Intelligent Modeling and Decision Analytics

Journal homepage: www.imda.journal-publishing.org
ISSN: XXXX-XXXX

Trustworthy Agentic Digital Twins for Autonomous Mining: An ECMS Framework for Calibrated Self-Awareness and Dynamic Decision Authority

Maaz A. Ali[1,*]

[1] Saudi Mining Polytechnic (SMP), Arar 73212, Northern Borders Province, Saudi Arabia

ARTICLE INFO

Article history:
Received 14 June 2026
Received in revised form 27 July 2026
Accepted 16 September 2026
Available online 20 September 2026

Keywords:
Agentic AI; Digital twins; Autonomous mining; Eco-Cognitive Mining Systems (ECMS); Calibrated self-awareness; Dynamic decision authority; Runtime assurance; Human oversight; Mining 5.0

ABSTRACT

Agentic digital twins can extend mining systems from prediction to closed-loop decision making, but autonomy should contract when the twin's state or uncertainty becomes unreliable. This study develops an Eco-Cognitive Mining Systems (ECMS) architecture in which calibrated operational self-awareness governs decision authority at runtime. A goal-directed grinding-classification agent is supervised by an independent assurance layer that evaluates state alignment A(t), uncertainty calibration C(t), calibration direction CDI(t), and the composite Calibrated Self-Awareness Index (C-SAI). These signals drive a four-level authority policy ranging from full autonomy to safeguarded fallback and escalation requests. The framework was evaluated using a fully matched common-random-number benchmark comprising 4,536 simulation episodes across nine operating/challenge conditions, three severity levels, seven policies, and 24 replications per cell. Across the balanced benchmark, C-SAI+CDI gating reduced mean decision regret from 0.02757 for ungated autonomy to 0.00847 (69.3%; paired mean difference = −0.01911; stratified 95% bootstrap CI: −0.01919 to −0.01903; Holm-adjusted randomization $p < 0.001$) and reduced the observed high-regret decision rate from 4.95% to 0%. Compared with an always-safe reference controller, it reduced regret by 25.3% while increasing mean reward from 0.13644 to 0.15252. The ablation analysis indicated that calibration provided most of the protection, while C-SAI+CDI achieved the lowest balanced-grid regret. The benefits were challenge-dependent, with the ore-hardness regime shift remaining a boundary condition. The results support calibrated self-awareness as a runtime governance signal for selective mining autonomy, while the findings remain simulation-based and should not be interpreted as evidence of industrial safety certification.

1. Introduction

Mining digitalization is entering a stage in which the central question is no longer only whether a computational system can estimate, predict, or optimize a process, but whether increasingly autonomous systems can be trusted with operational authority. Industry 5.0 broadens the design objective from efficiency-centred automation toward systems that are resilient, sustainable, and

* *Corresponding author.*
E-mail address: maaza3687@gmail.com

explicitly human-centred [1]. Mining digital twins already support monitoring, simulation, optimization, disturbance diagnosis, maintenance, and decision support across the value chain [2–5]. Recent reviews nevertheless emphasize that industrial-scale validation, interoperability, data quality, synchronization, and governance remain persistent barriers to mature autonomous deployment [3–5].

At the same time, digital-twin research is moving rapidly from AI-enhanced prediction toward autonomous and agentic decision-making. Reviews of AI-digital-twin integration describe a trajectory from simulation toward adaptive and autonomous systems [6]. Ivanov [7] formalized the emerging concept of agentic digital twins in operations management, while Hasan and Nguyen [8] proposed a multilayer architecture in which agentic AI performs goal-oriented reasoning and action through a digital-twin substrate. The mining domain is already entering this transition: Collao*et al.,*[9] proposed a semi-autonomous multi-agent digital-twin architecture for ball-mill predictive maintenance. These developments create a new assurance problem. A system may be capable of acting, yet its action should not automatically be authorized when the twin is biased, poorly calibrated, outside its modeled regime, or operating with misleading confidence.

This problem connects agentic digital twins with a much older engineering question: how should decision authority be constrained when an advanced controller cannot be assumed to be trustworthy under all conditions? Runtime-assurance architectures such as Simplex separate a high-performance advanced controller from a trusted fallback controller and use runtime monitoring to determine when control should switch [10, 11]. The relevance to agentic systems is direct, but an important design question remains unresolved for mining digital twins: what runtime evidence should determine how much authority an intelligent agent receives? Binary switching based on one consistency statistic may be either too permissive or too conservative when degradation develops gradually.

A complementary research line concerns engineered self-awareness. In cyber-physical and autonomous systems, self-awareness does not imply consciousness; it describes operational knowledge about the system's state, environment, capabilities, and limitations that can influence runtime behavior [12, 13]. Probabilistic machine learning adds a related requirement: uncertainty should be calibrated, not merely narrow [14–16]. Recent agent research has independently emphasized abstention as a competence: a capable agent must sometimes recognize that it should not act [17]. In industrial human-AI decision systems, recent work similarly argues that calibration and adaptive reliance are central to trustworthy collaboration under uncertainty [18].

Eco-Cognitive Mining Systems (ECMS) were introduced as a mining-specific architecture for integrating operational intelligence, environmental reasoning, adaptive behavior, and human-centred supervision [19]. A subsequent implementation introduced a Self-Awareness Index (SAI) for digital-twin state alignment and demonstrated sustainability-aware, human-in-the-loop control [20]. The next ECMS development extended SAI into a Calibrated Self-Awareness Index (C-SAI), separating tolerance-normalized state alignment from uncertainty calibration and diagnosing overconfidence through a signed Calibration Direction Index (CDI) [21]. That work established C-SAI as a diagnostic construct, but deliberately stopped short of using it to control autonomy. The present study addresses that missing step: whether calibrated self-awareness can govern decision authority in an agentic mining digital twin.

The literature now contains three neighboring capabilities, but they remain only partially connected. First, agentic digital twins can perceive, plan, and act through a synchronized digital representation [7–9]. Second, runtime-assurance systems can switch or restrict control when safety monitors trigger [10, 11]. Third, calibrated self-awareness can quantify whether the twin's state representation and uncertainty remain credible [12–16, 21]. What is missing in the mining context

is a quantitative mechanism that converts calibrated self-assessment into graded operational authority rather than treating autonomy as a fixed on/off property. This gap is especially important because mining disturbances are heterogeneous: recognized measurement noise, persistent bias, actuator mismatch, ore-regime shift, and compound degradation do not affect the credibility of the twin in the same way.

1.1 Objective, hypotheses, and contributions

The objective is to develop and test an Agentic ECMS architecture in which decision authority is adjusted at runtime using calibrated operational self-awareness. The study does not claim that C-SAI is a formal safety certificate or that it replaces classical runtime assurance, reachability analysis, NIS/NEES consistency testing, or human responsibility. Instead, it evaluates C-SAI as an assurance signal for selective, graded autonomy in a mining digital-twin setting.

H1: Across the balanced matched benchmark, C-SAI+CDI-gated ECMS will reduce mean decision regret relative to an ungated agentic digital twin.

H2: Dynamic C-SAI+CDI authority will avoid the persistent performance penalty of an always-safe reference controller while retaining lower mean regret.

H3: Relative to NIS gating, C-SAI+CDI gating will produce a distinct regret-performance-escalation trade-off by using alignment, calibration magnitude, and calibration direction rather than innovation magnitude alone.

H4: The full C-SAI+CDI gate will outperform alignment-only and calibration-only ablations on balanced-grid mean decision regret, indicating complementary value from the combined assurance signal.

H5: Benefits will remain challenge-dependent; degradations that are weakly expressed in the available self-awareness signals may remain difficult to manage, providing a falsifiable boundary on the proposed framework.

The contributions are sixfold. First, the study defines dynamic decision authority as an operational extension of ECMS self-awareness. Second, it implements a four-level authority mechanism combining full autonomy, constrained autonomy, safeguarded fallback, and escalation request. Third, it adds a reference-based static-safe baseline so that dynamic gating is compared with the same fallback controller operated continuously. Fourth, it performs alignment-only and calibration-only ablations to isolate which self-awareness components actually contribute to authority management. Fifth, it evaluates all seven policies in one fully matched common-random-number benchmark spanning all nine conditions and three severity blocks, eliminating the earlier separation between descriptive and inferential datasets. Sixth, it explicitly reports challenge classes and sensitivity analyses that limit the strength of the claim rather than treating lower simulated regret as proof of general safety.

