completed calibration, which resets drift to zero. The projection prevents the unbounded growth of a pure random walk, and the resulting drift process is stationary between calibrations.

Action space. Each agent selects

$$a_i(t) = \big[\omega_i,\ q_i^{\mathrm{CO_2}},\ q_i^{\mathrm{POME}},\ q_i^{\mathrm{H}},\ v_i,\ b_i,\ u_i\big]^{\top}, \tag{8}$$

where ω_i is the paddlewheel speed, $q_i^{\mathrm{CO_2}}$, q_i^{POME} and q_i^{H} are the CO_2-injection, effluent-feed, and harvest flow rates, $v_i \in [0,1]$ is the scrubber bypass valve position and $b_i(t) \in \{0,1\}$ the backwash trigger, and $u_i(t) \in \{0, 1\}$ requests a drone survey of pond i. Survey requests are queued and served subject to battery and airspace constraints, and each flight adds to the energy term E_t. Continuous actions are bounded by actuator limits, and a hard-coded safety layer clips any action that would violate pH, DO or hydraulic constraints before it reaches the actuators.

Hybrid action space. The action vector mixes continuous components (ω_i, the three flow rates, and v_i) with binary triggers (b_i and u_i). Each actor therefore has a Gaussian head for the continuous components, squashed to actuator bounds, and independent Bernoulli heads for the binary triggers, so that the policy factorizes as $\pi_{\theta i}(a_i \mid \tau_i) = \pi^{c}{}_{\theta i}(a^{c}{}_i \mid \tau_i)\, \pi^{b}{}_{\theta i}(a^{b}{}_i \mid \tau_i)$. This parameterization is compatible with policy-gradient CTDE methods such as MAPPO (Yu et al., 2022), whose likelihood ratios are computed from the product of the two heads.

4.3 The Finamatrix vGRE index

The vGRE summarizes how productively a pond has converted effluent into biomass on healthy equipment. Three properties are required for its use inside a reward: it must respond to recent deterioration rather than average over the entire history, it must not double-count current biomass, and it must be dimensionless so that it can scale other reward terms. We therefore define vGRE as a dimensionless rolling index:

$$G_t = \frac{\sum_{k=t-W+1}^{t} V_k \; H_k \left(\frac{B_k}{B_{\max}}\right)}{\sum_{k=t-W+1}^{t} V_k}, \qquad G_t \in [0,1], \tag{9}$$

where V_k is the POME volume processed in interval k, $H_k \in [0,1]$ is a composite hardware-health index (Section 5.3), B_k is biomass density, $B_{\max}$ is the design carrying capacity and W is the window length. Because each weight $V_k / \sum V_k$ is non-negative and sums to one, and $H_k B_k / B_{\max} \in [0,1]$, G_t is a convex combination and is therefore bounded in [0, 1]. A high G_t indicates that the pond has recently processed substantial effluent while maintaining dense cultures on healthy equipment.

4.4 Reward design

The reward balances biomass gain, energy use, culture-crash risk, and corrosion exposure. Golden-ratio constants are used as weights in a form that has a clear meaning: since $\phi^{-1} + \phi^{-2} = 1$, the pair $w_B = \phi^{-1} \approx 0.618$ and $w_E = \phi^{-2} \approx 0.382$ defines a convex weighting of biomass gain against energy use. All terms are normalized by reference values so that they are dimensionless:

$$r_t = \left(\frac{1}{2} + \frac{1}{2} G_t\right) \left[w_B \, \frac{\Delta B_t}{\Delta B_{\text{ref}}} - w_E \, \frac{E_t}{E_{\text{ref}}} \right] - \kappa \max\{0, \; \hat{p}_t - \bar{p}\} - \lambda \, \frac{S_t}{S_{\text{ref}}}. \tag{10}$$

Here ΔB_t is the estate-level biomass gain, E_t the electrical energy used, $\hat{p}_t$ the predicted probability of a culture crash within the forecast horizon (from a calibrated classifier trained on simulated trajectories), $\bar{p}$ a tolerated crash probability, S_t the SO_2 exposure of the gas-handling equipment, and $\kappa, \lambda > 0$ penalty weights. The factor $1/2 + 1/2 \, G_t \in [0.5,1]$ scales the productive component by recent operating quality without suppressing the learning signal entirely when G_t is low. The hinge term replaces the discontinuous indicator with a continuous penalty that grows with the excess risk. The weights $(w_B, w_E, \kappa, \lambda)$ are value judgments (Roijers et al., 2013); the golden-ratio pair is a transparent default, not an optimality result, and Section 5.4 specifies the sensitivity analysis over these weights.

The team objective is to find decentralized policies $\pi = (\pi_1, \ldots, \pi_N)$ that maximize

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{T} \gamma^t \; r_t \right]. \tag{11}$$

Policies are trained under CTDE: each actor $\pi_{\theta_i}(a_i \mid \tau_i)$ conditions only on its local action–observation history τ_i, while a centralized critic uses the global state during training (Lowe et al., 2017; Yu et al., 2022). This structure is what allows the actors to keep running when the estate network is unavailable.

4.5 Adaptive conformal deferral

Design requirements. Split conformal prediction provides a marginal *coverage* guarantee, $\Pr\{Y_{t+h} \in \hat{C}_\alpha(X_t)\} \geq 1 - \alpha$, under exchangeability (Lei et al., 2018). Two features of the deferral problem require modifications. First, with the usual absolute-residual score the interval width is

identical for every input, so a threshold on width cannot separate easy states from hard ones. Second, sensor drift, seasonality and contamination events violate exchangeability.

Revised procedure. Let $\hat{\mu}(X_t)$ be a forecaster of a target Y_{t+h} (here, biomass density h = 24 hours ahead, normalized by $B_{\max}$), and $\hat{\sigma}(X_t) > 0$ a model of its conditional spread, both fitted on a training split. On a calibration split of size n we compute normalized residual scores, in which each absolute residual is scaled by the estimated conditional spread (Lei et al., 2018):

$$R_j = \frac{\left|Y_j - \hat{\mu}(X_j)\right|}{\hat{\sigma}(X_j)}, \qquad j = 1, \ldots, n, \tag{12}$$

and form the interval

$$\hat{C}_{\alpha_t}(X_t) = \left[\hat{\mu}(X_t) \pm \hat{q}_{\alpha_t}\, \hat{\sigma}(X_t)\right], \qquad \hat{q}_{\alpha_t} = \text{the } \lceil (n+1)(1-\alpha_t) \rceil\text{-th smallest of } \{R_j\}, \tag{13}$$

whose width $w_t = 2\,\hat{q}_{\alpha_t}\,\hat{\sigma}(X_t)$ now varies with the input. To maintain calibration under shift, the working miscoverage level is updated online by adaptive conformal inference (Gibbs & Candès, 2021): If $\lceil (n+1)(1-\alpha_t) \rceil > n$, the quantile is set to infinity and the interval is unbounded. Calibration points receive equal weight, so this is not a weighted conformal procedure: adaptivity to the input comes only through $\hat{\sigma}$, and robustness to distribution shift comes from the online update below. Weighting calibration points by their similarity to X_t (Barber et al., 2023) is a possible extension.

$$\alpha_{t+1} = \Pi_{[0,1]}\left(\alpha_t + \eta(\alpha - \mathrm{err}_t)\right), \quad \mathrm{err}_t = \mathbf{1}\left\{Y_t \notin \hat{C}_{\alpha_{t-h}}(X_{t-h})\right\}, \tag{14}$$

where α is the target miscoverage (0.1 in our experiments) and $\eta > 0$ is a step size, and $\Pi_{[0,1]}$ projects the working level onto [0, 1]; when α_t reaches 0, the interval is unbounded and the safe state is entered. When intervals miss too often, α_t falls, and intervals widen; when they over-cover, intervals tighten. ACI guarantees that the long-run miscoverage rate converges to α without distributional assumptions. Because conformal intervals capture total predictive uncertainty, we refer to w_t as a predictive uncertainty signal rather than a measure of epistemic uncertainty alone.

