A FALSIFICATION STUDY OF COUPLED-CONSTRAINT TELEMETRY DIAGNOSTICS IN MULTI-AGENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We evaluate three author-specified telemetry signatures before their labels are mistaken for causes or safety certificates. A deliberately separated 200-run construct/unit test is followed by independently scheduled controller faults and then natural future outcomes. On the 480 independent-fault runs, the frozen T1–T3 instrument resolves 322 (67.1%); 280/322 resolved verdicts are correct (.870), 42/480 runs are wrongly resolved (.0875), and 32/120 healthy runs raise a false alarm (.267). It does not significantly outperform learned baselines under forced scoring, and its .2021 advantage over a no-fit severity rule comes entirely from obstruction recall. In an out-of-domain predictive-utility test of T1/T3 proxies, all 1,265 original windows are unresolved and 95–100% of budget-sweep windows are unresolved, although a no-fit current-cost persistence baseline is informative on the same target. Calibrated conjunctions equal a calibrated violation-only rule on both 2,876 held-out windows and a 17,040-window extension, providing no demonstrated incremental value from residual or dual gates; T2 and natural mechanism labels are unavailable there, so natural mechanism accuracy remains unidentified. Exact identifiability, split-conformal, detection-delay, and unlabeled-target results formalize the required assumptions. This is a case-specific falsification study of this instrument, not validation of a common diagnostic, natural causal diagnosis, or a safety certificate.

1 INTRODUCTION

Constrained MARL includes centralized-critic and multiplier-based methods (Yu et al., 2022; Gu et al., 2021; 2023). This paper studies one author-specified monitor built from the traces available in our implementations: a primal residual, realized budget violation, candidate-cost minima, and a clipped multiplier. We do not claim that the T1–T3 conjunctions are established or common diagnostics. ADMM residuals concern agreement in an optimization interface, not closed-loop safety (Boyd et al., 2011; Bastianello et al., 2021); a clipped multiplier reveals pressure at a configured cap, not the source of infeasibility; and a local action screen can be empty even under exact consensus.

The scientific question is whether this three-signature instrument retains meaning as label independence and target shift increase, before construct recognition is promoted to natural diagnosis or safety. A signature label is not a unique causal mechanism. Runtime verification motivates a rigorous monitoring vocabulary (Bartocci et al., 2018); safe MARL supplies the case-study setting. Our study preserves frozen decisions, matched information, abstention, and adverse controls at each escalation.

We make three contributions. First, we specify the author-defined T1–T3 instrument and label its perfectly separated 200-run result as a construct/unit test, not a natural-performance estimate. Second, we test the frozen instrument on independently scheduled faults and natural future outcomes: selective performance degrades on the former and coverage collapses on the latter despite an informative matched persistence rule. Held-out generator and separate- simulator stresses provide secondary sensitivity checks. Third, an exact identifiability characterization for finite telemetry conjunctions, together with misspecification, split-conformal, detection-delay, and unlabeled-target results, formalizes when these available signals can and cannot support mechanism labels. Table 5

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

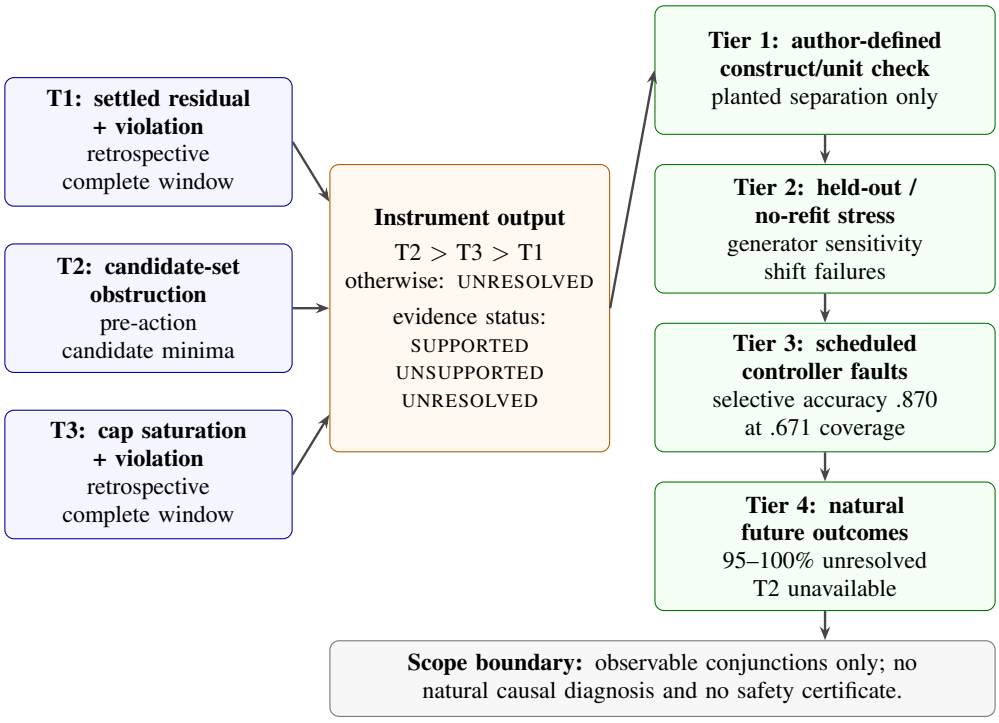

Figure 1: Overview of the instrument and evidence ladder. T1–T3 telemetry enters the frozen precedence map before progressively less favorable validation tiers. T2 requires candidate-cost telemetry that is unavailable in the natural datasets; T1 and T3 are retrospective. The outputs remain selective observational verdicts rather than mechanism identification or safety certification.

Table 1: Frozen operational specification. Q_k are independently evaluated base flags; P_k are exclusion-masked labels. Means use the last 12 steps. The display hierarchy prevents double counting but never deletes Q; no match is UNRESOLVED.

Test	Base flag Q_k / reported label P_k	Adverse control	Verdict / excluded inference
T1	$Q_1 = [\bar{r}_t \le .05 \wedge \bar{v}_t \ge .50]$; $P_1 = Q_1 \wedge \neg Q_2 \wedge \neg Q_3$	healthy learning, same budget/disturbance	stall signature; not its cause; retrospective
T2	$Q_2 = [o_t = 1]$; $P_2 = Q_2$	$\sum_i \ell_{i,t} \le d$ (nonnegative minima)	candidate obstruction; not global infeasibility
T3	$Q_3 = [\overline{\lambda_t/\lambda_{\max}} \ge .98 \wedge \bar{v}_t \ge .50]$; $P_3 = Q_3 \wedge \neg Q_2$	uncapped healthy dual response	cap/violation signature; not optimizer failure; retrospective

binds every claim to its evidence; the construct constants and thresholds were not preregistered, and the 20-seed extension follows only an internal paper-local plan. The contribution is a positive-to-negative case record, not a generally sufficient audit protocol, new signal, natural causal identifier, or optimizer.

2 Diagnostic Instrument

2.1 Constrained learning trace

For agents $i \in \{1, \dots, N\}$ with policies π_i, nonnegative costs C_i, and team budget d, consider

$$\max_{\pi_1,\dots,\pi_N} \sum_i J_i(\pi_i) \quad \text{s.t.} \quad \sum_i C_i(\pi_i) \le d. \tag{1}$$

A scalar budget interface introduces $b_i \geq 0$ with $C_i \leq b_i$ and $\sum_i b_i \leq d$. For $d \geq 0$, the planted simulator uses the radial budget normalization $R_d(q) = q_+$ when $\mathbf{1}^\top q_+ \leq d$, and $R_d(q) = dq_+/(\mathbf{1}^\top q_+)$ otherwise, where $q_+ = \max(q, 0)$ elementwise. This is not Euclidean simplex projection. With raw proposal q_t and allocation $b_t = R_d(q_t) \in \Delta(d) = \{b \geq 0 : \mathbf{1}^\top b \leq d\}$, we normalize their disagreement by the physical budget scale,

$$r_t = \frac{\|q_t - b_t\|_2}{\max(d, 10^{-12})}, \qquad v_t = \mathbf{1}\left[\sum_i c_{i,t} > d\right]. \tag{2}$$

Normalization makes the T1 threshold dimensionless; it neither turns residual convergence into feasibility nor makes the threshold comparable across implementations, as Section 4 shows. Section 4 reports both raw and normalized trajectories rather than only a terminal average.