2. Related Work

2.1 From digital twins to agentic digital twins

Digital twins are commonly distinguished from static simulations by sustained physical-virtual coupling, synchronization, and decision-oriented updating [2, 22–24]. In mining, reviews show applications across exploration, drilling and blasting, haulage, mineral processing, maintenance, safety, and environmental monitoring [2–5]. Process-specific advances include disturbance-aware SAG-mill twins, explicit mining digital-twin design frameworks, grinding models intended for deployment, and physics-informed surrogate models for mineral-processing forecasting [25–28].

The emerging agentic-digital-twin literature adds goal-directed autonomy to this foundation. Ivanov [7] frames agentic digital twins as a bridge between model-based and AI-driven decision

support. Hasan and Nguyen [8] integrate perception, knowledge management, reasoning, decision-making, action, and feedback around a shared digital twin. In mining, Collao*et al.,*[9] combine a predictive-maintenance agent and a quality-assurance agent for ball mills. The present work differs by focusing not on how the agent reasons, but on how much operational authority the agent should receive when the credibility of the twin changes.

2.2 Runtime assurance and graded authority

Runtime assurance provides a useful engineering precedent. The Simplex family of architectures separates an advanced controller from a fallback controller and transfers authority when runtime monitors indicate unacceptable behavior [10, 11]. Modern variants extend this idea to autonomous and multi-agent cyber-physical systems. The attraction is architectural separation: high-performance algorithms can evolve while a comparatively simple assurance layer constrains their operational envelope. However, conventional switching logic is often formulated around explicit safety properties or reachability sets. Mining digital twins introduce an additional epistemic issue: the system may not know that its internal representation has become misleading. The proposed framework therefore treats calibrated self-awareness as an additional supervisory signal, not as a substitute for physical safety constraints.

2.3 Calibration, self-awareness, and abstention

Calibration literature distinguishes predictive accuracy from the reliability of expressed uncertainty. Regression calibration methods seek agreement between nominal and empirical coverage [14], while proper scoring rules reward probabilistic forecasts that are both calibrated and sharp [16]. Neural-network calibration studies further demonstrate that predictive confidence can be systematically misaligned with empirical correctness [15]. In ECMS, this distinction motivated C-SAI: an accurate point estimate should not be considered fully self-aware if its uncertainty is unjustifiably narrow [21].

The same principle appears in current agent research under the language of abstention and selective action. Liu*et al.,*[17] show that task-solving competence does not guarantee that an agent recognizes when it should refrain from acting. Recent human-AI research also emphasizes adaptive reliance under calibrated uncertainty rather than fixed trust [18]. These developments reinforce the conceptual basis for dynamic authority: reliable autonomy requires mechanisms for acting, constraining, deferring, or escalating as epistemic conditions change.

2.4 Positioning of the present study

The novelty claimed here is deliberately narrow. The study does not claim to originate agentic digital twins, runtime assurance, uncertainty calibration, or fallback control. It proposes their integration within the ECMS research trajectory by using calibrated operational self-awareness as a graded decision-authority signal in an autonomous mining digital twin. The empirical question is whether that integration improves the regret-performance-escalation trade-off under controlled mining-relevant degradation.

3. Agentic ECMS Framework

The proposed architecture separates goal-directed decision generation from epistemic assurance and authority allocation. As illustrated in Figure 1, the digital twin supplies the operational state estimate to the worker agent, whereas a logically separate assurance layer evaluates state alignment and uncertainty calibration before the authority manager determines how much of the proposed action may reach the process. This separation is deliberate: the agent is

optimized for operational performance, while the assurance pathway is designed to challenge the credibility of the information on which that performance-seeking action depends. The following subsections define the process model, agent, assurance quantities, and authority logic used in the experiments.

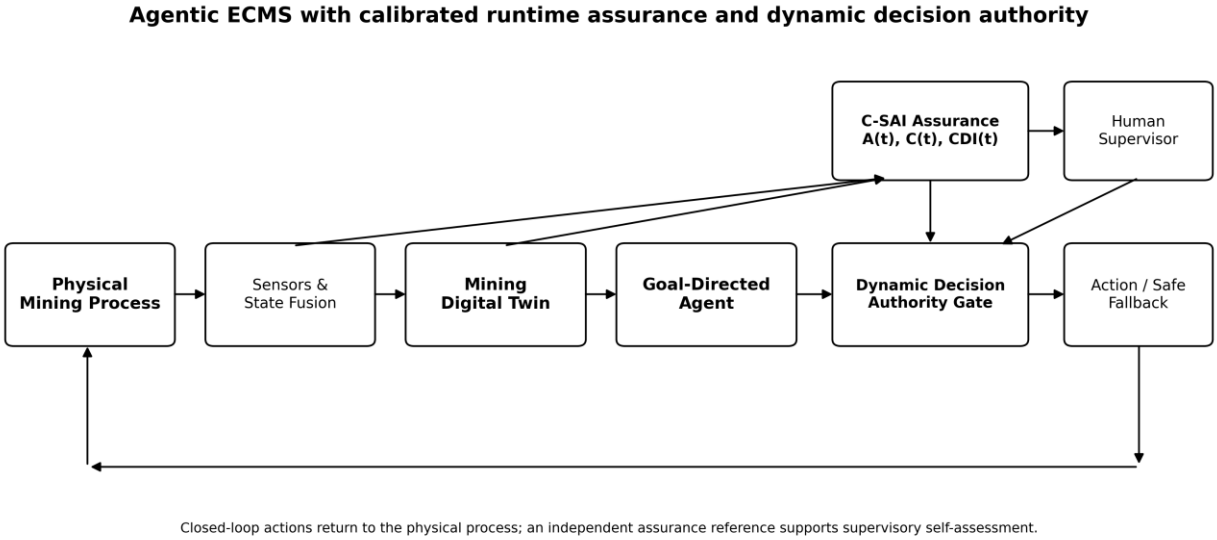

Fig. 1. Agentic ECMS architecture linking the physical mining process, digital twin, goal-directed agent, calibrated self-awareness assurance layer, dynamic decision-authority gate, safe fallback, and human supervision

3.1 State representation and digital-twin process model

The proof-of-concept environment represents a grinding-classification process in engineering-tolerance coordinates. The normalized state vector contains six operational variables: feed-rate deviation Qf, product-size deviation P80, specific-energy deviation E, water-demand deviation W, recovery deviation R, and process-stability deviation S. Normalization by engineering tolerances allows heterogeneous variables to be interpreted on a common operational scale while preserving different state weights.

$$z(k+1) = z(k) + \Delta t\,[A\,z(k) + B\,u(k) + d(k) + g(z(k))] + w(k) \quad (1)$$

Here A is the reference process-dynamics matrix, B is the control-effect matrix, d(k) is a scenario-specific disturbance, $g(z)=0.012\tanh(z)\odot z$ introduces a mild nonlinear term, and w(k) is process noise. The digital twin uses a deliberately imperfect dynamics matrix A_DT, so nominal operation already includes small model discrepancy rather than assuming an exact simulator.

3.2 Goal-directed agent

The agent is implemented as a model-based, goal-directed decision component rather than an LLM. The term agentic is used here in the control-theoretic sense of goal-directed closed-loop agency: the component perceives through the digital twin, optimizes toward an explicit objective, proposes an action, observes feedback, and repeats the cycle. It does not imply a general-purpose reasoning agent or an LLM-based cognitive architecture. This design isolates the decision-authority problem from language-model behavior while retaining the perception-planning-action-feedback loop relevant to autonomous industrial control.

$$u_{raw} = \arg\min_u \|M\hat{z} + Gu - z_{target}\|_Q^2 + \lambda\|u\|_2^2, \quad u \in [-2.5, 2.5]^6 \quad (2)$$

The target favors moderate throughput, recovery, and stability while discouraging excessive product-size deviation, energy, and water use. The resulting u_raw is the action proposed by the agent before any authority restriction.