Execution rule. Because network loss can coincide with very high uncertainty, and a HITL request cannot be delivered without a network, the execution rule (Figure 3) evaluates conditions in a fixed order of precedence, so that exactly one mode applies at each step, with $\tau_{\text{low}} < \tau_{\text{high}}$:

$$m_t = \begin{cases} \text{Safe state + signed HITL request}, & w_t \geq \tau_{\text{high}} \text{ and network available} \\ \text{Safe state + buffered alert}, & w_t \geq \tau_{\text{high}} \text{ and network unavailable} \\ \text{Conservative fallback policy } \pi^{\text{fb}}, & \tau_{\text{low}} < w_t < \tau_{\text{high}} \\ \text{Autonomous MARL policy } \pi_\theta, & w_t \leq \tau_{\text{low}} \end{cases}. \tag{15}$$

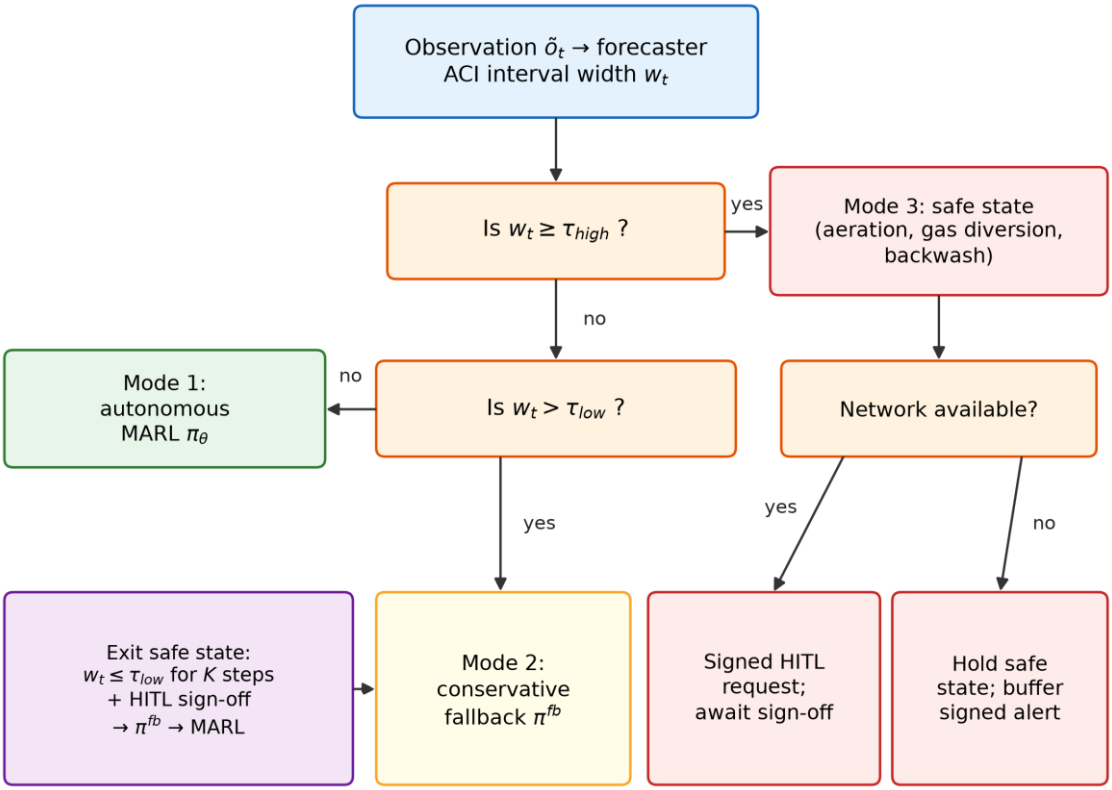
The safe state applies maximum aeration, diverts flue gas, requests a drone survey for aerial confirmation, and triggers backwashing. Signed requests are queued in the local buffer and transmitted when connectivity returns, so that every override is auditable. Because decentralized actors run locally, network loss alone does not force a change of mode; it only removes the HITL

channel. Thresholds τ_{low} and τ_{high} are set on held-out data as the 70th and 95th percentiles of w_t under normal operation.

Cost-sensitive thresholds. The percentiles above are initial values only. In the full evaluation, τ_{low} and τ_{high} are chosen on a validation set to minimize the expected cost $c_{\text{FN}}n_{\text{FN}} + c_{\text{FP}}n_{\text{FP}} + c_{\text{H}}n_{\text{H}}$, where n_{FN} counts crash episodes that were not deferred, n_{FP} counts deferrals not followed by a crash within the forecast horizon, n_{H} counts HITL requests, and $c_{\text{FN}} > c_{\text{FP}} > c_{\text{H}}$ reflect the relative costs of a culture crash, an unnecessary safe state, and operator time. The resulting confusion matrix is reported with the chosen thresholds.

Exit from the safe state. The system leaves the safe state when $w_t \leq \tau_{\text{low}}$ for K consecutive intervals and, if a HITL request was issued, the signed response has been received. It then passes through the conservative fallback policy before the MARL policy resumes, which prevents oscillation between modes. If the network remains unavailable, the safe state is held for at most T_{safe}; after that, the edge controller switches to the fallback policy, keeps the alert buffered, and repeats the backwash and calibration routine.

Figure 3. Precedence-ordered execution logic under adaptive conformal deferral



Note. w_t is the width of the adaptive conformal interval for the 24-hour-ahead biomass forecast. The high-uncertainty test is evaluated first, so the safe state is always entered when $w_t \geq \tau_{\text{high}}$, regardless of connectivity. HITL = human-in-the-loop. The system leaves the safe state only after the exit condition of Section 4.5 is met, passing through the fallback policy before MARL control resumes.

4.6 Macro-layer verification and settlement

Zero-burn status is evaluated over a rolling 30-day window for each registered estate polygon. An estate is flagged if any VIIRS 375 m or MODIS active-fire detection with nominal or high confidence falls within the polygon or a buffer around it, or if Sentinel-1 backscatter change

indicates a new burn scar. Flags trigger a drone inspection of the flagged location and HITL review rather than an automatic penalty, which limits the cost of false positives from, for example, flaring at the mill. Each harvested batch is hashed together with the estate identifier, period, and verification status, and the hash is checked against registry records before a credit is issued, which prevents the same batch from being claimed twice (Schneider et al., 2015; Franke et al., 2020). Outbound telemetry is stripped of personally identifiable information at the edge; operational data remains in Indonesia while financial settlement is executed in Singapore, consistent with the cross-border transfer provisions of both jurisdictions' data protection laws.

5 Proposed Evaluation Protocol

5.1 Simulation environment

The proposed evaluation uses a physical–biological simulator of ten open raceway ponds of 500 m³ each (operating depth 25 cm and surface area 2,000 m², about 2 ha in total), fed with pre-treated POME under Southeast Asian climatic conditions. Specific growth rate follows the multiplicative structure common in outdoor productivity models (Béchet et al., 2013):

$$\mu_i(t) = \mu_{\max} f_I\big(I(t)\big)\, f_T\big(T_i(t)\big)\, f_{\mathrm{pH}}\big(\mathrm{pH}_i(t)\big)\,\big(1-\chi(t)\big), \tag{16}$$

$$\frac{dB_i}{dt} = \big(\mu_i(t) - m - D_i(t)\big)B_i(t) - \frac{q_i^{POME}(t) + q_i^{rain}(t) - q_i^{evap}(t)}{V_i(t)}\, B_i(t), \tag{17}$$

where f_I, f_T and $f_{\mathrm{pH}} \in [0,1]$ are limitation functions for light, temperature, and pH; m is the maintenance (respiration) rate; $D_i(t)$ a stochastic loss rate representing grazing and contamination, and V_i the pond volume. Disturbances comprise diurnal temperatures of 24–38 °C, monsoonal cloud cover, irradiance spikes, fluctuations in heavy-metal and phenolic load, pH, and DO probe drift according to Equation (6), fouling according to Equation (7), and random 4-hour grid outages.

Equation (17) follows from mass balances for biomass and water. Harvesting removes culture at the pond concentration and therefore lowers the biomass stock without changing B_i, whereas inflows of POME and rain dilute the culture and evaporation concentrates it; the pond volume follows the water balance in Table A2, and an overflow weir keeps V_i within design limits. The water-chemistry states that enter the observation and the reward (pH, DO, dissolved inorganic carbon, nitrogen, phosphorus, and the toxicity index χ) follow the balance equations summarized in Table A2.

5.2 Baselines

The proposed evaluation compares Eagor-Core with three simple controllers: (i) a static rule-based schedule; (ii) independent single-loop PID controllers for pH, DO, and paddlewheel speed, tuned per pond; and (iii) a single-agent deep Q-network (DQN; Mnih et al., 2015) controlling all ponds with a discretized joint action space. Because these baselines are deliberately simple, the evaluation also includes three stronger comparators: MAPPO with independent actors and a centralized critic (Yu et al., 2022), QMIX value decomposition (Rashid et al., 2018), and a nonlinear model predictive controller that uses the simulator equations over a 24-hour horizon. The single-agent DQN mainly

illustrates the cost of a centralized controller, because its discretized joint action space grows exponentially with the number of ponds.

5.3 Performance metrics

To make the results reproducible, every metric is defined explicitly.

Biomass stability index (BSI). One minus the coefficient of variation of daily estate-level biomass productivity P_d over the evaluation horizon, truncated at zero and expressed as a percentage:

$$\mathrm{BSI} = 100 \times \max\left\{0,\ 1 - \frac{\mathrm{sd}(P_d)}{\mathrm{mean}(P_d)}\right\}. \tag{18}$$

Mean time between culture crashes (MTBCC). Total pond-days divided by the number of crash events, where a crash is a fall in biomass density of more than 50% within 48 hours.