For pre-action candidate costs $c^{\mathrm{cal}}_{i,t}(a)$, define

$$\ell_{i,t} = \max\left\{0, \min_{a \in \mathcal{A}_i} c^{\mathrm{cal}}_{i,t}(a)\right\}, \qquad o_t = \mathbf{1}\left[\sum_i \ell_{i,t} > d\right]. \tag{3}$$

If $o_t = 1$, every allocation in $\Delta(d)$ leaves at least one unfloored local candidate set empty: otherwise summing $\ell_{i,t} \leq b_i$ contradicts $\sum_i b_i \leq d$. This is a static candidate-set result, not proof that the underlying control problem is globally infeasible.

2.2 Predicates and verdict

T1 and T3 require a complete window; T2 is available pre-action immediately. A test is SUPPORTED only when its artifact verifies, its positive predicate holds, its adverse control does not, and the run is in the declared domain. A failed predicate is UNSUPPORTED; missing evidence or a domain mismatch is UNRESOLVED. Let $Q_t = (Q_{1,t}, Q_{2,t}, Q_{3,t})$ contain the three independently evaluated base flags in Table 1. Exclusion masking produces $P_{2,t} = Q_{2,t}$, $P_{3,t} = Q_{3,t} \wedge \neg Q_{2,t}$, and $P_{1,t} = Q_{1,t} \wedge \neg Q_{2,t} \wedge \neg Q_{3,t}$. The display map returns T2, T3, or T1 when the corresponding P_k is true, and UNRESOLVED otherwise. Hence raw overlap such as $Q_1 = Q_3 = 1$ remains reportable even though the displayed labels P_k are mutually exclusive; an all-zero Q is never called "healthy." The original .05/.50/.98/12 configuration is retrospective. A second synthetic study selects among 81 configurations using seeds 0–49 only and evaluates disjoint seeds 50–99; this calibrates the local generator, not cross-algorithm transport. Maximization uses a deterministic lexicographic key (training accuracy, then proximity to the original residual, violation, dual, and window values). The rendered submission specifies that rule but does not independently authenticate when the full candidate grid was fixed.

The reported T1 label denotes a settled-residual/violation signature after excluding obstruction and cap saturation. T2 denotes candidate-set obstruction. The reported T3 label denotes cap saturation with violation after excluding T2. These are observable conjunctions, not unique causal mechanisms: an executable counterexample assigns identical T1/T3 telemetry to distinct stall and saturation causes. The hierarchy prevents double counting but cannot restore causal identifiability. Proposition 4 gives the exact analytic boundary: a finite flag vector identifies a mechanism perfectly if and only if no output cell has positive probability under two mechanisms, and coarsening to the display verdict preserves Bayes accuracy if and only if each merged cell retains one common posterior-maximizing mechanism.

3 Controlled Planted-Failure Validation

3.1 Design

We use a local particle-style constrained-learning simulator with eight agents, 240 steps, binary safe/risky actions, additive realized cost, and budget $d = 1$. Each actor has a stochastic logistic policy and receives a local state perturbation. A scalar primal–dual update adapts the actors and team multiplier. At step 80, one of four interventions is applied: (i) *healthy*: no intervention and an uncapped effective dual response; (ii) *optimizer stall*: risky logits are shocked while actor and dual updates are frozen; (iii) *infeasible budget*: every safe candidate's base cost is raised so $\sum_i \ell_i > d$;

Table 2: Construct/unit test with deliberately separated planted conditions (50 seeds/class; 200 runs). Intervals are Wilson 95% intervals. Lead is the online sustained-failure declaration step minus first verdict; positive is early.

Test	Planted target	Sens.	Spec.	FPR	FNR	Median lead
T1	optimizer stall	1.00 $[.929, 1]$	1.00 $[.975, 1]$	0	0	-28
T2	infeasible budget	1.00 $[.929, 1]$	1.00 $[.975, 1]$	0	0	$+11$
T3	dual saturation	1.00 $[.929, 1]$	1.00 $[.975, 1]$	0	0	-3

Perfect observed separation is a construct check under deliberately separated interventions, not an estimate for natural failures. T1 and T3 are retrospective (negative lead); only pre-action T2 warns early.

Table 3: Same-budget, same-distribution intervention on the planted optimizer-stall class (50 paired seeds). Deltas are screened minus unscreened; 95% paired-t intervals use seeds as units.

Outcome	Unscreened	Screened	Paired delta [95% CI]
Violation rate	0.905	0.000	-0.905 $[-0.911, -0.899]$
Reward/step	7.309	4.640	-2.669 $[-2.680, -2.657]$

State, action-uniform, and realized-cost-noise streams are byte-for-byte shared within seed. The screen changes only action acceptance. Safety improves here at a measured reward cost; this is synthetic intervention evidence, not benchmark dominance.

or (iv) *dual saturation*: risky logits are shocked and the multiplier cap is reduced to 0.20. The detector receives only current and prior telemetry, never the intervention label, future state, or oracle outcome.

We execute 50 independent seeds per class. The statistical unit is a complete seed run. Sustained failure is declared online only after a complete 12-step window with violation rate at least .50. Lead time equals that declaration step minus the first diagnostic verdict; positive values are early. Wilson intervals quantify binomial rates. This is a 200-run construct/unit test of deliberately separated conditions; it does not estimate prevalence, difficulty, or accuracy for natural failures.

All 150 planted failures receive their assigned signature, while all 50 no-intervention controls receive UNRESOLVED; the construct-assignment matrix is diagonal under that explicit mapping (200/200, Wilson 95% CI $[.981, 1]$). This is unit-test-style separation by construction, not a natural diagnostic-accuracy estimate. Under selective semantics, healthy coverage is 0/50 because UNRESOLVED is an abstention. For each one-vs-rest T1–T3 predicate the denominators are 50 planted positives and 150 other runs: TP/FN/TN/FP are 50/0/150/0. The adverse controls are substantive: healthy runs share the same budget, stochastic state law, action sampler, and cost noise, but average post-intervention violation is .0103 rather than .905–1.000 for the planted failures.

Perfect separation here should be read together with timing. T2 fires at the intervention because it uses pre-action candidate minima and precedes the 12-step sustained-failure declaration by 11 steps in every infeasible run. T3 is typically three steps late and T1 28 steps late because both require realized violation and, for T1, residual settling. Thus T1/T3 are reliable retrospective labels in this construction, not warning devices. This negative timing result changes the operational decision: only T2 may trigger preventive recovery; T1/T3 should trigger investigation or shutdown after evidence accumulates. Proposition 5 gives an exact boundary only for its iid Bernoulli model. Plugging the simulator's measured average rates into that fitted surrogate puts the 12-step gate within 1.3 steps of its converse floor and gives a post-hoc floor of at least 62 steps at the pre-action channel's fitted budget. These are model-based comparisons, not guarantees for the adaptive, temporally dependent simulator streams.

3.2 SAME-BUDGET SCREENED INTERVENTION

For the optimizer-stall class we rerun every seed with a pre-action local screen. Within each pair, initial parameters, state sequence, Bernoulli action uniforms, realized-cost noise, budget, and update schedule are identical; only rejection of locally over-budget risky proposals differs. This is a same-distribution intervention rather than a comparison of separately visited trajectories.