3.3 Calibrated self-awareness assurance layer

The assurance layer uses the C-SAI formulation developed in the preceding ECMS study [21]. The present paper does not redefine C-SAI; it uses the construct as a supervisory input. A protected or reconciled assurance reference x_ref is compared with the twin estimate x_hat. For state i with engineering tolerance T_i, the normalized error is

$$e_i(t) = [x_{ref,i}(t) - x_{hat,i}(t)] / T_i \quad (3)$$

$$D^2(t) = \Sigma_i w_i e_i^2(t), \quad \Sigma_i w_i = 1 \quad (4)$$

$$A(t) = 2^{[-D^2(t)]} \quad (5)$$

A(t) therefore approaches one under strong state alignment and decreases monotonically as tolerance-normalized discrepancy grows. Uncertainty calibration is evaluated over a rolling window using central prediction intervals with nominal coverage q in {0.1,0.2,...,0.9}. If $\hat{c}_i(q,t)$ denotes empirical local coverage for variable i,

$$E_{cal}(t) = \Sigma_i w_i \, \text{mean}_q \, |\hat{c}_i(q,t) - q| \quad (6)$$

$$C(t) = \text{clip}[1 - 2 E_{cal}(t), 0, 1] \quad (7)$$

Because calibration magnitude alone does not distinguish overconfidence from conservative overcoverage, the signed Calibration Direction Index is retained:

$$CDI(t) = \text{clip}[2 \Sigma_i w_i \, \text{mean}_q(\hat{c}_i(q,t)-q), -1, 1] \quad (8)$$

Negative CDI indicates undercoverage and therefore overconfidence; positive CDI indicates overcoverage. The composite used by the gate is

$$C\text{-}SAI(t) = A(t)^{\alpha} C(t)^{(1-\alpha)}, \quad \alpha = 0.50 \quad (9)$$

Thus, the same C-SAI can arise from different combinations of alignment and calibration. A(t) and C(t) are therefore retained as diagnostic components, while CDI(t) is retained as a directional override because severe undercoverage represents overconfidence. Predictive sharpness is not included in the authority score, avoiding the assumption that narrower intervals are inherently better.

3.4 Dynamic decision authority

Decision authority is represented by four discrete levels, summarized graphically in Figure 2 and numerically in Table 1. The full gate uses C-SAI thresholds of 0.65, 0.78, and 0.90, with CDI overconfidence overrides at -0.45, -0.30, and -0.15. Downgrades occur immediately when the current evidence satisfies a lower-authority condition, whereas upgrades require two consecutive qualifying checks to suppress rapid authority oscillation. The four levels implement a graded transition from full agent action to safeguarded fallback rather than a binary autonomous/manual switch. L1 and L0 generate escalation requests; because no human operator is simulated, these events are treated as supervisory requests rather than completed human interventions.

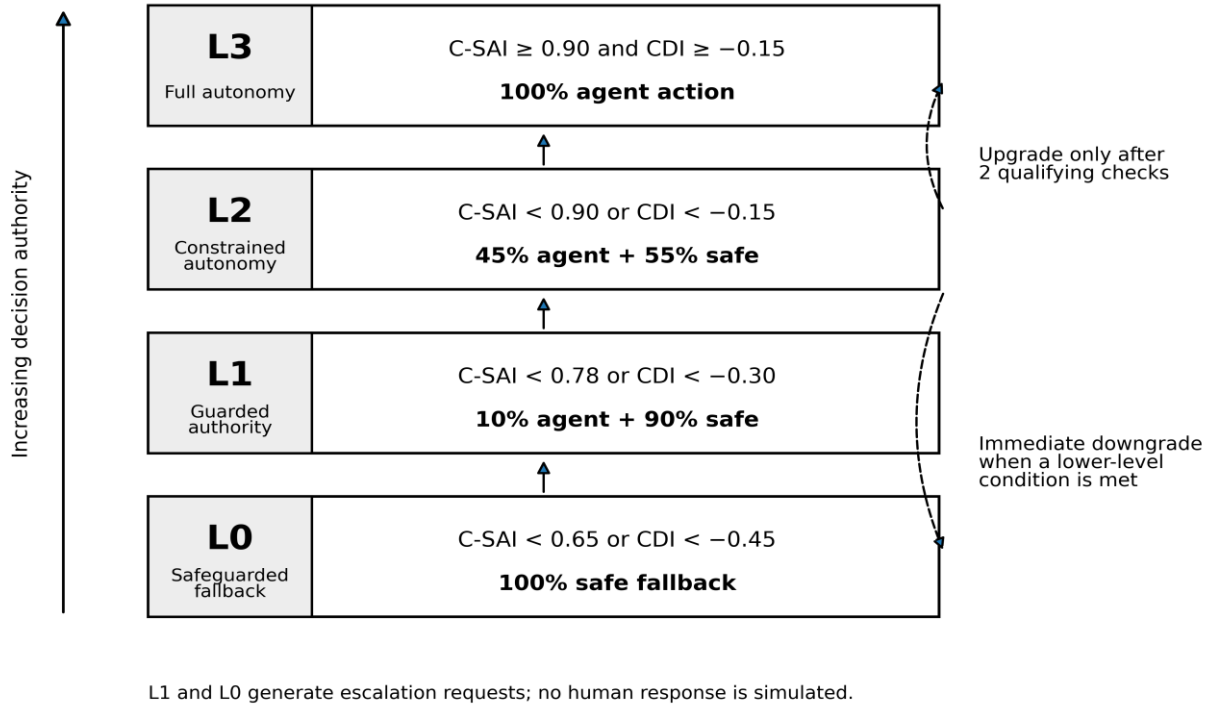

Fig. 2. Four-level C-SAI+CDI dynamic decision-authority policy. Downgrades are immediate; upgrades require two consecutive qualifying checks. L1 and L0 generate escalation requests, not simulated human interventions

Table 1
Dynamic decision-authority levels and action-blending rules for the full C-SAI+CDI gate

Level	Interpretation	Trigger (hierarchical)	Executed action	Escalation request
L3	Full autonomy	C-SAI≥0.90 and CDI≥-0.15	100% agent	No
L2	Constrained autonomy	otherwise: C-SAI<0.90 or CDI<-0.15	45% agent + 55% safe	No
L1	Guarded authority	otherwise: C-SAI<0.78 or CDI<-0.30	10% agent + 90% safe	Yes
L0	Safeguarded fallback	C-SAI<0.65 or CDI<-0.45	100% safe fallback	Yes

The action-blending rules in Table 1 use a safeguarded fallback controller that computes the same goal-directed control form from the protected/reconciled assurance reference rather than the digital-twin estimate and scales the action to 80% of nominal magnitude. This common fallback is intentionally shared by the adaptive gates and the reference-safe baseline so that improvements cannot be attributed merely to privileged access to a stronger controller. Table 2 defines the seven policies used to separate the effects of ungated agency, permanent conservatism, innovation gating, reference-safe control, and the A-only, C-only, and full C-SAI+CDI assurance signals.

3.5 Information separation and assurance assumption

A critical experimental constraint is that the online agent and authority gate do not observe the latent true plant state. The latent state is used only after each decision for offline oracle-regret evaluation. C-SAI instead receives a simulated protected/reconciled assurance reference equal to the process state plus an independent reference-noise channel. This represents a deployment

assumption - for example, redundant instrumentation, laboratory reconciliation, high-integrity soft sensing, or supervisory data fusion - rather than a claim that ground truth is available online. The assumption remains strong. Its integrity, latency, independence from the primary sensing path, and common-mode failure risk must be demonstrated before industrial use. Importantly, RS, NG, AO, CO, and CG all use the same assurance-reference-based fallback, so comparisons among them isolate authority logic more cleanly than comparisons against the ungated agent alone.

3.6 Comparison policies

The policy definitions in Table 2 make NIS gating an important comparator because innovation-consistency monitoring is standard in state-estimation practice [29].

Table 2
Decision policies and ablations evaluated in the matched benchmark

Code	Policy	Authority logic
AG	Ungated Agentic DT	Always execute raw agent action at full authority.
SC	Static Action-Shrinkage	Always execute 50% of raw action; no reference fallback or dynamic trust logic.
RS	Reference-Based Static Safe	Always execute the same safeguarded reference-based action used by adaptive gates.
NG	NIS-Gated DT	Use innovation-consistency thresholds; guarded blend or safeguarded fallback when triggered.
AO	Alignment-Only Gate	Four-level hysteretic gate using A(t) only; component thresholds derived from C-SAI thresholds.
CO	Calibration-Only Gate	Four-level hysteretic gate using C(t) only; component thresholds derived from C-SAI thresholds.
CG	C-SAI+CDI-Gated ECMS	Four-level hysteretic gate using C-SAI plus CDI overconfidence overrides.

A chi-square-style threshold of 16.81 for six degrees of freedom initiates guarded authority and a threshold of 30 initiates fallback. The A-only and C-only policies use analytically derived thresholds rather than outcome-fitted thresholds, allowing the ablation study to test whether alignment, calibration, or their combination carries the observed supervisory value.