Net biomass yield. Harvested dry biomass per hectare per year, net of biomass used for re-inoculation.

Fault recovery rate. The share of injected fault episodes (sensor drift, fouling, actuator failure, contamination, outage) after which the pond returned to within 10% of its pre-fault productivity within 72 hours without manual intervention. Episodes that triggered a HITL request are counted as not recovered autonomously.

Hardware-health index H_t. The mean of normalized health scores for probes (inverse of drift magnitude), paddlewheel drive (inverse of torque anomaly) and gas lines (inverse of cumulative SO_2 exposure), each scaled to [0, 1].

5.4 Protocol and statistical analysis (planned)

None of the following has yet been executed; it is the protocol proposed for future work. Each controller will be evaluated over at least ten independent random seeds and disturbance realizations, and results are reported as means with standard deviations and 95% bootstrap confidence intervals, because point estimates from a few RL runs are known to be unreliable (Henderson et al., 2018; Agarwal et al., 2021). Pairwise differences between Eagor-Core and each baseline are tested with the Mann–Whitney *U* test with Holm correction for multiple comparisons. The ablation study removes, one at a time, (a) the vGRE modulation, (b) conformal deferral, (c) the backwash and drift-detection actions, and (d) multi-agent decomposition. Sensitivity analysis varies w_B over [0.5, 0.8] (with $w_E = 1 - w_B$) and κ, λ over an order of magnitude. Conformal calibration is reported as empirical coverage against the 90% target, overall and within each disturbance regime. Two further ablations address the telemetry tiers: (e) removal of drone inputs, with z_i and ζ_i masked, and (f) delayed or corrupted drone maps. Scalability is tested by repeating the evaluation with 25 and 50 ponds, the latter matching the 10-hectare estate of Section 6. Fault-recovery results are reported as a confusion matrix of escalation decisions against crash outcomes.

Hypotheses. The protocol is designed to test three hypotheses about why the architecture might outperform the baselines; none is supported by evidence in this paper. H1: Decentralized actors

respond to local disturbances, such as fouling in one pond, more effectively than a single centralized controller. H2: including drift detection and backwashing in the action space lets the controller maintain its own sensors rather than act on corrupted readings. H3: Conformal deferral prevents the learned policy from acting in states far from its training distribution, which should appear as fewer crashes among episodes that were not escalated. Ablations (d), (c), and (b), respectively, isolate each mechanism, and the fault-escalation confusion matrix tests H3 directly.

6 Prospective Techno-Economic Scoping

6.1 Capital and operating costs

This section gives a prospective scoping estimate based on engineering cost assumptions and assumed yields; it is not an investment case, and no value in it comes from simulation or field data. Table 2 itemizes the CAPEX of converting a 10-hectare lagoon system, and Table 3 the annual OPEX. The ten line items sum to USD 382,000; with a 15% contingency of USD 57,300, total CAPEX is USD 439,300. Annual OPEX totals USD 93,000, including USD 8,000 for drone maintenance, battery replacement, and software subscriptions. Lining is the largest single item (Figure 4).

Table 2. Capital expenditure for a 10-hectare estate conversion

Item	Cost (USD)	Share (%)
Lagoon earthworks, grading and baffling	45,000	10.2
HDPE liners (1.5 mm, UV-stabilized, 10 ha)	85,000	19.3
Paddlewheels and redundant biogas/solar gear motors	38,000	8.7
Alkaline packed-tower flue-gas scrubbers and 316L headers	42,000	9.6
POME pre-treatment and alum metal-settling units	28,000	6.4
Self-cleaning dual sensor arrays and wiper assemblies	22,000	5.0
Drone docking station, charging pad, and two drones (multispectral and LiDAR)	18,000	4.1
Redundant edge controllers (single-board computer + microcontroller)	14,000	3.2
Solar drying beds and mechanical lipid-extraction presses	55,000	12.5
Degumming and bleaching pre-refining unit	35,000	8.0
Contingency (15% of the above)	57,300	13.0
Total CAPEX	**439,300**	**100.0**

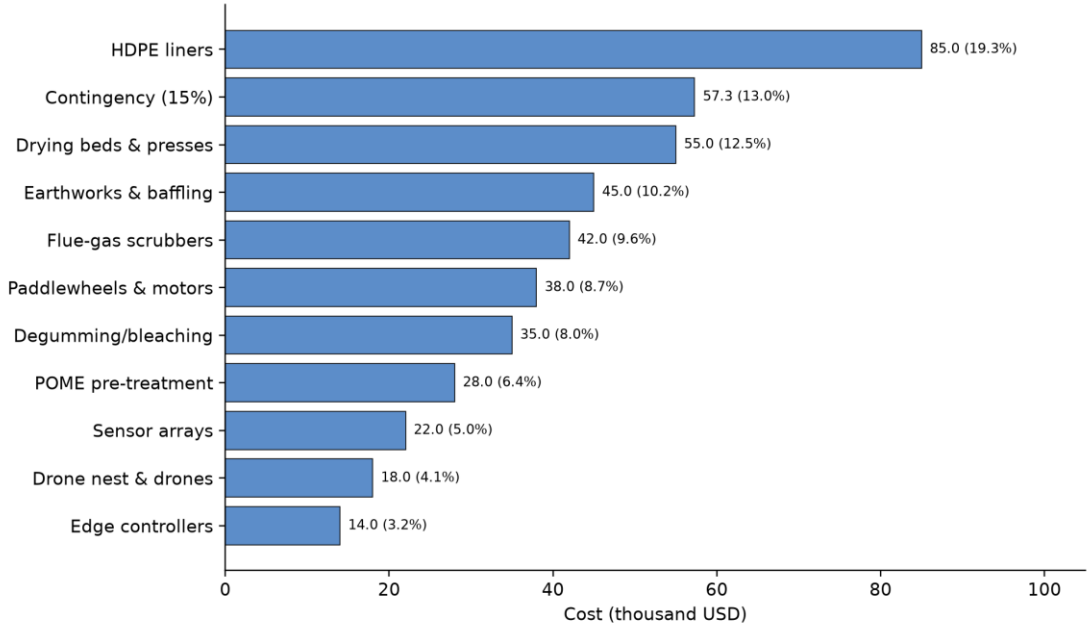
Note. Costs are engineering estimates in 2025–2026 US dollars for a retrofit of existing POME lagoons; land is not included. Shares may not sum exactly to 100 due to rounding.

Table 3. Annual operating expenditure for a 10-hectare operation

Item	Annual cost (USD)
Electrical backup and biogas power maintenance	12,000
Probe replacement, reagents, and scrubber alkali	15,000
Drone maintenance, battery replacement, and software subscriptions	8,000
Satellite connectivity and software maintenance	6,000
Local agronomist and technician labor	24,000
Community FPIC and smallholder profit-sharing reserve	18,000
Re-inoculation and contamination contingency	10,000
Total annual OPEX	**93,000**

Note. FPIC = free, prior, and informed consent. Cooperative dividends under Equation (3) are paid from net revenue and are not part of OPEX; the reserve covers administration and minimum community payments. Feedstock costs are excluded because POME and flue gas are mill by-products; this assumes the mill does not already sell them.

Figure 4. Composition of capital expenditure for the 10-hectare conversion



Note. Data from Table 2. Values are in thousand USD, with each item's share of total CAPEX (USD 439,300) in parentheses.

6.2 Annualized cost and break-even revenue

CAPEX is annualized with the capital recovery factor

$$\mathrm{CRF}(r,n) = \frac{r(1+r)^n}{(1+r)^n - 1}, \tag{19}$$

and the break-even revenue per tonne of dry biomass is

$$p^{\mathrm{BE}} = \frac{\mathrm{CAPEX} \cdot \mathrm{CRF}(r,n) + \mathrm{OPEX}}{Y \cdot A}, \tag{20}$$

where Y is the net biomass yield and A = 10 ha. Because no field or validated simulation yield is available, Y is treated as a scenario parameter with three values: 25, 35, and 45 t ha^{-1} yr^{-1}, corresponding to annual average areal productivities of about 6.8, 9.6, and 12.3 g m^{-2} d^{-1}. All three lie below the 25 g m^{-2} d^{-1} open-pond base case assumed by Davis et al. (2011), because long-term outdoor productivity is typically lower and more variable than such targets (Quinn & Davis, 2015). The analysis also assumes that productivity per hectare does not depend on the number of ponds; the 10-hectare estate corresponds to about 50 ponds of the size specified in Section 5.1, which the 50-pond scalability runs of Section 5.4 and Stage 2 of the rollout are designed to test. With an economic life of n = 10 years, Table 4 reports the break-even revenue at three discount rates and the three yield scenarios.