The screen eliminates observed violations but also removes risky progress, producing a precise reward loss. A distribution-free paired sign test agrees with the mean-based intervals: all 50 violation and reward differences are negative (two-sided $p = 2^{-49} = 1.78 \times 10^{-15}$); median differences are $-.90625$ and -2.66925, respectively. This establishes causal efficacy of this screen *inside the constructed feasible candidate regime*. It does not rescue the infeasible-budget class: when $\sum_i \ell_i > d$, no local acceptance rule under the same candidate set can make every unfloored set nonempty. The intervention is assigned by the planted regime rather than triggered prospectively by T1/T2, so it establishes no diagnostic action utility.

3.3 Implementation, controls, and inferential unit

The simulator generates all exogenous arrays before learning. Safe candidate cost is $.058 + .008|s|$ before intervention, while risky cost adds $.105 + .012 \max(s, 0)$. Risky reward exceeds safe reward by $.42 + .02s$. In responsive classes, actor and dual learning rates are .12 and .35. The optimizer-stall intervention adds 3.4 to risky logits and freezes both updates; dual saturation applies the same shock and lowers the cap to .20; infeasibility raises every safe base cost to $.142 + .010|s|$, making the eight-agent safe minimum exceed $d = 1$. These constants specify the implemented planted constructs; their selection was not preregistered, and we make no claim that they were chosen independently of prior classification checks.

For each integer seed, NumPy initializes one state array, one action-uniform array, and one realized-cost-noise array. The detector reads only current or prior telemetry. The screen pair reuses those arrays exactly and never reads the candidate action's realized noise. We average all timesteps within a run before cross-seed inference; neither agents nor timesteps are treated as independent replicates. The released summary therefore contains 200 diagnostic units and 50 matched intervention units, not inflated per-step sample sizes.

The initial validation answers a deliberately narrow construct-consistency question. Predicate cross-tabs are separate from the hierarchy: each T1–T3 row has TP/FN/TN/FP $= 50/0/150/0$ because the interventions deliberately instantiate separated conditions. This is an implementation check, not evidence of natural-failure accuracy. The next study deliberately removes perfect separation.

3.4 Held-out ambiguity and misspecification

A second local generator assigns labels from latent causes before sampling noisy, overlapping telemetry; detector outputs never define labels. Each 160-step trace changes at step 80. Before the change, violation probability is .08, the residual is $\max(.005, .18e^{-t/32} + \mathcal{N}(0, .012))$, and dual fraction is clipped noisy telemetry. Post-change violation probabilities for healthy/stall/saturation/infeasible are .08/.58/.58/.82; their dual means are .18/.42/.965/.93, and infeasibility independently activates obstruction with probability .88 per step. Mixed and unknown stress classes use probabilities .68/.62; unknown raises residual above .07 without obstruction or cap saturation. Telemetry noise streams are seeded as seed$+10{,}000\times$(mechanism index), so each (seed, mechanism) pair draws its own stream and detector output never enters label assignment; no further formal leakage audit is reported.

We calibrate residual, violation, dual, and window choices over 81 configurations on 50 seeds per single cause, then freeze the choice and evaluate disjoint seeds 50–99 (200 runs). We repeat this 50/50 procedure on five nonoverlapping 100-seed blocks. Separately, we fit every method once on nominal seeds 0–49 and, without refitting, cross signal-contrast factors $1, .75, .5$ with telemetry-noise factors $1, 1.5$ on seeds 50–99. Contrast contracts each post-event probability and dual mean toward the healthy value while preserving latent labels. We additionally freeze all methods trained on seeds 0–49 and evaluate 200 summaries from the separately implemented planted simulator (seeds 50–99), using final residual/dual and post-event violation/obstruction. This is out-of-generator locally, but both generators remain constructed and synthetic.

The frozen .05/.50/.98/12 rule reaches .715 held-out accuracy; calibration selects .05/.40/.94/24 and reaches .980, with per-class recall healthy/stall/saturation/infeasible $= 1/.98/.94/1$. Across five disjoint seed blocks, calibrated held-out accuracy is .980/.980/.975/.955/.955 and the selected window changes from 24 to 12 in two blocks. Under the frozen parameter-shift grid, rule accuracy falls from .980 to .690 at .75 contrast and .490 at .50 contrast; worst-case centroid/tree accuracy is

Table 4: Held-out ambiguous-telemetry stress test. Calibration uses 50 seeds/cause and evaluation uses disjoint seeds 50–99 (200 runs). All classifiers use the same four window means; selective NC calibrates its rejection radius on training only.

Evaluation	Result	Boundary
Frozen / calibrated rule	.715 / .980	143 / 196 correct
Rule recall H/S/D/I	1/.98/.94/1	class-level
Majority / NC / depth-3 tree	.250 / 1.000 / 1.000	rule is not best
Selective NC	1.000 at .965 coverage	unknown abstention 50/50
Five repeated splits	.955–.980	median .975
Parameter shift, frozen train	.490–.980 rule	NC/tree worst .415/.450
Separate simulator R/NC/tree	.750/.500/.500	stall recall 0
Selective NC, separate simulator	.000 coverage	rejects all 200
Replay direct–rule $\Delta V/\Delta R$	$-.226/-.564$	paired 95% exclude 0
Window 6 / 12 / 24	.855/.970/.980	sensitivity
Mixed raw $Q_1 + Q_3$	47/50 both flags	display hides overlap
Unknown mechanism	50/50 unresolved	0 known labels
T2 iid error $\sigma = .10$	P/R/FPR=.657/.623/.331	degraded
Conformal T2, exchangeable	P/R/FPR=.995/.421/.002	.276 resolved
Conformal T2, bias shift	P/R/FPR=.582/.998/.724	.360 wrong

.415/.450. At .75 contrast, saturation recall is zero. Window-only sensitivity ranges from .855 at 6 to .970 at 12 and .980 at 24. Thus all three classifiers are generator-sensitive.

For incremental value, nearest centroid and depth-3 CART score 1.000 versus .980 for the rule; selective centroid scores 1.000 at .965 coverage and rejects all 50 constructed unknowns. On the separate simulator, rule/centroid/tree score .750/.500/.500, selective coverage is zero, and rule stall recall is zero. Frozen rule/centroid/tree next select paired replay branches: they differ in one of 200 actions (Wilson 95% $[.0009, .0278]$) and yield (V, R) between $(.474, 5.446)$ and $(.478, 5.459)$. Direct $\bar{v} \geq .4 \wedge \bar{o} < .5$ yields $(.253, 4.896)$; direct-minus-rule seed-block differences are $-.226$ in violation (95% $[-.227, -.224]$) and $-.564$ in reward ($[-.567, -.561]$). Thresholds .2–.8 give the same point; without the obstruction guard, reward falls to 4.795 at unchanged violation. Thus no taxonomy action advantage or dominance is demonstrated, and this retrospective synthetic audit is not online recovery. Mixed raw $Q_1 + Q_3$ remains 47/50; all 50 constructed unknowns remain UNRESOLVED.

The implemented summaries contain the same tuple $(\bar{r}, \bar{v}, \bar{\lambda}, o) = (.02, .75, .99, 0)$ under runs carrying distinct stall and saturation labels, so the displayed T1–T3 tuple is not one-to-one with those sampled labels. The text does not fully specify the two class-conditional telemetry laws, however, and therefore makes no positive-density claim from that point witness. Proposition 4 now supplies the exact law-level error and the necessary-and-sufficient condition for generator-wide identification; Proposition 3 supplies a separate lower bound when target labels are absent. For T2, iid candidate-cost noise degrades exact-cost precision/recall 1.00/1.00 to .657/.623 at $\sigma = .10$, with FPR .331. We next calibrate the corrected split-conformal rank $\lceil (n+1)(1-\alpha) \rceil = 476$ on $n = 500$ absolute total-cost errors. On 2,000 exchangeable draws, the radius .093 yields precision/recall/FPR .995/.421/.002, resolves .276, and makes wrong resolved verdicts in .001 of all draws. Proposition 2 gives marginal error at most $\alpha = .05$ under exchangeability; it does not guarantee conditional precision. Raising error bias from .025 to .180 violates the assumption: interval coverage falls to .019, precision to .582, FPR rises to .724, and wrong resolved verdicts to .360. Thus the guarantee is decision-relevant but non-transportable, and T2 still inherits candidate-model shift.