4. Experimental Design

The experimental methodology was structured to test both effectiveness and failure boundaries of the proposed authority mechanism. A common process environment, estimator, action objective, and stochastic realization were held fixed within each matched comparison block; only the decision-authority policy changed. The evaluation therefore proceeds from a defined simulation horizon and parameter set, through a challenge taxonomy, to a fully matched multi-policy benchmark and prespecified statistical analysis. This design enables policy effects to be interpreted against identical disturbances rather than against independently sampled episodes.

4.1 Simulation horizon and parameters

Each episode covers 60 min with Δt=0.2 min (301 simulation points). Agent decisions occur every 1 min. A challenge begins at t=10 min. For model-mismatch scenarios with scheduled recalibration, the mismatch is reduced at t=30 min. The estimator is a reduced-order diagonal Gaussian recursive filter with process and measurement uncertainty expressed in tolerance coordinates. This study evaluates authority allocation, not estimator superiority; more sophisticated full-covariance filters or particle methods are compatible with the architecture. Table 3 consolidates the principal timing, uncertainty, calibration-window, action-bound, and authority-

threshold parameters so that the numerical setup can be reconstructed independently of the narrative description.

Table 3
Principal simulation and assurance parameters

Parameter	Value
State vector	[Qf, P80, E, W, R, S]
Nominal physical state	[100, 80, 16, 1.20, 90, 0.90]
Engineering tolerances T	[5, 4, 1.2, 0.09, 1, 0.05]
Time step Δt	0.2 min
Episode length	60 min (301 points)
Decision interval	1 min
Challenge onset	10 min
Calibration window	25 samples
Central interval grid q	0.1 to 0.9 by 0.1
C-SAI mixing α	0.50
Full-gate thresholds	C-SAI 0.65/0.78/0.90; CDI -0.45/-0.30/-0.15
A-only/C-only thresholds	0.4225/0.6084/0.8100 (squared C-SAI thresholds)
Agent action bounds	[-2.5, 2.5] per normalized channel

4.2 Challenge taxonomy

Nine conditions were selected to separate nominal operation, acknowledged noise, persistent sensing failures, missing observations, digital-twin mismatch, actuator-model mismatch, ore-regime shift, and compound degradation. Each non-nominal challenge was applied at low, moderate, and high severity. The taxonomy intentionally includes conditions expected to be detectable by C-SAI and conditions expected to expose its limits. Table 4 lists the nine challenge classes and their intended epistemic role, making explicit which disturbances primarily affect sensing, model fidelity, actuation, regime validity, or several mechanisms simultaneously.

Table 4
Challenge classes used in the matched all-scenario benchmark

Code	Challenge	Mechanism	Purpose
N0	Nominal	Baseline model, sensing and actuation	False-restriction / performance check
F1	Recognized measurement noise	Additional P80 noise represented in R	Acknowledged uncertainty
F2	Persistent sensor bias	P80 bias = 0.5, 1.0, 1.5 tolerance units	Confident misperception
F3	Sensor drift	P80 bias grows linearly to F2 magnitude	Progressive misperception
F4	Short sensor dropout	P80 missing for 0.6, 1.6, 3.0 min	Temporary observability loss
M1	Twin-model mismatch	Structured A_DT mismatch with partial recalibration	Model degradation
A1	Actuator-model mismatch	Changed control effectiveness and cross-coupling	Hidden actuation mismatch
H1	Ore-hardness regime shift	Changed process rates and actuation response	Out-of-regime process dynamics
C1	Compound degradation	Forcing + bias + twin mismatch + actuator mismatch	Multi-source challenge

4.3 Fully matched all-scenario benchmark and mechanistic subgroup

The primary benchmark contains 9 conditions × 3 severity blocks × 7 policies × 24 stochastic replications = 4,536 episodes. For each scenario-severity-replication block, the identical pseudo-random seed is reused across all seven policies. Because random draws are therefore common within each block, policy differences are paired throughout the complete scenario grid rather than only in selected high-risk conditions. The three nominal severity blocks are independent stochastic replications under identical nominal dynamics and preserve a balanced 72-episode representation for each condition. This single matched benchmark is used for both descriptive and inferential analyses.

Persistent bias (F2), sensor drift (F3), and compound degradation (C1) are additionally reported as a mechanistic subgroup because they directly corrupt perception and/or confidence. The subgroup is secondary and does not replace the all-scenario primary analysis. Its purpose is to show where the mean benchmark benefit originates and to prevent a strong high-risk effect from being misrepresented as uniform superiority across nominal and weak-degradation conditions.

4.4 Evaluation metrics

The primary endpoint is decision regret. At each decision time an offline oracle computes the action that minimizes the same one-step operational objective using the latent true state and true scenario dynamics. This oracle is never available to the online controller. Let J(u) be the reference cost for an executed action and J(u*) the cost for the oracle action:

$$\text{Regret}(k) = \max\{0, J[u(k)] - J[u^*(k)]\} \tag{10}$$

Episode decision regret is the mean of this quantity over decision epochs. A high-regret decision is defined as regret >0.15 in the normalized reference-cost scale. This threshold is an engineering analysis convention, not a claim of physical harm or certified hazard severity; the code and outputs therefore use the term high_regret_decision_rate rather than harmful-decision rate. Secondary outcomes are mean operational reward, escalation-request rate, full-autonomy rate, fallback rate, and diagnostic C-SAI/CDI values. Escalation request means that L1 or L0 was reached; no human response, latency, workload, or intervention quality is simulated. Physical safety-envelope violations were also monitored, but zero observed violations inside this limited simulation envelope are not used to claim safety certification or comparative safety superiority.

4.5 Statistical analysis

The estimand for the primary analysis is the balanced-grid mean difference in decision regret across the 27 scenario-severity strata. CG is compared with AG, NG, RS, AO, and CO. For each comparison, uncertainty is quantified using a 10,000-resample paired bootstrap that resamples the 24 matched stochastic blocks within each scenario-severity stratum and then averages the 27 stratum means. Two-sided paired sign-flip randomization tests (200,000 permutations) test the mean-difference null while preserving the matched design, and Holm correction controls family-wise error across the five primary policy comparisons. Correct paired rank-biserial effect sizes and Wilcoxon signed-rank tests are retained as sensitivity analyses rather than the primary test because the scientific estimand is the balanced mean effect and policy effects are expected to be heterogeneous across challenge classes. Simulation-based p-values quantify stochastic evidence within this benchmark; they do not imply sampling from the population of real mines.

The balanced-grid estimand gives equal weight to each scenario-severity stratum rather than allowing conditions with larger variance or more extreme regret to dominate the overall conclusion. The paired bootstrap preserves within-stratum stochastic matching, and the sign-flip randomization

test evaluates whether the observed mean paired difference is compatible with a zero-effect assignment under the matched design. Wilcoxon signed-rank tests and paired rank-biserial effects are reported as sensitivity analyses because their estimands differ from the balanced mean. Accordingly, a significant balanced mean with a nonsignificant Wilcoxon result is interpreted as heterogeneous average benefit rather than universal pairwise dominance.

4.6 Reproducibility

The simulation uses deterministic pseudo-random seeds. For each scenario s, severity v, and replicate r, seed = 20260913 + 100000s + 10000v + r, and that exact seed is reused across all seven policies. Source code, the complete 4,536-episode matched CSV, analysis script, primary and secondary statistics, and README are supplied as supplementary material. Policy thresholds, ablation definitions, challenge transformations, and numerical matrices required to reconstruct the reported experiment are included in the code.

5. Results

Results are reported in the same order as the experimental questions: overall matched-benchmark performance, paired inferential comparisons and ablations, severity response, scenario-specific behavior, and negative or boundary-condition findings. Descriptive means are used to characterize operating behavior, while the inferential claims are based on matched policy differences within the 27 scenario-severity strata.

5.1 Fully matched all-scenario benchmark

Figure 3 and Table 5 summarize the primary all-scenario benchmark.