Table 4. Annualized cost and break-even biomass revenue (scoping estimate)

Discount rate *r*	CRF	Annualized CAPEX (USD yr^{-1})	Total annual cost (USD yr^{-1})	Break-even revenue at Y = 25 / 35 / 45 (USD t^{-1} biomass)
8%	0.1490	65,469	158,469	634 / 453 / 352
10%	0.1627	71,494	164,494	658 / 470 / 366
12%	0.1770	77,749	170,749	683 / 488 / 379

Note. Y = assumed net biomass yield in t ha^{-1} yr^{-1}; these are scenario values, not measured or simulated yields. Assumes a 10-year life and zero salvage value. Taxes, insurance, and cost inflation are not included.

The break-even revenue is compared with oil and meal revenue only; zero-burn payments are treated as a separate transfer whose minimum size is given by Equation (2). At a lipid content of 20–30% of dry weight, the three yield scenarios correspond to roughly 5.0–13.5 t of crude algae oil per hectare per year.

6.3 Revenue scenarios

Revenue per tonne of dry biomass is $R = \ell\, p_{\mathrm{CAO}} + (1 - \ell)\, p_{\mathrm{meal}}$, where ℓ is the extractable lipid share and p_{CAO} and p_{meal} are the prices of crude algae oil and defatted meal. Neither product has an established market price at scale, so Table 5 benchmarks them against crude palm oil and soybean meal, whose 2023–2025 annual average prices ranged from USD 886 to 1,007 t^{-1} and from USD 366 to 541 t^{-1}, respectively, and stood at USD 1,180 and 447 t^{-1} in September 2026 (World Bank, 2026). The low scenario prices CAO at half the 2023 palm oil price and meal at 60% of the 2025 soybean meal price, reflecting quality discounts for unproven products; the central scenario prices CAO at 80% of the 2025 palm oil price and meal at par with 2025 soybean meal; the high scenario uses September 2026 prices at par. Results are pre-tax, exclude zero-burn payments, and use r = 10%.

Table 5. Revenue scenarios and annual net margin for the 10-hectare estate (scoping estimate)

Scenario	Lipid share ℓ	CAO price (USD t^{-1})	Meal price (USD t^{-1})	Revenue (USD t^{-1} biomass)	Net margin at Y = 25 / 35 / 45 (USD yr^{-1})
Low	0.20	443	220	265	−98,344 / −71,884 / −45,424
Central	0.25	806	366	476	−45,494 / 2,106 / 49,706
High	0.30	1,180	447	667	2,231 / 68,921 / 135,611

Note. Net margin = R × Y × 10 ha − total annual cost at r = 10% (USD 164,494; Table 4), for assumed yields Y of 25, 35, and 45 t ha^{-1} yr^{-1}. Prices are benchmarks derived from the World Bank (2026), not forecasts. CAO = crude algae oil.

Under these assumptions, the estate would break even at central prices only if the yield reaches about 35 t ha^{-1} yr^{-1}, and at 25 t ha^{-1} yr^{-1}, only the high price scenario would cover costs. The economic case, therefore, depends as much on field productivity as on prices, which is why Stage 1 measures yield before the Stage 2 investment.

7 Discussion and Research Agenda

7.1 Policy implications

The incentive model in Section 3 shows that the economic case for the system rests on two parameters that policy can influence: the detection probability p and the net margin ΔR. Satellite-based verification raises p at very low marginal cost, and a profitable algae operation raises ΔR without a fiscal transfer. Where ΔR alone does not cover the cost advantage of fire, Equation (2) gives the minimum conditional payment, which can be financed by voluntary carbon buyers, SAF off-takers seeking traceable feedstock, or public programs. Because payments are conditional on a rolling verification window and reviewed by humans before penalties are applied, the design follows established principles for PES schemes (Engel et al., 2008; Jack et al., 2008).

The cooperative extension matters because independent smallholders manage a large share of the planted area but are poorly served by certification (Jelsma et al., 2017; Carlson et al., 2018). Routing algae revenue through the mill gives them a stake in non-burning without requiring them to invest or certify individually. The village-level condition in Equation (3) creates peer monitoring, and the graduated, plot-attributable rule in Section 3 limits the losses that compliant farmers suffer from others' fires.

7.2 Future extension: vertical photobioreactors and firm low-carbon power

Open raceways remain exposed to contamination, weather, and land constraints. A longer-term extension of the framework would replace or supplement lagoon raceways with automated vertical photobioreactor (PBR) towers. One such concept, the AMTS-IoT Vertical Class proposed by EVES Energy, stacks modular glass and stainless-steel PBR units vertically and couples them to an IoT sensing layer and automated culture regulation with continuous partial harvesting and culture return. Each unit maps directly onto an agent of the Dec-POMDP in Section 4.2, so the control and deferral layers of Eagor-Core would apply without structural change. A further concept, Project AAI-Green Nexus, would pair such towers with a small modular reactor (SMR) supplying continuous low-carbon electricity for pumping, lighting, and control, and waste heat for biomass drying, harvesting, and lipid extraction. Its intended outputs are algae oil as a sustainable aviation fuel feedstock, protein-rich biomass, and biologically captured CO_2.

Neither concept has been built or tested, and both raise questions this paper does not address. Closed PBRs have historically had much higher production costs than open ponds; in the base case of Davis et al. (2011), the cost of lipid production was USD 18.10 per gallon for tubular PBRs against USD 8.52 for open ponds. Artificial lighting adds a large energy penalty. SMR deployment would require licensing, siting and safety assessment, and public acceptance in Singapore and Indonesia, neither of which currently operates a nuclear power plant. We therefore treat the vertical PBR–SMR pathway as a research direction for the framework's control and governance layers. It is not part of the techno-economic scoping in Section 6, which is limited to lagoon retrofits.

7.3 Limitations and threats to validity

This study has several limitations. First, the paper reports no performance results: neither the simulator nor the controllers have yet been evaluated, and the yields in Section 6 are assumptions; field productivity of outdoor ponds is typically lower and more variable than modeled, and contamination dynamics are particularly hard to simulate (Day et al., 2012; Wang et al., 2013). Second, the hypotheses H1–H3 remain untested, so whether each component contributes anything is unknown. Third, the claim that the system is "carbon-negative" cannot be supported without a life-cycle assessment: carbon fixed in algae oil is released when the fuel is combusted, so the net balance depends on displaced fossil fuel, avoided methane from POME lagoons, and avoided peat fires, none of which are quantified here. We therefore describe the system as low-carbon pending such an assessment. Fourth, CAO is a feedstock, not a fuel: it must be hydroprocessed in a HEFA facility, and the finished blend must meet ASTM D7566 (ASTM International, 2026). The impurity limits cited in the appendix are illustrative of off-taker requirements, not certified specifications. Fifth, the architecture relies on MODIS products for continuity, but the MODIS instruments are near the end of their missions; VIIRS should be treated as the primary active-fire source. Sixth, the cooperative

dividend has not been calibrated to actual FFB volumes and prices, and tokenized membership units may raise regulatory questions. Seventh, the drone tier adds cost and raises airspace and privacy issues: flights over smallholder plots require community consent, imagery must be limited to the pond grid and fire inspection, and its contribution to control performance has not yet been isolated (ablations (e) and (f) in Section 5.4). Finally, the open-source claim can be verified only once the repository described under Declarations is public.

Research agenda. The claims of this paper become testable in four steps: (i) the multi-seed protocol of Section 5.4, with MAPPO, QMIX, and nonlinear MPC baselines, ablations (a)–(f), the fault-escalation confusion matrix, and the 25- and 50-pond scalability runs; (ii) a Stage 1 pilot on an operating mill to replace the assumed yields with measured field productivity; (iii) a life-cycle assessment to test carbon claims; and (iv) a cooperative field study to calibrate FFB volumes, dividend shares, and HITL review times.

8 Conclusion

Eagor-Core is a conceptual framework that links three things usually studied separately: fault-tolerant control of microalgae ponds, calibrated rules for deferring decisions to humans, and conditional payments for verified non-burning. Its contributions are a Dec-POMDP formulation with sensor-drift, fouling, and water-chemistry states; a bounded vGRE-modulated reward; adaptive conformal deferral with a precedence-ordered execution rule and safe-state exit conditions; an incentive model with a graduated smallholder dividend; and an evaluation protocol. The evaluation protocol and the techno-economic scoping estimate show how the framework would be assessed, not whether it works or pays. The research agenda in Section 7 sets out the multi-seed evaluation, field pilot, and life-cycle assessment needed to test it.

Declarations

Conflict of interest. Lanz CWJ Chan is affiliated with Finamatrix Quant-Lab, which developed the vGRE indicator and the Eagor-Core software. The policy brief in Appendix B requests public co-funding for a demonstration facility in which Finamatrix Quant-Lab may participate. Shin-Hung Pan declares no competing interests. The vertical photobioreactor and SMR concepts discussed in Section 7.2 (AMTS-IoT Vertical Class and Project AAI-Green Nexus) are proposed by EVES Energy and Finamatrix, with which Lanz CWJ Chan is associated.