Proposition 3 strengthens the natural-label boundary. For a frozen selective binary diagnostic with balanced resolved predictions, two target laws can have exactly the same telemetry distribution, coverage, and class prevalence while resolved accuracy is respectively one and zero. Consequently, every estimator based only on unlabeled target telemetry has worst-case expected absolute accuracy error at least $1/2$, regardless of sample size. Independent target labels or an identifying assumption are therefore information-theoretically necessary for natural-failure accuracy claims.

Table 5: Claim-to-evidence map. Every evidentiary row is source/result/test bound in the anonymous content-addressed manifest; archival rows provide context only. "Fit" denotes training-only calibration.

Result	Evidence class	Freeze / fit boundary	Size	Comparator	Manifest claim
Planted T1–T3	Author-defined construct/unit check	Predicates frozen; no test fit	200 runs	Class confusion	C1
Ambiguous telemetry	Calibrated in-generator test	50 train / 50 held-out seeds	400 tests	Rule, centroid, tree	C2
Separate simulator	Out-of-generator stress	No target refit	200 tests	Same three models	C3
Planted screen	Action intervention	Paired streams; fixed gate	50 pairs	No screen	C1
MPE2 action overwrite	Intervention-labeled construct check	T1–T3 frozen across envs	480 + 360 pairs	Copied labels, tree, centroid	C4
Budget outcomes	Natural-outcome test	Frozen windows and labels	8 arms	Within-method strata	C5
Internal faults	Independent-label stress	T1–T3 frozen; untouched seeds	480 + 360 pairs	Direct, tree, centroid	C6
Grid / RCBS / MAPPO	Archival context only	Heterogeneous protocols	Listed locally	None valid cross-row	C8

C1–C8 follow manifest order; the manifest is identified in Appendix A. Legacy external rows are not used to support diagnostic accuracy, transfer, superiority, or recovery claims.

4 Residual Scale and Trajectory Evidence

In the retained $N = 20$, $d = 100$, ten-seed grid trace, primal residual falls by over four orders of magnitude through episodes 21–40 while violation stays .7425 and cost/budget stays about 1.282. The grid ADMM residual differs from the simulator's radial-normalization residual, so its scale cannot justify T1's .05 threshold. Appendix N reports both plots, all trajectory aggregates and the averaging protocol. The trace supports only the warning that an agreement metric can decrease while violation persists, not a transported diagnostic threshold.

5 Escalating Validation Tiers

The case-study ladder deliberately escalates from a definitional construct check, to held-out and no-refit generator stress, to independently scheduled controller faults, and finally to natural future outcomes. Within this instrument, promotion requires that labels be defined outside extraction, thresholds be frozen before the target split, direct and learned feature-matched baselines be retained, and abstention and adverse results remain visible. These requirements organize the falsification record; they are not claimed necessary or sufficient for auditing arbitrary diagnostics. T1–T3 are the instrument under test, not a proposed optimizer.

Cross-system boundary probes. Heterogeneous grid, RCBS, MAPPO, and adapter-defined MPE2 rows only establish that residual convergence or a saturated dual can coexist with violation. They are moved to Appendix E; incompatible budgets, training exposure, endpoints, and missing T2 channels prohibit accuracy, scaling, or optimizer comparisons.

Independently scheduled controller faults. The controlled fault study assigns labels only through a hidden scheduler: after a common cost shift it freezes both safety-bias and dual learning, lowers the dual cap, or lowers budget below latent candidate minima. Actions are not overwritten, and extraction reads only parameter updates, noisy candidate/selected costs, budget, and dual state. T1–T3 stays fixed on 20 untouched seeds (480 runs). Under the selective scoring reported next, the instrument resolves 322/480 runs and is wrong on 42/480. Under the older forced mapping of abstention to healthy, its .7667 accuracy is not significantly different from the leave-one-environment-out tree or centroid. The .2021 gap over the fixed no-fit severity argmax comes entirely from obstruction recall (.1917 versus 1.00); healthy, stall, and saturation recalls are identical. Correct-alarm fault removal triggers in 304/360 fault pairs and improves both descriptive outcomes. The scheduler faults remain engineered rather than natural causes; a preliminary smoke check produced no claim or artifact.

Table 6: Selective-verdict re-aggregation of both MPE2 studies (480 runs each: 120 healthy, 360 fault). Abstention (UNRESOLVED) is neither correct nor wrong. Cov. is $n_{\text{resolved}}/480$; Acc|res. is $n_{\text{correct}}/n_{\text{resolved}}$ with model-seed bootstrap 95% interval; Wrong is $n_{\text{wrong}}/480$; H-FA is $n_{\text{healthy alarm}}/120$; Fault acc. is $n_{\text{correct fault}}/360$. Naive maps abstention to HEALTHY. The stored direct-comparator rows and tree always resolve; the labeled-transfer comparator copies instrument labels rather than independently classifying.

Study	Method	Cov.	Acc|res. [95%]	Wrong	H-FA	Fault acc.	Naive
Independent faults	Frozen T1–T3	0.671	0.870 [0.852,0.886]	0.0875	0.267	0.778	0.767
Independent faults	Direct telemetry	1.000	0.565 [0.533,0.598]	0.435	0.267	0.508	0.565
Independent faults	Depth-3 tree	1.000	0.742 [0.710,0.771]	0.258	0.692	0.886	0.742
Independent faults	Selective centroid	0.948	0.767 [0.733,0.801]	0.221	0.292	0.750	0.740
Labeled transfer	Frozen T1–T3	0.762	0.984 [0.970,0.994]	0.013	0.050	1.000	0.988
Labeled transfer	Copied instrument labels	1.000	0.988 [0.977,0.996]	0.013	0.050	1.000	0.988
Labeled transfer	Depth-3 tree	1.000	0.990 [0.981,0.996]	0.010	0.042	1.000	0.990
Labeled transfer	Selective centroid	0.971	0.983 [0.972,0.993]	0.017	0.067	1.000	0.983

Selective-verdict re-aggregation. The stored analyses, and hence the accuracies above and in Tables 9–10, map UNRESOLVED to the HEALTHY label, so an abstention on a healthy run counts as a correct healthy prediction and the instrument's "healthy recall" is its abstention rate on healthy runs. That contradicts the selective semantics the instrument is built on. Table 6 recomputes every method from the stored runs with abstention scored as neither correct nor wrong, refitting the baselines leave-one-environment-out with the original functions; the naive column reproduces the stored numbers exactly. On independently scheduled faults, T1–T3 resolves 322/480 runs (67.1%), is correct on 280/322 resolved runs (.870; model-seed 95% interval in the table), is wrong on 42/480 total runs (.0875), and raises a false alarm on 32/120 healthy runs (.267). The depth-3 tree resolves all 480 runs, is correct on 356/480 (.742), is wrong on 124/480 (.258), and raises false alarms on 83/120 healthy runs (.692). On labeled transfer the instrument resolves 76.2% with .984 accuracy given resolution. The selective centroid had the same conflation (abstention mapped to healthy); recovered, it resolves 94.8% at .767. The .7667 value counts abstention as a correct HEALTHY prediction and is therefore not our headline estimand; we report the selective quantities instead. The nonsignificance test concerns forced-scoring accuracy, not these coverage-dependent estimands. A post-hoc calibration-targeted reanalysis of the same previously inspected records now supplies one foldwise coverage-matched comparison: frozen T1–T3 has risk $42/322 = .1304$ at $322/480$ coverage, while the depth-3 tree has $43/315 = .1365$ at $315/480$; tree-minus-frozen risk is $+.0061$ with paired model-seed-cluster 95% interval $[-.0199, .0313]$. Only the tree passes the $\pm.05$ achieved-coverage gate in every environment fold; direct telemetry and centroid fail despite pooled coverage gaps within .05. This does not establish superiority or equivalence, and it does not imply that all baselines were matched. Abstention remains the instrument's dominant behavior on healthy runs.