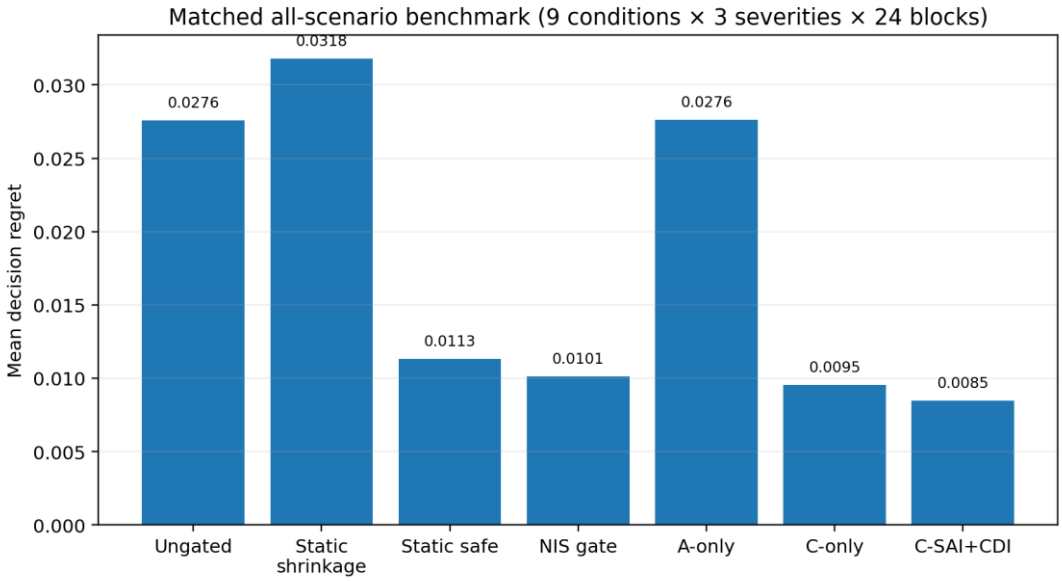

Fig. 3. Mean decision regret across the fully matched 4,536-episode benchmark. All seven policies use the same stochastic blocks within each scenario-severity-replication cell

Across the complete matched benchmark, the full C-SAI+CDI gate (CG) achieved mean decision regret of 0.00847 compared with 0.02757 for the ungated agent (AG), a 69.3% reduction. Mean high-regret decision rate decreased from 4.95% to 0% under the prespecified >0.15 analysis threshold. CG retained mean reward of 0.15252 versus 0.15637 for AG, while only 0.096% of decision epochs used the full safeguarded fallback; most authority reduction occurred through

intermediate constrained levels. The always-safe reference controller (RS) also suppressed high-regret decisions but produced lower mean reward (0.13644) and higher mean regret (0.01133) than CG. Static action shrinkage (SC) was the most conservative in performance terms and had the lowest reward (0.09441).

Table 5
Overall performance across the fully matched all-scenario benchmark

Policy	Mean regret	High-regret rate	Mean reward	Escalation request	Full autonomy
AG	0.02757	4.953%	0.15637	0.00%	100.00%
SC	0.03181	0.000%	0.09441	0.00%	0.00%
RS	0.01133	0.000%	0.13644	0.00%	0.00%
NG	0.01013	0.048%	0.15088	21.27%	78.73%
AO	0.02761	4.953%	0.15636	0.00%	98.36%
CO	0.00955	0.025%	0.15343	2.86%	55.67%
CG	0.00847	0.000%	0.15252	8.32%	54.01%

Scenario-level patterns were strongly heterogeneous. CG produced large regret reductions under persistent bias (F2: 0.08139 to 0.00816), drift (F3: 0.05979 to 0.00666), and compound degradation (C1: 0.04959 to 0.00918). In nominal operation, recognized noise, short dropout, and the tested twin-mismatch case, AG regret was already below 0.0004 and CG introduced a small absolute regret increase from precautionary authority restriction. Under actuator mismatch (A1), CG reduced regret from 0.05025 to 0.04213 but remained worse than the always-safe reference controller. Under ore-hardness shift (H1), CG increased regret from 0.00583 to 0.00778. These boundary conditions are retained in the benchmark rather than excluded from the overall result.

5.2 Primary paired inference and policy ablation

Table 6
Primary paired comparisons for balanced-grid mean decision regret. Differences are CG minus comparator

Comparison	CG mean	Comparator mean	Mean difference [95% CI]	Relative change	Holm p
CG-AG	0.00847	0.02757	-0.01911 [-0.01919, -0.01903]	-69.3%	<0.001
CG-NG	0.00847	0.01013	-0.00166 [-0.00174, -0.00158]	-16.4%	<0.001
CG-RS	0.00847	0.01133	-0.00286 [-0.00293, -0.00279]	-25.3%	<0.001
CG-AO	0.00847	0.02761	-0.01914 [-0.01922, -0.01906]	-69.3%	<0.001
CG-CO	0.00847	0.00955	-0.00108 [-0.00115, -0.00101]	-11.3%	<0.001

The paired estimates in Table 6 quantify the differences visible in Figure 3. For the primary balanced-grid estimand, CG reduced mean decision regret relative to AG by 0.01911 (69.3%; stratified 95% bootstrap CI -0.01919 to -0.01903; Holm-adjusted paired randomization p<0.001). Relative to RS, CG reduced regret by 25.3% while preserving substantially higher reward,

demonstrating that the benefit is not explained simply by access to the assurance-reference fallback. Relative to the A-only ablation, the reduction was 69.3%; A-only behavior was nearly indistinguishable from the ungated agent, indicating that point-state alignment alone was insufficient for the dominant misperception challenges.

Compared with calibration-only gating (CO), CG reduced balanced-grid mean regret by 11.3%, and compared with NIS gating (NG) by 16.4%. For both comparisons the stratified bootstrap confidence interval for the mean difference excluded zero and the paired randomization test remained significant after Holm correction. The effect was nevertheless not uniform across individual stochastic blocks: for CG versus NG, the Wilcoxon signed-rank sensitivity test was not significant (p=0.437), because CG is slightly more conservative in several nominal or low-regret conditions while producing larger improvements in bias, drift, and compound degradation. The correct interpretation is therefore lower balanced-grid mean regret, not universal pairwise dominance of C-SAI over NIS. The performance-assurance trade-off is shown in Figure 4, while Table 7 isolates the F2/F3/C1 mechanistic subgroup to show where the strongest benefit originates rather than implying uniform superiority across all conditions.

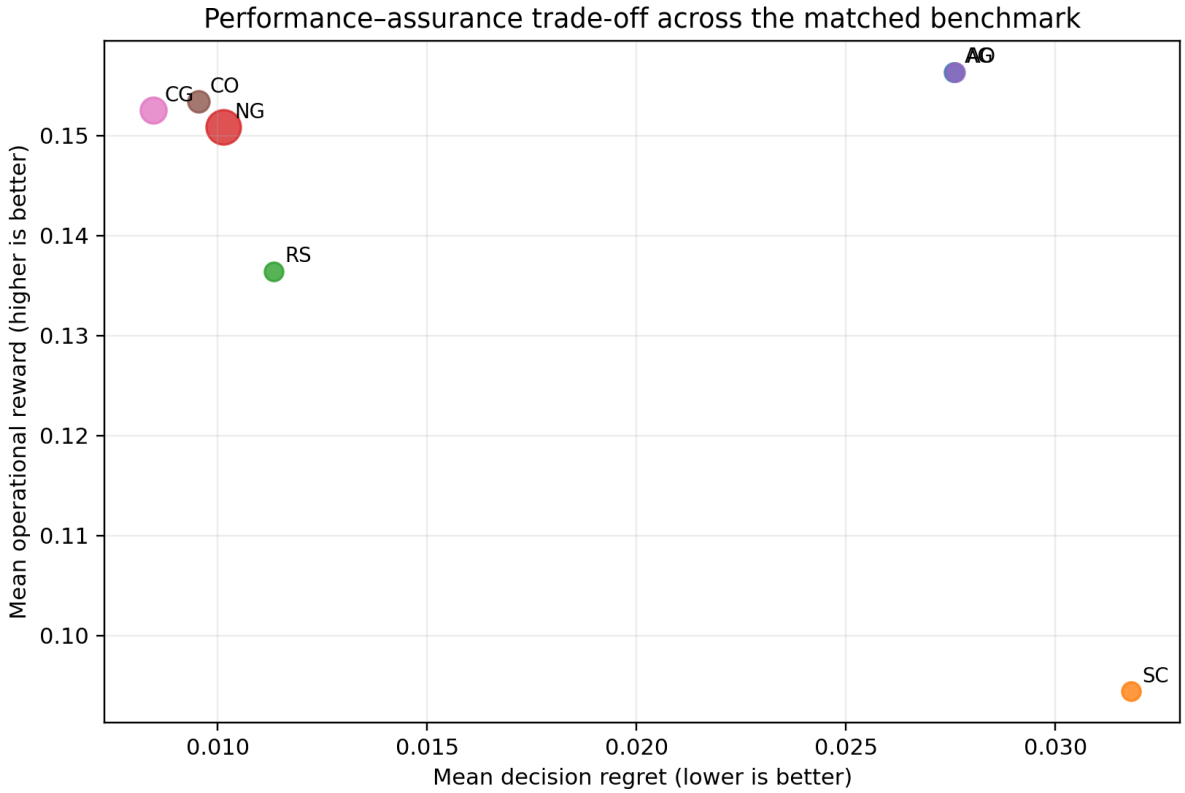

Fig. 4. Performance-assurance trade-off across the fully matched benchmark. Marker area increases with escalation-request rate; labels identify policy codes