Data and code availability. No datasets were generated or analyzed in this conceptual study. The simulator specification is given in Section 5.1 and Table A2; code implementing the framework and the evaluation protocol will be deposited in a public repository with a persistent identifier before any results are reported.

References

ASEAN. (2002). *ASEAN Agreement on Transboundary Haze Pollution*. ASEAN Secretariat.

ASTM International. (2026). *Standard specification for aviation turbine fuel containing synthesized hydrocarbons* (ASTM D7566-26a). ASTM International.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., & Bellemare, M. G. (2021). Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems, 34*, 29304–29320.

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

Bagchi, S. K., Patnaik, R., & Prasad, R. (2021). Feasibility of utilizing wastewaters for large-scale microalgal cultivation and biofuel productions using hydrothermal liquefaction technique: A comprehensive review. *Frontiers in Bioengineering and Biotechnology, 9*, Article 651138. https://doi.org/10.3389/fbioe.2021.651138

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics, 51*(2), 816–845. https://doi.org/10.1214/23-AOS2276

Béchet, Q., Shilton, A., & Guieysse, B. (2013). Modeling the effects of light and temperature on algae growth: State of the art and critical assessment for productivity prediction during outdoor cultivation. *Biotechnology Advances, 31*(8), 1648–1663. https://doi.org/10.1016/j.biotechadv.2013.08.014

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*(2), 169–217. https://doi.org/10.1086/259394

Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research, 27*(4), 819–840. https://doi.org/10.1287/moor.27.4.819.297

Berti Suman, A. (2022). The role of information in multilateral governance of environmental health risk: Lessons from the Equatorial Asian haze case. *Journal of Risk Research, 25*(8), 959–975. https://doi.org/10.1080/13669877.2019.1617338

Brennan, L., & Owende, P. (2010). Biofuels from microalgae—A review of technologies for production, processing, and extractions of biofuels and co-products. *Renewable and Sustainable Energy Reviews, 14*(2), 557–577. https://doi.org/10.1016/j.rser.2009.10.009

Carlson, K. M., Heilmayr, R., Gibbs, H. K., Noojipady, P., Burns, D. N., Morton, D. C., Walker, N. F., Paoli, G. D., & Kremen, C. (2018). Effect of oil palm sustainability certification on deforestation and fire in Indonesia. *Proceedings of the National Academy of Sciences, 115*(1), 121–126. https://doi.org/10.1073/pnas.1704728114

Chan, L. C. W., & Pan, S.-H. V. (2026). *Multi-attribute decision frameworks for autonomous agent operations in bio-industrial bioprocessing* [Working paper]. Finamatrix Quant-Lab.

Chan, L. C. W., & Wong, W.-K. (2016). *vGRE-Lite: Volume-weighted golden ratio estimator for safe-haven regime detection* [Working paper]. Finamatrix Quant-Lab.

Chan, L. C. W., & Wong, W.-K. (2026a). *EAGOR: Energy–AI governance orchestration in recursive decision architectures* [Working paper]. Finamatrix Quant-Lab.

Chan, L. C. W., & Wong, W.-K. (2026b). Volume-weighted golden ratio estimator (vGRE) for drawdown and tail-risk control for wealth portfolios. *International Journal of Wealth*

Management, 2, 1–23. https://ijwm.scibiz.world/volume-weighted-golden-ratio-estimator-vgre-for-drawdown-and-tail-risk-control-for-wealth-portfolios

Chan, L. C. W., Pan, S.-H. V., & Wong, W.-K. (2026). *Human–AI collaborative decision-making under uncertainty: The HACDU framework and benchmark suite* [Working paper]. Finamatrix Quant-Lab.

Cheah, W. Y., Show, P. L., Juan, J. C., Chang, J.-S., & Ling, T. C. (2018). Microalgae cultivation in palm oil mill effluent (POME) for lipid production and pollutants removal. *Energy Conversion and Management, 174*, 430–438. https://doi.org/10.1016/j.enconman.2018.08.057

Chisti, Y. (2007). Biodiesel from microalgae. *Biotechnology Advances, 25*(3), 294–306. https://doi.org/10.1016/j.biotechadv.2007.02.001

Chisti, Y. (2013). Constraints to commercialization of algal fuels. *Journal of Biotechnology, 167*(3), 201–214. https://doi.org/10.1016/j.jbiotec.2013.07.020

Civil Aviation Authority of Singapore. (2024). *Singapore sustainable air hub blueprint*. https://www.caas.gov.sg/sustainability/sustainable-air-hub-blueprint

Civil Aviation Authority of Singapore. (2026, September 3). *CAAS to start sustainable aviation fuel levy for origin-destination passengers, general and business aviation flights from 1 October 2026 for flights departing from 1 January 2027* [Press release]. https://www.caas.gov.sg/resources/media-and-publication/newsroom/caas-to-start-sustainable-aviation-fuel-levy-for-origin-destination-passengers-from-1-october-2026-for-flights-departing-from-1-january-2027

Davis, R., Aden, A., & Pienkos, P. T. (2011). Techno-economic analysis of autotrophic microalgae for fuel production. *Applied Energy, 88*(10), 3524–3531. https://doi.org/10.1016/j.apenergy.2011.04.018

Day, J. G., Slocombe, S. P., & Stanley, M. S. (2012). Overcoming biological constraints to enable the exploitation of microalgae for biofuels. *Bioresource Technology, 109*, 245–251. https://doi.org/10.1016/j.biortech.2011.05.033

Doliente, S. S., Narayan, A., Tapia, J. F. D., Samsatli, N. J., Zhao, Y., & Samsatli, S. (2020). Bio-aviation fuel: A comprehensive review and analysis of the supply chain components. *Frontiers in Energy Research, 8*, Article 110. https://doi.org/10.3389/fenrg.2020.00110

Engel, S., Pagiola, S., & Wunder, S. (2008). Designing payments for environmental services in theory and practice: An overview of the issues. *Ecological Economics, 65*(4), 663–674. https://doi.org/10.1016/j.ecolecon.2008.03.011

Euler, M., Schwarze, S., Siregar, H., & Qaim, M. (2016). Oil palm expansion among smallholder farmers in Sumatra, Indonesia. *Journal of Agricultural Economics, 67*(3), 658–676. https://doi.org/10.1111/1477-9552.12163

Franke, L., Schletz, M., & Salomo, S. (2020). Designing a blockchain model for the Paris Agreement's carbon market mechanism. *Sustainability, 12*(3), 1068. https://doi.org/10.3390/su12031068

García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research, 16*(42), 1437–1480.

Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems, 34*, 1660–1672.

Giglio, L., Schroeder, W., & Justice, C. O. (2016). The collection 6 MODIS active fire detection algorithm and fire products. *Remote Sensing of Environment, 178*, 31–41. https://doi.org/10.1016/j.rse.2016.02.054

Gil, J. D., Del Rio Chanona, E. A., Guzmán, J. L., & Berenguel, M. (2026). *From learning to control: Data-driven multi-agent reinforcement learning for multivariable control in a microalgae bioprocess* (arXiv:2609.09313). arXiv. https://doi.org/10.48550/arXiv.2609.09313

Glauber, A. J., Moyer, S., Adriani, M., & Gunawan, I. (2016). *The cost of fire: An economic analysis of Indonesia's 2015 fire crisis* (Indonesia Sustainable Landscapes Knowledge Note No. 1). World Bank.

Hariz, H. B., Takriff, M. S., Mohd Yasin, N. H., Ba-Abbad, M. M., & Mohd Hakimi, N. I. N. (2019). Potential of the microalgae-based integrated wastewater treatment and CO_2 fixation system to treat palm oil mill effluent (POME) by indigenous microalgae; *Scenedesmus* sp. and *Chlorella* sp. *Journal of Water Process Engineering, 32*, 100907. https://doi.org/10.1016/j.jwpe.2019.100907

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://doi.org/10.1609/aaai.v32i1.11694

Huijnen, V., Wooster, M. J., Kaiser, J. W., Gaveau, D. L. A., Flemming, J., Parrington, M., Inness, A., Murdiyarso, D., Main, B., & van Weele, M. (2016). Fire carbon emissions over maritime southeast Asia in 2015 largest since 1997. *Scientific Reports, 6*, 26886. https://doi.org/10.1038/srep26886

Jack, B. K., Kousky, C., & Sims, K. R. E. (2008). Designing payments for ecosystem services: Lessons from previous experience with incentive-based mechanisms. *Proceedings of the National Academy of Sciences, 105*(28), 9465–9470. https://doi.org/10.1073/pnas.0705503104

Jelsma, I., Schoneveld, G. C., Zoomers, A., & van Westen, A. C. M. (2017). Unpacking Indonesia's independent oil palm smallholders: An actor-disaggregated approach to identifying environmental and social performance challenges. *Land Use Policy, 69*, 281–297. https://doi.org/10.1016/j.landusepol.2017.08.012