Natural future-outcome corpus and recorded budget sweep. Failure is independently defined by the next disjoint window's mean cost-to-budget ratio. The original RCBS/MAPPO corpus yields 400 future-failure and 865 future-safe windows, but T1–T3 abstains on 1,265/1,265. A recorded eight-arm budget sweep places both outcomes within methods (Table 13) yet still abstains on 95–100% of windows: MAPPO sensitivity is .00–.03 at limits 3–10 and .56 (10/18) at 20 with precision .45; RCBS alerts on none. Alerts occur after the first future-failure label where present; this label is not a physical crossing timestamp. T2 telemetry is absent, and overlapping windows remain descriptive. Thus outcome transfer fails even though the intervention-labeled construct test succeeds.

This is not merely a change to an impossible target. A matched no-fit baseline alerts when the *current* 12-unit mean cost/budget exceeds its physical boundary one, and predicts the same next-window label. At MAPPO limits 5 and 10 its sensitivity/specificity is $1.00/.84$ and $1.00/.939$, versus $0/.939$ and $.030/.947$ for T1–T3. At RCBS limit 10 it attains $1.00/.951$, versus $0/1.00$. The limit-3 arm is degenerate ($858/865$ future failures), and performance is not uniform: at MAPPO limit 20 persistence sensitivity is .333 versus .556 for T1–T3. Thus the matched comparison isolates a coverage failure of these frozen proxy conjunctions, not a universal failure of calibrated rules or a claim that mechanism labels should equal future outcomes. A separate retrospective seed split trains on MAPPO seeds 100–101 and RCBS seeds 100–104, holding every other seed out across all four budgets. Calibrated conjunctions improve the binary forecast, but equal a calibrated violation-

only rule on all 2,876 test windows. Their mean error differences from cost persistence are $-.0183$ for MAPPO and approximately zero for RCBS; both paired seed t intervals include zero. An extension to 20 new seeds per method keeps the fitted rules unchanged: pooled error differences are $-.0155$ $[-.0196, -.0115]$ for MAPPO and $-.0009$ $[-.0102, .0084]$ for RCBS (pointwise seed-t 95% intervals). The union and calibrated violation-only predictor agree on all 17,040 new windows. Appendix M.1 separates budget-specific evidence from inconclusive comparisons; no multiple-test control, natural mechanism labels or T2 observations are introduced.

Finite-horizon T2 monitoring. Trajectory-max calibration conditionally controls any wrong resolved verdict over a fixed horizon. Matched-law simulations reduce run-level errors to 0.05–0.8% but resolve only 1.47–8.56% of time points; under shift, every run has a wrong resolution. Appendix L derives the guarantee and reports these limitations.

6 Limitations and Conclusion

The 200-run synthetic suite and the action-overwrite MPE2 suite are only construct checks. The independent fault scheduler removes label leakage and direct action rewriting, but its faults and cost shift remain engineered; neither forced-scoring .7667 accuracy nor 280/322 selective accuracy among resolved runs is natural causal identification. The gain is against the explicitly weak no-fit severity argmax, is confined to obstruction precedence, and does not appear against the learned baselines. Natural budget outcomes still have no mechanism labels and remain 95–100% unresolved under the frozen instrument, even though a matched no-fit persistence baseline shows that the same future-outcome target is predictable on several arms. T1/T3 detect realized violation, and paired faults persist until removal by design.

T2 requires exact or exchangeably covered candidate costs; its conformal result is marginal and fails under the tested shift. Proposition 1 is static, and Proposition 3 is only a necessary unlabeled-target check. Content hashes rather than repository history authenticate the anonymous bundle. We claim no human, hardware, cyber-physical, deployment, or safety-certificate evidence.

Conclusion. The studies do not establish superiority over learned baselines; the gap over the fixed severity rule comes from obstruction precedence. Natural outcomes remain mostly unresolved despite an informative persistence baseline. This is a falsification record for one instrument, not evidence of natural causal diagnosis, a general audit methodology or deployment safety.

REFERENCES

Ezio Bartocci, Yliès Falcone, Adrian Francalanza, and Giles Reger. Introduction to runtime verification. In *Lectures on Runtime Verification*, pp. 1–33. Springer, 2018. doi: 10.1007/978-3-319-75632-5_1. URL https://doi.org/10.1007/978-3-319-75632-5_1.

Nicola Bastianello, Ruggero Carli, Luca Schenato, and Marco Todescato. Asynchronous distributed optimization over lossy networks via relaxed ADMM: Stability and linear convergence. *IEEE Transactions on Automatic Control*, 66(6):2620–2635, 2021. doi: 10.1109/TAC.2020.3011358.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: 10.1561/2200000016.

Zhiyuan Cai, Huanhui Cao, Wenjie Lu, Lin Zhang, and Hao Xiong. Safe multi-agent reinforcement learning through decentralized multiple control barrier functions. arXiv:2103.12553, 2021.

Tsung-Hui Chang, Mingyi Hong, Wei-Cheng Liao, and Xiangfeng Wang. Asynchronous distributed ADMM for large-scale optimization–part I: Algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016. doi: 10.1109/TSP.2016.2537271.

Ingy Elsayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2021. URL https://arxiv.org/abs/2101.11196.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL https://jmlr.org/papers/v24/21-0048.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1705.08500.

Shangding Gu, Jakub Grudzien Kuba, Muning Wen, Ruiqing Chen, Ziyan Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. Multi-agent constrained policy optimisation. arXiv:2110.02793, 2021.

Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319(C):103905, 2023. doi: 10.1016/j.artint.2023.103905.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11694.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8767–8775, 2021. doi: 10.1609/aaai.v35i10.17062.

Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. Learning safe multi-agent control with decentralized neural barrier certificates. arXiv:2101.05436, 2021. URL https://arxiv.org/abs/2101.05436.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jaeger. Overcoming common flaws in the evaluation of selective classification systems, 2024. URL https://arxiv.org/abs/2407.01032.

Xuefeng Wang, Lei Zhang, Henglin Pu, Husheng Li, and Ahmed H. Qureshi. Safe continuous-time multi-agent reinforcement learning via epigraph form. arXiv:2602.17078, 2026.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative, multi-agent games. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2103.01955.

Lijun Zhang, Lin Li, Wei Wei, Huizhong Song, Yaodong Yang, and Jiye Liang. Scalable constrained policy optimization for safe multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-4400. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fa76985f05e0a25c66528308dda33de0-Abstract-Conference.html.

Songyuan Zhang, Oswin So, Kunal Garg, and Chuchu Fan. GCBF+: A neural graph control barrier function framework for distributed safe multiagent control. *IEEE Transactions on Robotics*, 41: 1533–1552, 2025. doi: 10.1109/TRO.2025.3530348.

A Reproduction Contract

The anonymous supplement includes `simulation/f42_reproduction_manifest.json`. From its root, `python-msimulation.f42_reproduction_verify--check-dependencies` verifies 55 files, 18 exact MPE2 versions, eight evidence classes, and 32 headline metrics; any missing file, changed byte, unbound claim, value mismatch, or version mismatch exits nonzero. The manifest binds source, protocols, raw or row-level results, tests, execution logs, and independent verifier receipts. Content hashes replace repository-history authentication in the anonymous package.