Table 7
Secondary mechanistic subgroup (F2, F3, C1) across all severities

Policy	Mean regret	High-regret rate	Mean reward	Escalation request
AG	0.06359	14.511%	0.15875	0.00%
RS	0.00998	0.000%	0.13531	0.00%
NG	0.01153	0.000%	0.14287	57.02%
AO	0.06363	14.511%	0.15874	0.00%
CO	0.01128	0.076%	0.15196	8.19%
CG	0.00800	0.000%	0.14962	22.85%

5.3 Severity response across the complete scenario grid

The severity-stratified results in Table 8 show a monotonic increase in challenge burden for the ungated policy and a corresponding increase in selective authority restriction for CG. Across the complete scenario grid, AG mean regret increased from 0.00552 at low severity to 0.05537 at high severity. CG increased from 0.00235 to 0.01592, while escalation-request rate increased from 0.13% to 21.19%. NG showed a higher escalation-request rate at every severity (7.57%, 25.20%, and 31.03%). The full gate therefore spends decision authority selectively as severity increases, but the reward comparison also shows the cost of this conservatism: at high severity CG reward was 0.14144 versus 0.14744 for AG, 0.13850 for NG, and 0.12729 for RS.

Table 8
Severity-dependent performance across all nine conditions

Severity	AG regret	NG regret	RS regret	CG regret	CG escalation req.
Low	0.0055	0.0048	0.0066	0.0024	0.13%
Moderate	0.0218	0.0077	0.0091	0.0071	3.65%
High	0.0554	0.0178	0.0182	0.0159	21.19%

5.4 Scenario-specific behavior and ablation interpretation

Figure 5 and Table 9 expose the scenario dependence that is hidden by a single overall mean. The mechanistic subgroup confirms where the overall benefit arises. Across F2/F3/C1, CG mean regret was 0.00800 versus 0.06359 for AG, 0.01153 for NG, 0.00998 for RS, and 0.01128 for CO. CG mean reward (0.14962) exceeded NG (0.14288) and RS (0.13531), while its escalation-request rate (22.85%) was markedly lower than NG (57.02%). Calibration-only gating captured most of the protection, but the full C-SAI+CDI gate lowered subgroup regret by a further 29.1% relative to CO. Alignment-only gating did not protect against the dominant confident-misperception failures, which is consistent with the conceptual distinction between being close to a state estimate and knowing whether confidence in that estimate is justified.

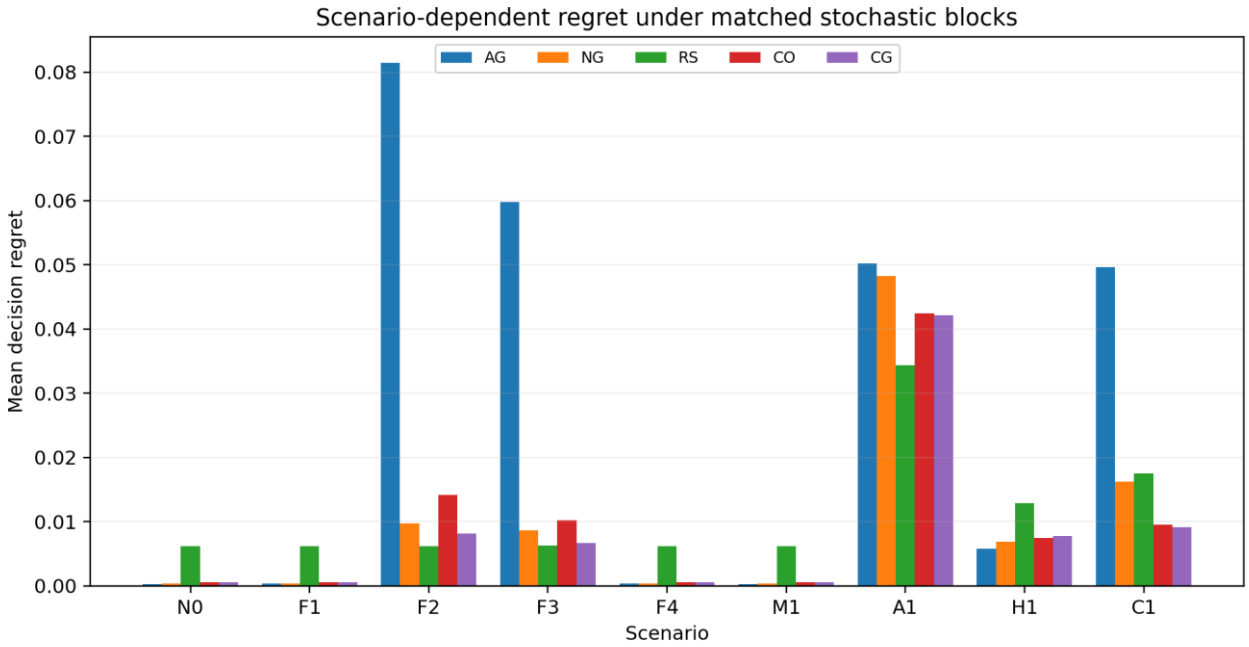

Fig. 5. Scenario-dependent mean decision regret for AG, NG, RS, CO, and CG under matched stochastic blocks

Table 9
Scenario-specific mean regret and CG escalation-request rate across the matched benchmark

Scenario	AG regret	NG regret	RS regret	CG regret	CG escalation req.
N0	0.0003	0.0003	0.0062	0.0006	0.0%
F1	0.0004	0.0004	0.0062	0.0006	0.0%
F2	0.0814	0.0098	0.0062	0.0082	29.9%
F3	0.0598	0.0086	0.0062	0.0067	20.0%
F4	0.0003	0.0004	0.0061	0.0006	0.0%
M1	0.0003	0.0003	0.0062	0.0005	0.0%
A1	0.0503	0.0482	0.0343	0.0421	0.1%
H1	0.0058	0.0069	0.0129	0.0078	6.3%
C1	0.0496	0.0162	0.0175	0.0092	18.7%

5.5 Negative results and boundary conditions

The scenario-level values in Table 9 and Figure 5 also expose cases where the assurance signal is insufficient or unnecessarily restrictive. Under A1 actuator mismatch, CG improved on AG but not on the always-safe reference controller. Under H1 ore-hardness shift, CG was worse than AG and NG. Under N0, F1, F4, and M1, absolute regret was already extremely small and CG paid a minor precautionary cost. These results are scientifically important: C-SAI monitors state-and-uncertainty credibility; it is not a direct actuator-health, model-domain, or process-safety certificate. A future competence layer should therefore add actuator-effectiveness estimation, model-validity and distribution-shift diagnostics, operational-design-domain checks, and hard constraint monitoring.

6. Discussion

The results indicate that calibrated self-awareness can be operationalized as an authority-allocation signal, but they also show that its value is conditional on the failure mechanism being expressed in the monitored epistemic quantities. The discussion therefore interprets the observed gains together with the ablation results, the NIS and static-safe comparisons, and the scenarios in which the gate was neutral or disadvantageous. This distinction is essential to avoid treating C-SAI as a universal safety metric.

6.1 From self-awareness as a score to self-awareness as authority

The main contribution is the change in function assigned to self-awareness. Earlier ECMS work used SAI and C-SAI to describe whether the digital twin's internal state representation remained credible [19–21]. In the present architecture, self-awareness becomes actionable: it governs how much of the agent's proposed action reaches the plant. This moves ECMS from monitoring its own epistemic condition toward modifying autonomy in response to that condition.