Koh, L. P., Miettinen, J., Liew, S. C., & Ghazoul, J. (2011). Remotely sensed evidence of tropical peatland conversion to oil palm. *Proceedings of the National Academy of Sciences, 108*(12), 5127–5132. https://doi.org/10.1073/pnas.1018776108

Koplitz, S. N., Mickley, L. J., Marlier, M. E., Buonocore, J. J., Kim, P. S., Liu, T., Sulprizio, M. P., DeFries, R. S., Jacob, D. J., Schwartz, J., Pongsiri, M., & Myers, S. S. (2016). Public health impacts of the severe haze in Equatorial Asia in September–October 2015: Demonstration of a new framework for informing fire management strategies to reduce downwind smoke exposure. *Environmental Research Letters, 11*(9), 094023. https://doi.org/10.1088/1748-9326/11/9/094023

Lam, M. K., & Lee, K. T. (2011). Renewable and sustainable bioenergies production from palm oil mill effluent (POME): Win–win strategies toward better environmental protection. *Biotechnology Advances, 29*(1), 124–141. https://doi.org/10.1016/j.biotechadv.2010.10.001

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association, 113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Low, S. S., Bong, K. X., Mubashir, M., Cheng, C. K., Lam, M. K., Lim, J. W., Ho, Y. C., Lee, K. T., Munawaroh, H. S. H., & Show, P. L. (2021). Microalgae cultivation in palm oil mill effluent (POME) treatment and biofuel production. *Sustainability, 13*(6), Article 3247. https://doi.org/10.3390/su13063247

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems, 30*, 6379–6390.

Ma, Y., Noreña-Caro, D. A., Adams, A. J., Brentzel, T. B., Romagnoli, J. A., & Benton, M. G. (2020). Machine-learning-based simulation and fed-batch control of cyanobacterial-phycocyanin production in *Plectonema* by artificial neural network and deep reinforcement learning. *Computers & Chemical Engineering, 142*, 107016. https://doi.org/10.1016/j.compchemeng.2020.107016

Marlier, M. E., DeFries, R. S., Voulgarakis, A., Kinney, P. L., Randerson, J. T., Shindell, D. T., Chen, Y., & Faluvegi, G. (2013). El Niño and health risks from landscape fire emissions in southeast Asia. *Nature Climate Change, 3*(2), 131–136. https://doi.org/10.1038/nclimate1658

Ministry of Sustainability and the Environment. (2026, June 29). *Singapore and Indonesia sign Memorandum of Understanding on Environmental Cooperation* [Press release]. https://www.mse.gov.sg/latest-news/singapore-and-indonesia-sign-memorandum-of-understanding-on-environmental-cooperation

Ministry of Trade and Industry. (2026, July 6). *Singapore and Indonesia sign Memorandum of Understanding on Carbon Credits Collaboration* [Press release]. https://www.mti.gov.sg/newsroom/singapore-and-indonesia-sign-memorandum-of-understanding-on-carbon-credits-collaboration/

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review, 56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning* (pp. 278–287). Morgan Kaufmann.

Oliehoek, F. A., & Amato, C. (2016). *A concise introduction to decentralized POMDPs*. Springer. https://doi.org/10.1007/978-3-319-28929-8

Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D., & del Rio-Chanona, E. A. (2020). Reinforcement learning for batch bioprocess optimization. *Computers & Chemical Engineering, 133*, 106649. https://doi.org/10.1016/j.compchemeng.2019.106649

Quinn, J. C., & Davis, R. (2015). The potentials and challenges of algae based biofuels: A review of the techno-economic, life cycle, and resource assessment modeling. *Bioresource Technology, 184*, 444–452. https://doi.org/10.1016/j.biortech.2014.10.075

Rashid, T., Samvelyan, M., Schroeder de Witt, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning.

In *Proceedings of the 35th International Conference on Machine Learning* (PMLR 80, pp. 4295–4304).

Republic of Indonesia. (2009). *Law No. 32 of 2009 on Environmental Protection and Management*.

Republic of Indonesia. (2022). *Law No. 27 of 2022 on Personal Data Protection*.

Republic of Singapore. (2012). *Personal Data Protection Act 2012* (Act No. 26 of 2012).

Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research, 48*, 67–113. https://doi.org/10.1613/jair.3987

Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems, 32*, 3543–3553.

Schneider, L., Kollmuss, A., & Lazarus, M. (2015). Addressing the risk of double counting emission reductions under the UNFCCC. *Climatic Change, 131*(4), 473–486. https://doi.org/10.1007/s10584-015-1398-y

Schroeder, W., Oliva, P., Giglio, L., & Csiszar, I. A. (2014). The New VIIRS 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment, 143*, 85–96. https://doi.org/10.1016/j.rse.2013.12.008

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal, 3*(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Tacconi, L. (2016). Preventing fires and haze in Southeast Asia. *Nature Climate Change, 6*(7), 640–643. https://doi.org/10.1038/nclimate3008

Tan, K. T., Lee, K. T., Mohamed, A. R., & Bhatia, S. (2009). Palm oil: Addressing issues and towards sustainable development. *Renewable and Sustainable Energy Reviews, 13*(2), 420–427. https://doi.org/10.1016/j.rser.2007.10.001

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas Traver, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., . . . Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment, 120*, 9–24. https://doi.org/10.1016/j.rse.2011.05.028

Treloar, N. J., Fedorec, A. J. H., Ingalls, B., & Barnes, C. P. (2020). Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLOS Computational Biology, 16*(4), e1007783. https://doi.org/10.1371/journal.pcbi.1007783

Varkkey, H. (2013). Patronage politics, plantation fires and transboundary haze. *Environmental Hazards, 12*(3–4), 200–217. https://doi.org/10.1080/17477891.2012.759524

Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic learning in a random world* (2nd ed.). Springer. https://doi.org/10.1007/978-3-031-06649-8

Wang, H., Zhang, W., Chen, L., Wang, J., & Liu, T. (2013). The contamination and control of biological pollutants in mass cultivation of microalgae. *Bioresource Technology, 128*, 745–750. https://doi.org/10.1016/j.biortech.2012.10.158

World Bank. (2026). *World Bank commodities price data (The Pink Sheet), October 2026*. https://thedocs.worldbank.org/en/doc/74e8be41ceb20fa0da750cda2f6b9e4e-0050012026/related/CMO-Pink-Sheet-October-2026.pdf

Wu, T. Y., Mohammad, A. W., Jahim, J. M., & Anuar, N. (2010). Pollution control technologies for the treatment of palm oil mill effluent (POME) through end-of-pipe processes. *Journal of*

Environmental Management, 91(7), 1467–1490. https://doi.org/10.1016/j.jenvman.2010.02.008

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems, 35*, 24611–24624.

Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, & D. Cansever (Eds.), *Handbook of reinforcement learning and control* (pp. 321–384). Springer. https://doi.org/10.1007/978-3-030-60990-0_12

Appendix A. Low-CAPEX Palm Oil Estate Transformation Manual

A practical field guide for plantation owners, engineers, and cooperatives

A.1 Overview and financial rationale

This manual describes how to convert existing POME lagoons into microalgae cultivation units. Repurposing mill infrastructure (lagoons, boilers, biogas generators) avoids land acquisition and much of the civil work of a greenfield plant. A drone docking station and two drones are added to the standard retrofit to support Tier 2 monitoring.

Table A1. Comparison of a conventional CPO estate with an Eagor-Core transformed estate

Metric	Conventional CPO estate (fire-assisted expansion)	Eagor-Core transformed estate
Primary feedstock cost	High (land acquisition, fertilizer)	Near zero (POME and boiler CO_2 as by-products)
Additional land required	New land for expansion	None (existing lagoon footprint)
Annual oil yield	About 3–4 t CPO ha^{-1} yr^{-1}	About 5.0–13.5 t CAO ha^{-1} yr^{-1} (assumed yield scenarios; 20–30% lipid)
Environmental compliance	Exposure to sanctions under Law No. 32/2009	Reduced effluent discharge; carbon balance subject to LCA
Product	Standard CPO	CAO as HEFA feedstock (illustrative limit ≤ 15 ppm total metals and phosphorus)
Smallholder inclusion	Marginalized; pressure to expand by burning	Cooperative dividends from mill algae revenue, conditional on zero-burn verification
Settlement latency	Months (conventional supply chains)	Under 12 hours after verification (smart contract)

Note. CAO yield is derived from the assumed biomass yields of 25–45 t ha^{-1} yr^{-1} (Section 6) and an assumed lipid content of 20–30%; it is not a measured or simulated value. LCA = life-cycle assessment; HEFA = hydroprocessed esters and fatty acids.