`python-msimulation.run_planted_diagnostic_validation` rebuilds the planted and paired outputs; `python-msimulation.run_diagnostic_robustness_validation` rebuilds ambiguity, calibration, learned comparators, replay, shift, and candidate-error results. The repository reproduction record gives the exact MPE2 full-run and verification commands and maps every evidence class to its output. The budget sweep result embeds hashes for all 24 raw source files. MPE2 uses CPython 3.11.15; core audit modules use CPython 3.14.6 plus NumPy 2.5.1 because they contain Python 3.12+ type aliases. Both requirement files are hash-bound. Legacy heterogeneous rows are explicitly archival context, not evidence for accuracy, transfer, superiority, or recovery.

B Related-work and benchmark-transport boundaries

External rows are boundary tests rather than replications of the synthetic classes. Grid protocols differ in budget and evaluation mode exactly as Table 5 records; high-N active-budget rows are adverse, and the MPE2 result uses an adapter-defined endpoint. There is no independent replication, method-faithful MACPO comparison, official MPE2 safety endpoint, or deployment study.

Constrained-MARL optimizers return policies or optimization guarantees Gu et al. (2021; 2023); Lu et al. (2021); Zhang et al. (2024); Wang et al. (2026); safety filters and shields return interventions or certificates Qin et al. (2021); Zhang et al. (2025); Elsayed-Aly et al. (2021); Cai et al. (2021). ADMM residuals measure optimization agreement Boyd et al. (2011); Bastianello et al. (2021); Chang et al. (2016), while runtime verification motivates a rigorous monitoring vocabulary Bartocci et al. (2018). Our narrower output is an empirical diagnosis of emitted telemetry, with seed-level rather than episode-level inference Henderson et al. (2018); it is neither policy optimality nor closed-loop invariance.

Selective versus generalized risk. The resolved/wrong verdict counts reported above instantiate standard selective-classification quantities rather than a newly invented evaluation format Geifman & El-Yaniv (2017); Franc et al. (2023). With selection indicator $g = \mathbf{1}\{\text{resolved}\}$ and coverage $\mathrm{E}[g]$, the *selective risk* conditions on resolution, $\mathrm{E}[\ell\, g]/\mathrm{E}[g]$, while the *generalized risk* is the unconditional error-and-selected rate, $\mathrm{E}[\ell\, g]$; the two differ by the coverage factor and are not interchangeable. Our $42/322 = .1304$ is the selective risk at $322/480$ coverage, and $42/480 = .0875$ is the corresponding full-sample generalized risk; we do not call the latter a selected-only or false-selection rate. T1–T3 supplies a single frozen operating point per run, not a continuous confidence

ranking, so no risk–coverage curve or AURC/AUGRC statistic is computed here. The distinction between selective and generalized risk underlying AUGRC addresses a related but not identical evaluation issue Traub et al. (2024); our correction (Table 6) scores abstention as neither correct nor wrong rather than constructing an AURC/AUGRC curve. The foldwise matched-coverage comparison above applies this standard selective-risk definition to previously collected verdicts and established no superiority or equivalence claim; it is a reanalysis under a cited definition, not a new evaluation method.

C Analytic Candidate-Set and Identifiability Results

Proposition 1 (Candidate-set obstruction with bounded cost error) *For $d \geq 0$ and finite nonempty action sets with finite candidate costs, take $\ell_i = \max\{0, \min_a c_i^{\mathrm{cal}}(a)\}$. If $\sum_i \ell_i > d$, every $b \in \Delta(d)$ leaves some $F_i(b_i) = \{a : c_i^{\mathrm{cal}}(a) \leq b_i\}$ empty. More generally, suppose observable bounds $0 \leq L_i \leq \ell_i \leq U_i$ hold jointly (negative lower endpoints may be replaced by zero). If $\sum_i L_i > d$, obstruction is certain; if $\sum_i U_i \leq d$, $b_i = \ell_i$ witnesses simultaneous nonempty sets in $\Delta(d)$; for an equality interface $\mathbf{1}^\top b = d$, assign unused slack $d - \sum_i \ell_i$ to any coordinate. Otherwise either state is compatible with the bounds and the verdict must remain unresolved. If a floored gate uses $\widetilde{b}_i = \max(b_i, m)$, accepted actions imply only $\sum_i c_i^{\mathrm{cal}}(a_i) \leq d + \sum_i (m - b_i)_+$.*

Proof. If all F_i were nonempty, $\ell_i \leq b_i$ for every agent, so $\sum_i L_i \leq \sum_i \ell_i \leq d$, contradicting the first condition. Under the second condition, $b_i = \ell_i$ belongs to the inequality simplex and admits an argmin action for each finite set. If equality is required, increasing one b_i by the nonnegative unused slack preserves every admitted argmin and makes the sum exactly d. In the intermediate case, choosing admissible minima on opposite sides of d constructs both states. Summing accepted floored inequalities gives the last claim. □

Proposition 2 (Finite-sample conformal T2 validity) *Let $S = \sum_i \ell_i$ with the nonnegative effective minima above, fix $0 < \alpha < 1$, and let $\widehat{S}$ be its estimate, and let $E_j = |\widehat{S}_j - S_j|$ for n calibration runs. Suppose $(E_1, \ldots, E_n, E_{n+1})$ are exchangeable. For $k = \lceil (n+1)(1-\alpha) \rceil$, let q be the kth calibration order statistic (or $+\infty$ if $k > n$). Report obstruction when $\widehat{S} - q > d$, feasibility when $\widehat{S} + q \leq d$, and* UNRESOLVED *otherwise. Then the marginal probability of an incorrect resolved verdict on run $n+1$ is at most α.*

Proof. Either incorrect resolved verdict implies $|\widehat{S}_{n+1} - S_{n+1}| > q$: a false obstruction has $S \leq d < \widehat{S} - q$, and a false feasibility verdict has $\widehat{S} + q \leq d < S$. By exchangeability, the probability that the test score exceeds the corrected calibration order statistic is at most $(n+1-k)/(n+1) \leq \alpha$; ties only reduce it. The statement is marginal, not conditional on resolution, and does not apply after nonexchangeable shift. □

Proposition 3 (Unlabeled-target validation lower bound) *Let X contain every telemetry variable observed on a target population and let a frozen selective diagnostic be $h(X) \in \{0, 1, \perp\}$. Write $R = \{h(X) \neq \perp\}$ and suppose $\Pr(R) = q > 0$ and $\Pr(h(X) = 1 \mid R) = 1/2$. For every target sample size n, there exist two laws P_+ and P_- on (X, Y) with the same marginal law of X, the same coverage q, and $\Pr(Y = 1) = \Pr(Y = 1 \mid R) = 1/2$, but*

$$\Pr_{P_+}(h(X) = Y \mid R) = 1, \qquad \Pr_{P_-}(h(X) = Y \mid R) = 0.$$

Consequently, for every estimator $\widehat{a}(X_1, \ldots, X_n)$ based only on unlabeled target telemetry,

$$\max_{\nu \in \{+,-\}} \mathbb{E}_{P_\nu} |\widehat{a} - a(P_\nu)| \geq \frac{1}{2}, \quad a(P) = \Pr_P(h(X) = Y \mid R).$$

Proof. Keep the same law of X in both worlds. On R, set $Y = h(X)$ under P_+ and $Y = 1 - h(X)$ under P_-. On R^c, draw $Y \sim \mathrm{Bernoulli}(1/2)$ conditionally independently in both worlds. Balanced resolved predictions give resolved and overall positive prevalence $1/2$ under both laws, while their resolved accuracies are one and zero. Because every unlabeled sample has the same distribution in both worlds, $\widehat{a}$ has the same distribution; pointwise, $|\widehat{a} - 1| + |\widehat{a}| \geq 1$, so one expected error is at least $1/2$. □

The executable four-cell witness uses coverage $q = .5$, equal mass on two resolved and two abstained telemetry cells, and writes both confusion matrices. It is a lower bound for unlabeled target validation, not evidence that any particular natural-failure corpus realizes either adversarial law.