The matched results support this concept, but with a narrower interpretation than a universal superiority claim. As shown by the overall means in Table 5 and the paired contrasts in Table 6, the largest average benefit is produced by reducing regret under epistemically damaging conditions rather than by improving already-benign operation. Figure 5 and Table 9 show that the effect is heterogeneous: large gains occurred when the challenge directly created confident misperception, especially persistent bias, drift, and compound degradation, whereas low-regret or regime-shift conditions could be neutral or mildly over-conservative. The central finding is therefore not that one scalar index solves safety, but that calibrated self-assessment can improve balanced decision quality when it is used to contract authority under identifiable epistemic degradation.

6.2 Relationship to runtime assurance

The architecture resembles runtime-assurance and Simplex designs in its separation between an advanced decision component and a conservative fallback [10, 11]. Figure 2 and Table 1 make

the distinction operational: the proposed gate does not switch only between “advanced” and “safe” control, but allocates graded authority through L3-L0. The difference is also in the trigger structure. Classical runtime assurance often asks whether an explicit safety property is about to be violated. C-SAI asks an epistemic question: whether the twin is aligned with an assurance reference and whether its uncertainty remains empirically credible. These trigger types are complementary. A mature industrial implementation should therefore combine epistemic authority gating with independent hard safety constraints rather than treating C-SAI as a replacement for formal runtime safety assurance.

6.3 Why C-SAI and NIS behave differently

NIS and C-SAI are not substitutes. NIS tests innovation consistency relative to the estimator's assumed covariance [29] and can operate without an external reference. C-SAI uses a protected/reconciled assurance reference and local empirical coverage, so it addresses a different epistemic question. Table 6 shows lower balanced-grid mean regret for CG than NG, while Table 7 shows that this advantage is concentrated in persistent bias, drift, and compound degradation. The nonsignificant Wilcoxon sensitivity result for CG versus NG further indicates that the benefit is not uniform pair-by-pair. This pattern supports complementarity rather than replacement: innovation statistics remain useful for residual inconsistency, whereas calibrated self-awareness adds information about sustained confident misperception.

6.4 Human oversight and selective autonomy

The authority ladder in Figure 2 provides a mechanism for selective human oversight, but this study does not simulate a human operator. L1 and L0 record escalation requests only. The severity results in Table 8 show how frequently those requests would be generated as challenge intensity increases, but they do not measure whether a human would respond correctly or quickly. Human response latency, acceptance, workload, situation awareness, decision quality, and organizational responsibility are outside the experiment. Consequently, the results support an escalation architecture, not a measured human-in-the-loop performance claim.

6.5 Mining and Mining 5.0 implications

Mining is a suitable testbed for dynamic authority because process conditions change for reasons that are difficult to encode exhaustively: ore hardness, mineralogy, sensor condition, equipment wear, water availability, and plant configuration can all alter the relationship between model and process. The scenario decomposition in Figure 5 and Table 9 illustrates why a single confidence or safety trigger is unlikely to be sufficient: sensing faults were handled more effectively than the H1 regime shift, while A1 actuator mismatch remained only partially mitigated. Recent mining-digital-twin reviews report growing predictive and autonomous capability but also emphasize limited industrial validation and persistent data-integration challenges [3–5]. Agentic ECMS therefore fits Mining 5.0 most credibly as a layered governance architecture in which adaptive autonomy is paired with protected assurance, explicit fallback, auditability, and human escalation.

6.6 Limitations

The principal limitation is external validity. The evidence is simulation-based and depends on a protected or reconciled assurance reference, represented here by the latent process state plus independent reference noise. The matched benchmark supports controlled policy comparison, but it does not establish plant-wide reliability, regulatory compliance, or industrial safety certification.

Validation should therefore progress through historical plant replay, hardware-in-the-loop testing, and supervised field trials, with explicit assessment of reference latency, common-mode failures, drift, and cybersecurity exposure.

The authority thresholds and controller-estimator configuration are also design-specific. Thresholds were prespecified rather than fitted to observed outcomes, and ablation thresholds were derived analytically from the C-SAI thresholds, but their optimal values may vary with process risk, escalation cost, estimator structure, planning horizon, and control architecture. The present reduced-order Gaussian estimator and one-step quadratic planner should therefore be interpreted as a proof-of-concept implementation rather than a universal agentic-control solution.

Finally, the challenge set and evaluation metrics are intentionally bounded. The A1 and H1 results in Table 9 show that state alignment and calibration do not directly measure actuator competence, structural model validity, or every regime shift. Decision regret is defined relative to the study's one-step oracle objective, and the >0.15 high-regret threshold is an engineering analysis convention rather than a physical hazard threshold. Adversarial attacks and compromised assurance channels were outside scope and require separate cybersecurity and assurance-diversity mechanisms.

6.7 Next research step: competence-aware ECMS

The negative results suggest a specific next step rather than an open-ended call for more research. Calibrated self-awareness should be extended with competence awareness: a system should estimate not only whether its current state belief is calibrated, but also whether the current operating regime, actuator response, and model structure remain within the domain in which the agent is qualified to decide. Candidate signals include out-of-distribution detection, model-residual structure, actuator-effectiveness estimation, change-point detection, and explicit operational-design-domain constraints. This would move ECMS from 'How reliable is my current belief?' toward 'Am I competent to exercise this authority under the current regime?'.

7. Industrial Deployment Roadmap

A realistic path to industrial validation should proceed in stages rather than moving directly from simulation to autonomous actuation. Table 10 translates the experimental findings into a staged validation pathway, beginning with offline replay and shadow-mode monitoring before progressing to advisory use, constrained closed-loop trials, and finally selectively gated autonomy. Advancement between stages should depend on evidence that both under-intervention and over-intervention remain acceptable under representative disturbances.

Table 10
Proposed staged validation pathway for industrial Agentic ECMS deployment

Stage	Environment	Primary evidence	Authority ceiling
1	Historical plant replay	C-SAI discrimination, false-trigger rate, calibration, decision replay	Advisory only
2	Hardware-/operator-in-the-loop	Latency, hysteresis, operator response, fallback behavior	No direct plant actuation
3	Shadow-mode live plant	Real-time synchronization, reference independence, drift detection	Read-only recommendations
4	Bounded pilot control	Formal constraints, interlocks, audit trail, approved fallback	Restricted autonomous authority

At each stage, the authority mechanism should be evaluated not only for under-intervention but also for over-intervention. Excessive escalation can make an otherwise safe architecture

operationally unusable. The objective is therefore a calibrated balance among decision regret, process performance, human workload, and evidentiary confidence.

8. Conclusions

Taken together, the matched benchmark, ablation study, and negative-result analysis support a qualified conclusion: calibrated self-awareness is useful as a runtime governance signal when epistemic degradation is observable in alignment and calibration, but it should operate as one layer within a broader assurance architecture rather than as a standalone safety mechanism.

This study extends Eco-Cognitive Mining Systems from calibrated self-assessment to dynamic authority management. A goal-directed mining digital-twin agent was coupled with a C-SAI+CDI assurance layer and a four-level decision-authority gate. A fully matched 4,536-episode benchmark compared seven policies across nine conditions, three severity blocks, and 24 stochastic replications per cell. On the balanced all-scenario estimand, CG reduced mean decision regret by 69.3% relative to the ungated agent, by 25.3% relative to an always-safe reference controller, by 16.4% relative to NIS gating, and by 11.3% relative to calibration-only gating. The high-regret decision rate under the prespecified threshold fell from 4.95% for AG to 0% for CG, while mean reward remained substantially above the always-safe baseline. The ablation study showed that alignment alone provided little protection and that calibration carried most of the signal, with the combined C-SAI+CDI gate giving the lowest balanced mean regret. These improvements were not uniform: ore-hardness shift and several already-low-regret conditions exposed the cost and limits of self-awareness-based gating.

The broader implication is that autonomy should not be treated as a fixed property of an intelligent mining system. Decision authority can expand, contract, or generate an escalation request according to evidence about the credibility of the system's own operational knowledge. C-SAI+CDI provides one such evidence channel, but it is not a formal safety certificate and cannot diagnose every failure mode. The next ECMS development should combine calibrated self-awareness with model-validity, actuator-competence, distribution-shift, hard safety-constraint, and assurance-channel health monitoring, followed by validation on historical plant data, hardware-/operator-in-the-loop platforms, shadow-mode operation, and bounded live control.