A.2 Smallholder inclusion and cooperative business model

Independent smallholders manage about 40% of Indonesia's oil palm area (Jelsma et al., 2017). Because they have no mills of their own, certification-based policies often exclude them, and expansion by burning remains tempting. Eagor-Core therefore uses a mill–cooperative aggregation structure:

1. *Mill-level aggregation.* The palm oil mill (*pabrik kelapa sawit*, PKS) builds the raceway facility on its central POME lagoon footprint, so smallholders bear no land or capital cost.

2. *Cooperative membership.* Smallholders who supply FFB to the mill are registered as cooperative members and receive digital membership units recorded on the settlement ledger.

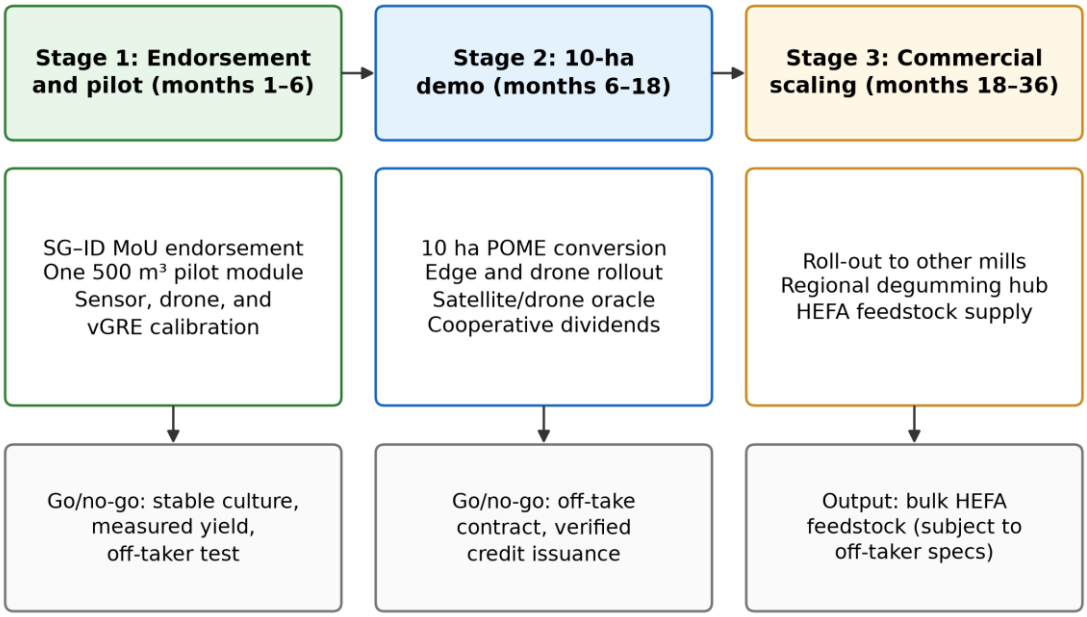
3. *Dividend mechanics.* A share θ of net revenue from CAO sales and verified credits is distributed quarterly in proportion to FFB deliveries, following Equation (3).

4. *Conditionality.* Dividends are reduced in proportion to the verified burned share of the village geofence (Equation 3), and fires that drones attribute to a registered plot reduce only that member's dividend. Each reduction follows HITL review and can be appealed before settlement.

If the membership units are transferable tokens, they may fall under securities or crypto-asset regulation in Indonesia; a non-transferable ledger entry avoids this question.

A.3 Three-stage commercial rollout

Figure 5. Three-stage commercial rollout roadmap



Note. Durations are planning targets. Each stage proceeds only if the go/no-go criteria shown at its end are met. HEFA = hydroprocessed esters and fatty acids.

1. *Stage 1: Bilateral endorsement and pilot (months 1–6).* Seek endorsement under existing bilateral instruments, such as the Singapore–Indonesia Memorandum of Understanding (MoU) on Environmental Cooperation (Ministry of Sustainability and the Environment, 2026) and the MoU on carbon credits collaboration under Article 6 of the Paris Agreement (Ministry of Trade and Industry, 2026). Retrofit a single 500 m³ lagoon module (about 0.2 ha). Validate culture stability, calibrate the edge controllers, drones, and vGRE parameters under local weather, test CAO quality against off-taker feedstock limits, and measure field productivity to replace the simulated yield in the payback assessment.

2. *Stage 2: 10-hectare demonstration (months 6–18).* Expand to the complete 10-hectare facility, deploy redundant edge controllers, wet scrubbers, and drones, integrate drone and satellite verification into the settlement oracle, sign an off-take agreement, and activate cooperative dividend settlement.

3. *Stage 3: Commercial scaling (months 18–36).* Use the demonstration results to attract private investment and extend the system to other mills, and establish a regional pre-refining hub that aggregates CAO from ten or more neighboring mills and performs centralized degumming and bleaching, supplying bulk HEFA feedstock to refiners whose jet fuel blends are certified under ASTM D7566.

A.4 Risk management framework

Biological and contamination risks.

- *Wild strains, viruses, and ciliates.* Apply periodic pulse-dosing of mild organic acids or salinity shocks scheduled by Eagor-Core. Maintain high-density inocula in closed photobioreactor (PBR) seed tanks to reseed ponds immediately after contamination (Wang et al., 2013).

- *Heavy metals and POME toxicity.* Pre-treat POME through anaerobic digestion traps, alum coagulation, and settling basins to remove heavy metals (Cr, Pb, Cd) and excess solids before the effluent enters the ponds (Wu et al., 2010).

- *Seasonal light and temperature variability.* Adjust culture depth dynamically, for example, toward 15 cm on overcast days to improve light availability and toward 30 cm during intense heat to buffer temperature and limit photoinhibition.

Hardware and infrastructure degradation.

- *Sensor fouling and probe drift.* Install dual-redundant pH and DO probes with mechanical wipers and pressurized backwash rings. Eagor-Core detects drift through cross-probe divergence (Equation 6).

- *Paddlewheel failures and power outages.* Pair low-wattage gear motors with dual-feed power switching to mill biogas generators and solar battery backup. Fit shear-pin protection on paddlewheel axles.

- *Flue-gas scrubber and pipe corrosion.* Route flue gas through an alkaline packed-tower scrubber to remove SO_2 and capture acidic condensate before it reaches the PVC headers and 316L stainless steel nozzles.

- *Loss of connectivity.* Edge controllers run fully autonomously on local SQLite state databases. When the satellite link fails, decentralized policies continue, and audit logs are buffered until connectivity resumes.

- *Drone failure and airspace restrictions.* Treat drone data as optional inputs: agents fall back on Tier 1 and Tier 3 data when flights are grounded by weather, battery faults, or airspace restrictions, and the age indicator in the observation flags stale maps.

Macro, financial, legal, and compliance risks.

- *Double counting of credits.* Link satellite-verified zero-burn status to batch-level production logs through cryptographic hashes, and check every claim against registry records before issuance (Schneider et al., 2015).

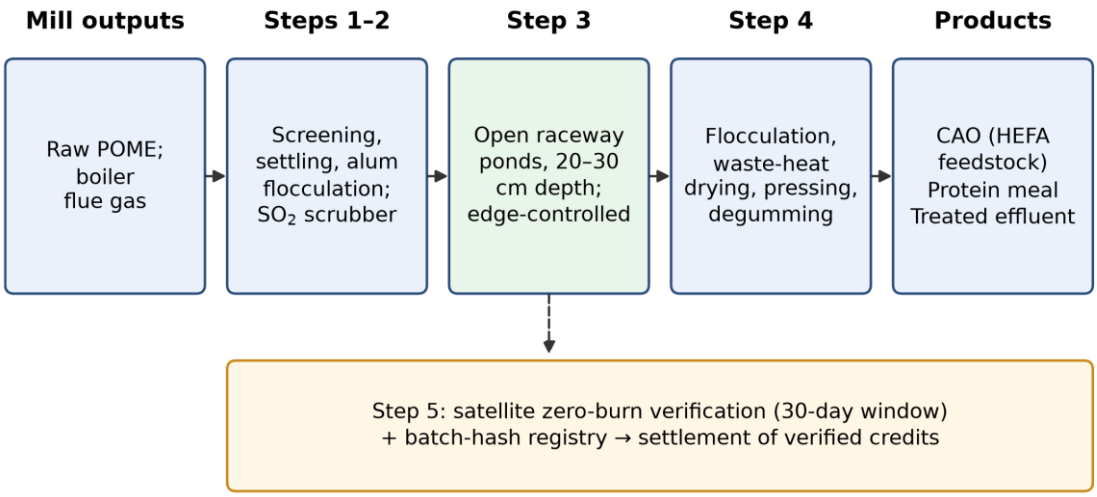
- *Regulatory and community governance.* Comply with Law No. 32/2009 and the ASEAN Agreement on Transboundary Haze Pollution. Establish free, prior, and informed consent (FPIC) agreements and profit-sharing schemes with local communities.

- *Off-taker quality requirements.* Subject harvested CAO to centrifugal degumming and bleaching to meet the off-taker's feedstock limits (for example, ≤ 15 ppm total metals and phosphorus and < 0.5% free fatty acids). These limits are illustrative of off-taker requirements, not certification criteria; ASTM D7566 applies only to the finished, hydroprocessed jet fuel blend.