Proposition 4 (Exact finite-telemetry identifiability boundary) *Let the latent mechanism $M \in \{1, \dots, K\}$ have prior $\pi_m > 0$, let Z contain all available telemetry, and let $Q = f(Z)$ be any finite-valued flag or verdict produced by threshold conjunctions and precedence. Write μ_m for the law of Z conditional on $M = m$, $F_q = \{z : f(z) = q\}$, and $\lambda = \sum_m \mu_m$, with densities $p_m = d\mu_m/d\lambda$. The minimum mechanism-classification errors using Z and using only Q are exactly*

$$e_Z^* = 1 - \int \max_m \{\pi_m p_m(z)\}\, d\lambda(z), \qquad e_Q^* = 1 - \sum_q \max_m \{\pi_m \mu_m(F_q)\}.$$

Consequently, $e_Q^* \geq e_Z^*$, with equality if and only if, for every non-null cell F_q, some single m_q maximizes $\pi_m p_m(z)$ for λ-almost every $z \in F_q$. Moreover, $e_Q^* = 0$ if and only if every output cell has positive probability under at most one mechanism. For two equiprobable mechanisms this matching bound is

$$e_Q^* = \frac{1}{2} \sum_q \min\{\mu_1(F_q), \mu_2(F_q)\} = \frac{1 - \mathrm{TV}(f_{\#}\mu_1, f_{\#}\mu_2)}{2}.$$

Thus a shared positive-probability flag cell is sufficient to rule out perfect mechanism identification, while disjoint output supports are sufficient; the result concerns specified mechanism laws, not causal discovery from unlabeled telemetry.

Proof. Given $Q = q$, a Bayes classifier chooses a mechanism maximizing $\pi_m \mu_m(F_q)$, so its total probability of a correct decision is $\sum_q \max_m \pi_m \mu_m(F_q)$; this proves the formula for e_Q^*. With all of Z, the same pointwise Bayes argument and the common dominating measure λ give correct-decision probability $\int \max_m \pi_m p_m \, d\lambda$, proving the formula for e_Z^*.

For each cell,

$$\int_{F_q} \max_m \{\pi_m p_m\}\, d\lambda \;\geq\; \max_m \int_{F_q} \pi_m p_m \, d\lambda = \max_m \{\pi_m \mu_m(F_q)\}.$$

Summing proves $e_Q^* \geq e_Z^*$. Choose an index m_q attaining the right-hand maximum. Equality on that cell holds exactly when the nonnegative integrand $\max_m \pi_m p_m - \pi_{m_q} p_{m_q}$ vanishes almost everywhere; this is precisely the stated common-maximizer condition. Because the cellwise gaps are nonnegative, equality in the sum holds exactly when it holds in every non-null cell.

Finally, the total prior mass in a cell is $\sum_m \pi_m \mu_m(F_q)$. The difference between this quantity and its largest summand is zero exactly when at most one summand is positive. Summing over q, and using $\pi_m > 0$, proves the zero-error characterization. In the binary equiprobable case the error in each cell is one half of the smaller conditional mass. Summing these minima and applying $\mathrm{TV}(f_{\#}\mu_1, f_{\#}\mu_2) = \frac{1}{2} \sum_q |\mu_1(F_q) - \mu_2(F_q)|$ yields the displayed identity. □

Proposition 5 (Post-hoc detection-delay frontier and the pre-action bypass) *Fix a horizon H, a change point τ, rates $0 < p_0 < p_1 \leq 1$, and write $n = H - \tau + 1$. Let the violation stream $v_1, \dots, v_H$ be independent, with P_0 making every v_t Bernoulli(p_0) and P_τ making v_t Bernoulli(p_1) for $t \geq \tau$ and Bernoulli(p_0) otherwise. A* post-hoc monitor *is a stopping time T for $\mathcal{F}_t = \sigma(v_1, \dots, v_t)$ obeying the per-step false-alarm budget $\max_{t \leq H} P_0(T = t) \leq \bar{\alpha}$; its delay is $D = T - \tau + 1$ when $\tau \leq T \leq H$ and $D = n$ otherwise. Let $\beta(\varepsilon, i)$ be the randomized Neyman–Pearson power at level ε for $\mathrm{Ber}(p_0)^{\otimes i}$ against $\mathrm{Ber}(p_1)^{\otimes i}$, and let* KL *denote Kullback–Leibler divergence.*

(a) Converse. *Every post-hoc monitor within budget satisfies*

$$\mathbb{E}_\tau[D] \;\geq\; \underline{D}(\bar{\alpha}) := 1 + \sum_{i=1}^{n-1} \big\{1 - \beta(\min(i\bar{\alpha}, 1), i)\big\}.$$

(b) Achievability. *The window monitor* $T_{w,k} = \min\{t \geq w : \sum_{u=t-w+1}^{t} v_u \geq k\}$ *obeys the budget* $\alpha_{w,k} = \Pr\{\mathrm{Bin}(w, p_0) \geq k\}$*, and with* $\pi_{w,k} = \Pr\{\mathrm{Bin}(w, p_1) \geq k\} > 0$ *and* $m = \lfloor n/w \rfloor$*,*

$$\mathbb{E}_\tau[D_{w,k}] \;\leq\; \frac{w}{\pi_{w,k}} + n\big\{(1-\pi_{w,k})^m + (\tau-1)\alpha_{w,k}\big\}.$$

(c) Joint horizon–budget tightness. *For fixed* $p_0, p_1 \in (0,1)$ *and fixed* τ*, along any sequence* $\bar{\alpha} \to 0$ *and* $n = n(\bar{\alpha}) \to \infty$ *with* $n \geq \log^2(1/\bar{\alpha})$ *and* $n\bar{\alpha} = o(\log(1/\bar{\alpha}))$*, both* $\underline{D}(\bar{\alpha})$ *and the least in-budget window delay equal* $(1+o(1))\log(1/\bar{\alpha})/\mathrm{KL}(p_1\|p_0)$*.*

(d) Blindness floor and bypass. *Every post-hoc monitor within budget has* $P_\tau(\tau \leq T \leq H) \leq \min(n\bar{\alpha}, 1)(p_1/p_0)^n$*, and*

$$\underline{D}(\bar{\alpha}) \;\geq\; 1 + \tfrac{1}{2}\min\{n-1, i^\star\}, \qquad i^\star = \max\{i \geq 0 :\; i\bar{\alpha}(p_1/p_0)^i \leq \tfrac{1}{2}\}.$$

For fixed n*, this displayed bound approaches* $1 + \frac{1}{2}(n-1)$ *as* $\bar{\alpha} \to 0$*; it does not diverge. It diverges only jointly with the horizon, for example under the regime in (c). The pre-action monitor* $T^{\mathrm{pre}} = \min\{t : \sum_i \ell_{i,t} > d\}$ *reads candidate minima before the action at* t*, so under the infeasible-budget construction it alarms at* $t = \tau$*, before any post-change violation is realized: its delay is* 0*, while every post-hoc monitor has* $D \geq 1$ *and* $\mathbb{E}_\tau[D] \geq \underline{D}(\bar{\alpha})$*.*

Proof. (a) Since $1 \leq D \leq n$, $\mathbb{E}_\tau[D] = \sum_{j=1}^{n} P_\tau(D \geq j)$ and $P_\tau(D \geq 1) = 1$. For $2 \leq j \leq n$ the event $\{D \leq j-1\}$ is exactly $\{\tau \leq T \leq \tau + j - 2\}$, which lies in $\mathcal{F}_{\tau+j-2}$. Write $i = j-1$. The per-step budget and a union bound over i time points give $P_0(\tau \leq T \leq \tau+i-1) \leq i\bar{\alpha}$. On $\mathcal{F}_{\tau+i-1}$ the density of P_τ with respect to P_0 depends only on $v_\tau, \ldots, v_{\tau+i-1}$, and the remaining coordinates are identically distributed under both laws, so they act as randomization rather than evidence. The Neyman–Pearson lemma therefore bounds the P_τ-probability of any such event by $\beta(\min(i\bar{\alpha}, 1), i)$, and summing over i proves (a).