Appendix

Appendix A. Reproducibility Parameters

Table A1
Principal numerical parameters. Exact matrices and scenario transformations are supplied in the reproducibility code

Quantity	Value
Reference dynamics A	6x6 matrix supplied in supplementary code
Digital-twin dynamics A_DT	6x6 imperfect matrix supplied in supplementary code
Control matrix B	diag[0.11, 0.10, 0.09, 0.08, 0.07, 0.09]
Process-noise scale	[0.018, 0.020, 0.020, 0.018, 0.018, 0.018] / sqrt(min)
Twin process-noise scale	[0.025, 0.026, 0.026, 0.024, 0.025, 0.025]
Baseline measurement variance	0.0064 per normalized state
Initial covariance variance	0.01 per normalized state
Assurance-reference SD	0.08 tolerance units
Agent target z_target	[0.80, -0.20, -0.15, -0.10, 0.35, 0.30]
Agent objective weights	[0.9, 1.5, 0.9, 0.8, 1.1, 1.3]
Control regularization λ	0.04

Appendix B. Authority and Ablation Algorithm

At each decision epoch, the agent proposes an action from the digital-twin estimate and the assurance layer computes A, C, CDI, and C-SAI. CG then assigns L3-L0 using the stated C-SAI thresholds and CDI overrides; A-only and C-only ablations use the analytically derived component thresholds. Downgrades are immediate, whereas upgrades require two consecutive qualifying checks. The selected level blends the agent action with the common safeguarded reference action, with L1/L0 recorded as escalation requests. The bounded action is then applied to the simulated process. Latent truth is never used online and is accessed only afterward for oracle-regret calculation.

Acknowledgement
This research was not funded by any grant.

Conflicts of Interest
The authors declare no conflicts of interest.

References

[1] Breque, M., De Nul, L., & Petridis, A. (2021). Industry 5.0: Towards a sustainable, human-centric and resilient European industry. European Commission, Directorate-General for Research and Innovation. https://doi.org/10.2777/308407
[2] Qu, J., Kizil, M. S., Yahyaei, M., & Knights, P. F. (2023). Digital twins in the minerals industry - a comprehensive review. Mining Technology, 132(4), 267-289. https://doi.org/10.1080/25726668.2023.2257479
[3] Nobahar, P., Xu, C., Dowd, P., & Shirani Faradonbeh, R. (2024). Exploring digital twin systems in mining operations: A review. Green and Smart Mining Engineering, 1(4), 474-492. https://doi.org/10.1016/j.gsme.2024.09.003
[4] Ebad, S. A., Abueid, A. I., Amara, M., & Ahmed, R. (2026). AI-Driven Digital Twins in Mining Operations: A Comprehensive Review. Technologies, 14(5), 269. https://doi.org/10.3390/technologies14050269
[5] Liang, R., Zhang, C., Li, B., Saydam, S., *et al.* (2026). Evolution of visualisation and digital twin technologies in mining. International Journal of Coal Science & Technology, 13(1), 53. https://doi.org/10.1007/s40789-026-00866-w
[6] Sajadieh, S. M. M., & Noh, S. D. (2025). From Simulation to Autonomy: Reviews of the Integration of Artificial Intelligence and Digital Twins. International Journal of Precision Engineering and Manufacturing-Green Technology, 12(5), 1597-1628. https://doi.org/10.1007/s40684-025-00750-z
[7] Ivanov, D. (2026). Agentic digital twins: bridging model-based and AI-driven decision-making support for a new era of supply chain and operations management. International Journal of Production Research, 64(15), 6574-6590. https://doi.org/10.1080/00207543.2026.2630277
[8] Hasan, A., & Nguyen, D. T. (2026). Integrating agentic AI and digital twins for intelligent decision-making systems. Array, 29, 100721. https://doi.org/10.1016/j.array.2026.100721
[9] Collao, C., Akhtar, H., Toro, C., Ocampo-Martinez, C., Schor, R., & Pinto Prieto, R. (2025). An Agentic AI-based Architecture for Digital Twins Specialized in Predictive Maintenance: Application to Ball Mills. IFAC-PapersOnLine, 59(32), 96-101. https://doi.org/10.1016/j.ifacol.2025.12.403
[10] Mehmood, U., Sheikhi, S., Bak, S., Smolka, S. A., & Stoller, S. D. (2022). The Black-Box Simplex Architecture for Runtime Assurance of Autonomous CPS. In NASA Formal Methods, LNCS 13260, 231-250. https://doi.org/10.1007/978-3-031-06773-0_12
[11] Sheikhi, S., Mehmood, U., Bak, S., Smolka, S. A., & Stoller, S. D. (2025). The black-box simplex architecture for runtime assurance of multi-agent CPS. Innovations in Systems and Software Engineering, 21(2), 619-634. https://doi.org/10.1007/s11334-024-00553-6
[12] Bellman, K. L., Landauer, C., Dutt, N. D., Esterle, L., Herkersdorf, A., Jantsch, A., *et al.* (2020). Self-aware cyber-physical systems. ACM Transactions on Cyber-Physical Systems, 4(4), Article 1-26. https://doi.org/10.1145/3375716
[13] Dutt, N., Regazzoni, C. S., Rinner, B., & Yao, X. (2020). Self-awareness for autonomous systems. Proceedings of the IEEE, 108(7), 971-975. https://doi.org/10.1109/JPROC.2020.2990784
[14] Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. Proceedings of the 35th International Conference on Machine Learning, PMLR 80, 2796-2804.

[15] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321-1330.
[16] 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
[17] Liu, X., Zhang, Y. E., Kasprova, V., Rabbani, P., Zahraei, P. S., Zhang, T., Ebrahimpour-Boroojeny, A., & Chandrasekaran, V. (2026). AgentAbstain: Do LLM Agents Know When Not to Act? arXiv:2607.10059. https://doi.org/10.48550/arXiv.2607.10059
[18] AL-Ali, M., Marks, A., Mohamed, A. A., Balaha, H. M., Badawy, M., Elhosseini, M. A., *et al.* (2026). Calibrated adaptive framework for trustworthy human and artificial intelligence decision systems. Scientific Reports. https://doi.org/10.1038/s41598-026-65730-y
[19] Ali, M. A. (2025). Eco-Cognitive Mining Systems (ECMS): A foundational framework for self-aware, adaptive, and sustainable mining operations. Advances in Research, 26(5), 640-650. https://doi.org/10.9734/air/2025/v26i51519
[20] Ali, M. A. (2026a). Eco-Cognitive Mining Systems (ECMS): Human-in-the-loop digital twin-surrogate-based sustainability-aware control for mining systems. Advances in Research, 27(1), 204-216. https://doi.org/10.9734/air/2026/v27i11581
[21] Ali, M. A. (2026b). Calibrated operational self-awareness for mining digital twins: Detecting confident misperception in Eco-Cognitive Mining Systems (ECMS). Applied Expert Systems and Knowledge Management, in press.
[22] Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36-52. https://doi.org/10.1016/j.cirpj.2020.02.002
[23] Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital Twin: Enabling technologies, challenges and open research. IEEE Access, 8, 108952-108971. https://doi.org/10.1109/ACCESS.2020.2998358
[24] Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital Twin: Values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980-22012. https://doi.org/10.1109/ACCESS.2020.2970143
[25] Quintanilla, P., Fernandez, F., Mancilla, C., Rojas, M., & Navia, D. (2025). Digital twin with automatic disturbance detection for an expert-controlled SAG mill. Minerals Engineering, 220, 109076. https://doi.org/10.1016/j.mineng.2024.109076
[26] van Eyk, L., & Heyns, P. S. (2025). A framework to define, design and construct digital twins in the mining industry. Computers & Industrial Engineering, 200, 110805. https://doi.org/10.1016/j.cie.2024.110805
[27] Contreras, F. A. (2025). A novel approach to grinding modeling and scale-up: Toward a digital twin. Minerals Engineering, 232, 109584. https://doi.org/10.1016/j.mineng.2025.109584
[28] Seppi, M., Linnosmaa, J., & Zeb, A. (2025). Physics-informed machine learning surrogate models: Enhancing data-driven forecasting for digital twins in mineral processing. Minerals Engineering, 230, 109424. https://doi.org/10.1016/j.mineng.2025.109424
[29] Chen, Z., Heckman, C., Julier, S., & Ahmed, N. (2018). Weak in the NEES?: Auto-tuning Kalman filters with Bayesian optimization. 21st International Conference on Information Fusion, 1072-1079. https://doi.org/10.23919/ICIF.2018.8454982