- *Cross-border data privacy.* Strip personally identifiable information at the edge, store operational data in Indonesia, and execute financial settlement in Singapore, in line with Singapore's PDPA 2012 and Indonesia's Law No. 27 of 2022.

A.5 Step-by-step implementation guide

Figure 6. Process flow of the five-step estate conversion



Note. Steps correspond to Steps 1 to 5 of Section A.5. The dashed arrow indicates that batch-level production logs from the ponds feed the verification and settlement step.

Step 1: Lagoon retrofitting and hydraulic engineering

1. *Pond re-engineering.* Convert four- to five-stage anaerobic POME lagoons into shallow open raceway ponds with a uniform depth of 20–30 cm.

2. *Baffling and lining.* Install HDPE liners (1.0–1.5 mm) and central divider baffles to maintain a uniform channel flow.

3. *Paddlewheels.* Mount locally fabricated four-blade paddlewheels driven by redundant gear motors powered by mill biogas generators and battery backup. Maintain a channel velocity of 0.2–0.3 m s^{-1}.

4. *Flue-gas scrubbing and sparging.* Pass boiler exhaust through an alkaline packed-tower scrubber to remove SO_2, then distribute it through PVC piping and 316L stainless steel spargers along the pond floor.

Step 2: Strain selection and effluent pre-treatment

1. *Pre-treatment.* Pass raw POME through a 200-mesh vibrating screen, an anaerobic settling trap, and an alum flocculator to remove heavy metals, coarse solids, and excess phenolics.

2. *Inoculation.* Inoculate with robust strains such as *Chlorella vulgaris* or *Arthrospira platensis* (Spirulina) adapted to 30–38 °C. Maintain PBR seed tanks for rapid recovery.

3. *Nutrient balance.* Use the nitrogen and phosphorus already present in POME as the nutrient source, supplementing only where analysis shows a deficit.

Step 3: Installing and calibrating Eagor-Core edge gateways

1. *Hardware.* Install a redundant edge pair in an IP65 enclosure at each pond cluster, for example, a single-board computer (Raspberry Pi 4) as the primary controller and an ESP32-based programmable logic controller as a watchdog and fallback. Policies deployed to the microcontroller should be distilled or quantized to fit its memory.

2. *Sensors.* Connect self-cleaning sensor arrays with mechanical wipers: redundant gel-filled pH probes (0–14), dual optical-density sensors (OD_{680}) for biomass, galvanic DO sensors, and waterproof PT100 temperature sensors.

3. *Drones.* Install a docking and charging station with one multispectral and one LiDAR drone, register the aircraft and remote pilots with the civil aviation authority, and define flight corridors over the pond grid and the estate boundary.

Step 4: Low-energy harvesting and refining

1. *Flocculation.* Use low-cost organic flocculants, or induce flocculation by raising pH under Eagor-Core control, to concentrate the slurry without energy-intensive centrifugation.

2. *Drying.* Dewater the biomass on solar drying beds assisted by boiler waste heat.

3. *Fractionation.* Extract lipids by mechanical pressing followed by centrifugal degumming to produce CAO for HEFA processing, and dry the residual biomass (typically 40–50% protein) for sale as aquaculture feed.

Step 5: Satellite zero-burn verification and settlement

1. *Registration.* Register the estate perimeter on the Eagor-Core registry with personal data removed.

2. *Verification and double-counting check.* The macro-layer ingests VIIRS and MODIS active-fire products and Sentinel-1 SAR imagery and cross-checks registry records (Section 4.6).

3. *Disbursement.* After 30 consecutive days without verified fire detections and with verified production records, smart contracts disburse credits to the multi-signature wallets of the estate and of the smallholder cooperative.

A.6 Simulator state variables and balance equations

Table A2 summarizes the state variables of the simulator and the structure of their balance equations, which complete the transition kernel P of the Dec-POMDP in Section 4.2.

Table A2. State variables and balance equations of the simulator

State variable	Balance equation (structure)	Main drivers
Biomass density B	$\frac{dB}{dt} = (\mu - m - D)B - \frac{q^{POME} + q^{rain} - q^{evap}}{V} B$	Light, temperature, pH, toxicity, dilution
Pond volume V	$\frac{dV}{dt} = q^{POME} + q^{rain} - q^{evap} - q^{H}$	Effluent feed, rain, evaporation, harvest
Dissolved inorganic carbon C	$\frac{dC}{dt} = \frac{\eta_{CO_2} q^{CO_2}}{V} + k_L a\,(C^* - C) - Y_C \mu B - \frac{q^{in}}{V} C$	CO_2 injection, gas exchange, uptake
pH	Algebraic: carbonate equilibrium given C, alkalinity, and temperature	Inorganic carbon, alkalinity
Dissolved oxygen O	$\frac{dO}{dt} = Y_O \mu B - r_R B + k_L a_O\,(O^* - O) - \frac{q^{in}}{V} O$	Photosynthesis, respiration, mixing
Nitrogen N, phosphorus P	$\frac{dN}{dt} = \frac{q^{POME}}{V}(N_{in} - N) - Y_N \mu B$ (P analogous)	Effluent composition, uptake
Toxicity index χ	$\frac{d\chi}{dt} = \frac{q^{POME}}{V}(\chi_{in} - \chi) - k_\chi \chi$	Phenolics and metals in feed, degradation
Probe drift δ, fouling φ	Equations (6) and (7)	Calibration, backwash

Note. $k_L a$ = gas–liquid mass-transfer coefficient, which increases with paddlewheel speed; Y = yield coefficients; η_{CO_2} = CO_2 transfer efficiency; r_R = respiration rate; asterisks denote saturation concentrations; $q^{in} = q^{POME} + q^{rain} - q^{evap}$ is the net dilution flow. Parameter values and their sources are released with the simulator code.

Appendix B. Policy Brief: Bilateral Co-Investment in a 10-Hectare Demonstration Facility

To: Ministry of Sustainability and the Environment (MSE), National Environment Agency (NEA), and Economic Development Board (EDB), Singapore

From: Lanz CWJ Chan (Finamatrix Quant-Lab, Singapore) and Shin-Hung Pan (Chaoyang University of Technology, Taiwan)

Subject: Bilateral co-sponsorship of a 10-hectare Eagor-Core demonstration facility in Indonesia to reduce transboundary haze

B.1 Rationale

Transboundary haze imposes recurring health and economic costs on Singapore that originate outside its jurisdiction (Koplitz et al., 2016; Glauber et al., 2016). Section 3 shows that conditional payments make non-burning privately rational only when verification is credible and the payment covers the cost advantage of fire. A demonstration facility would test, under field conditions, whether mill-level algae revenue and three-tier verification can meet that condition. This brief is based on a conceptual framework and assumption-based engineering estimates. It is not a field-validated investment proposal, and field performance and carbon benefits have not been demonstrated.

B.2 Bilateral alignment

Public health. Lower fire activity in participating estates reduces one source of haze affecting Singapore. The demonstration would measure the avoided burned area within participating geofences rather than claim regional elimination of haze.

Sustainable aviation fuel. Singapore has set a 1% SAF uplift target, with the SAF levy applying to flights departing from 1 January 2027, and aims to raise the target to 3–5% by 2030, subject to global developments (Civil Aviation Authority of Singapore, 2024, 2026). CAO from POME could supply HEFA feedstock if it meets refiner specifications and the sustainability criteria Singapore applies to SAF (Doliente et al., 2020).

High-integrity carbon credits. Singapore and Indonesia signed an MoU on carbon credits collaboration under Article 6 of the Paris Agreement on 6 July 2026 (Ministry of Trade and Industry, 2026). Satellite- and drone-verified zero-burn records linked to registry entries would allow credits from the facility to be checked against double-counting (Schneider et al., 2015).

Institutional fit. The MoU on Environmental Cooperation, signed on 29 June 2026, covers transboundary pollution, climate change, and joint pilot projects (Ministry of Sustainability and the Environment, 2026), providing a framework under which the demonstration could be endorsed.

B.3 Proposed phasing

The phasing follows Figure 5 and Section A.3: bilateral endorsement and a pilot module (months 1–6); a 10-hectare demonstration in Riau or Central Kalimantan with smallholder onboarding and drone and satellite verification (months 6–18); and private roll-out to other mills once the

demonstration meets its go/no-go criteria (months 18–36). The capital requirement of the demonstration is USD 439,300, and its annual operating cost is USD 93,000 (Tables 3 and 4).

B.4 Requests

We request (i) co-funding of the demonstration CAPEX, (ii) facilitation of a host-site agreement with the relevant Indonesian ministries and a participating mill, and (iii) agreement on data-sharing arrangements that comply with both countries' personal data protection laws (Republic of Indonesia, 2022; Republic of Singapore, 2012).