(b) Since $\{T_{w,k} = t\} \subseteq \{\sum_{u=t-w+1}^{t} v_u \geq k\}$, the per-step false-alarm probability is at most $\alpha_{w,k}$; and $P_\tau(T < \tau) = P_0(T < \tau) \leq (\tau-1)\alpha_{w,k}$ because $\{T < \tau\} \in \mathcal{F}_{\tau-1}$. On $\{T \geq \tau\}$ let G be the first index g whose block $v_{\tau+(g-1)w}, \ldots, v_{\tau+gw-1}$ contains at least k ones. Blocks $g \leq m$ are disjoint and entirely post-change, so G is geometric with success probability $\pi_{w,k}$, and $G \leq m$ forces $T \leq \tau + Gw - 1$, that is $D \leq Gw$. The event $\{T \geq \tau\}$ lies in $\mathcal{F}_{\tau-1}$ and is independent of those blocks, so $\mathbb{E}_\tau[D\mathbf{1}\{T \geq \tau\}] \leq w/\pi_{w,k} + n(1-\pi_{w,k})^m$; charging $\{T < \tau\}$ at the cap adds $n(\tau-1)\alpha_{w,k}$.

(c) For any level-ε test on i post-change observations and any $\eta > 0$, splitting on the log-likelihood ratio and applying a Chernoff bound under P_1 gives $\beta(\varepsilon, i) \leq \varepsilon e^{i(\mathrm{KL}+\eta)} + e^{-ic(\eta)}$ with $c(\eta) > 0$, the moment generating function being finite because the log-likelihood ratio is bounded for $p_0, p_1 \in (0,1)$. Put $I = \lfloor (1-\eta)\log(1/\bar{\alpha})/(\mathrm{KL}+\eta) \rfloor$, which is at most $n-1$ in the stated regime. Then $\sum_{i \leq I} \beta(i\bar{\alpha}, i) \leq I^2\bar{\alpha}^\eta + (1-e^{-c(\eta)})^{-1} = o(I)$, so $\underline{D}(\bar{\alpha}) \geq I(1-o(1))$; letting $\eta \downarrow 0$ gives the lower rate. For the upper rate fix $\rho \in (p_0, p_1)$ and take $w = \lceil \log(1/\bar{\alpha})/\mathrm{KL}(\rho\|p_0) \rceil$ and $k = \lceil w\rho \rceil$; Chernoff bounds give $\alpha_{w,k} \leq e^{-w\mathrm{KL}(\rho\|p_0)} \leq \bar{\alpha}$ and $1 - \pi_{w,k} \leq e^{-w\mathrm{KL}(\rho\|p_1)}$, so (b) yields delay $w(1+o(1))$; letting $\rho \uparrow p_1$ and using continuity of $\mathrm{KL}(\cdot\|p_0)$ matches the constant, and the converse in (a) cannot exceed it.

(d) Applying the argument of (a) with $i = n$ and the elementary bound $P_\tau(A) = \mathbb{E}_0[(dP_\tau/dP_0)\mathbf{1}_A] \leq (p_1/p_0)^n P_0(A)$ gives the detection ceiling. The same likelihood-ratio bound gives $\beta(\min(i\bar{\alpha}, 1), i) \leq \min\{1, i\bar{\alpha}(p_1/p_0)^i\}$, whose right side increases in i and is at most $\frac{1}{2}$ for every $i \leq i^\star$; substituting into (a) proves the floor, and $i^\star \geq \log(1/2\bar{\alpha})/\log(p_1/p_0) - O(\log\log(1/\bar{\alpha}))$ diverges. Hence the displayed minimum saturates at $n-1$ for fixed n, while it diverges under the joint regime in (c). The pre-action statement follows from Proposition 1: the infeasible-budget intervention makes $\sum_i \ell_{i,t} > d$ from $t = \tau$ onwards, and $\ell_{i,t}$ is read before the action at t. □

Where the instrument sits on the frontier. To contextualize the observed delays, we fit Proposition 5's iid Bernoulli model by plugging in timestep-averaged rates from the adaptive, temporally dependent planted simulator: $p_0 = .0103$, $p_1 \in \{.905, 1\}$, $H = 240$, and $\tau = 80$. Thus every tail,

converse, and expected-delay number in this paragraph is an exact calculation under the fitted iid surrogate, not a guarantee for the simulator. Under that surrogate the $w = 12, k = 6$ declaration gate has per-step false-alarm probability 1.046×10^{-9} and expected delay 6.56 steps at $p_1 = .905$ and 5.93 at $p_1 = 1$, against converse values 5.24 and 4.63. It is therefore within 1.3 surrogate steps of the floor. The simulator's observed T3 and T1 lags remain three and 28 steps past the declaration clock; comparing them with the surrogate floor gives factors about 1.8 and 6.6, not guarantees for post-hoc telemetry. The same fitted calculation bounds the pre-action per-step false-alarm probability by 1.24×10^{-125} using the specified Gaussian state law, and gives post-hoc delay floors 73.7 and 62.0, while the constructed pre-action delay is zero. T2 is off the post-hoc frontier only within this surrogate comparison. At fixed H the delay gap is capped; divergence requires the joint horizon–budget limit in part (c).

An earlier form of (d) claimed exactly zero detection below $\bar{\alpha} < p_0^n/n$; that is false because pre-change observations supply free randomization. As an exact counterexample within the iid model, take $p_0 = .2$, $p_1 = .9$, $H = 14$, $\tau = 11$, and $\bar{\alpha} = 2 \times 10^{-4}$. Stopping at τ when the first five observations and v_τ are all one has per-step false alarm 6.4×10^{-5} and detection probability $2.88 \times 10^{-4} > 0$. The quantitative ceiling in (d) replaces that claim. The script `simulation/verify_delay_barrier.py` recomputes the surrogate and counterexample arithmetic exactly and checks the theorem on 520 randomly drawn parameter instances: 3,853 exact checks pass, the tightest converse slack is .032, and three deliberate mutations fail. Those checks verify the iid theorem and arithmetic, not iid calibration of the adaptive simulator.

D Complete Synthetic Update Specification

At each step, actor i samples $a_{i,t} \sim \text{Bernoulli}(p_{i,t})$ with $p_{i,t} = \sigma(\theta_{i,t} + .25 s_{i,t})$. Let $\Delta c_{i,t}$ be risky minus safe candidate cost and $g_t = \sum_i c_{i,t} - d$. The responsive update, in execution order, is

$$\lambda_{t+1} = \min\{\lambda_{\max}, \max(0, \lambda_t + .35 g_t)\}, \tag{4}$$

$$\theta_{i,t+1} = \text{clip}_{[-5,5]}(\theta_{i,t} + .12[.42 - \lambda_{t+1}\Delta c_{i,t}] p_{i,t}(1 - p_{i,t}))\,, \tag{5}$$

$$q_{i,t} = c_{i,t}^{\text{safe}} + p_{i,t}\Delta c_{i,t} + .02 \sin(i + t/7), \quad b_t = R_d(q_t), \quad r_t = \|q_t - b_t\|_2/d. \tag{6}$$

Stall freezes the first two updates after the logit shock; saturation applies the same actor update with $\lambda_{\max} = .20$; infeasibility changes the safe-cost law stated in the main text. All exogenous arrays are drawn before these updates. Calibration enumerates the stated Cartesian grid in product order and applies the fixed tie key above. The anonymous package's hash-bound source fully specifies the ambiguous-telemetry generator, selective-centroid calibration, replay construction, and candidate-cost error models; duplicating those implementations here would not add a second specification.

E Frozen-Threshold Transport Measurement

Table 7: Adverse MAPPO-Lagrangian neural-anchor audit after 200,000 environment steps per seed. Actors execute decentralized policies; training uses centralized value and cost critics. High-N rows are not evidence of safety, scaling, optimizer failure, or CTDE impossibility.

Regime	N	Seeds	Final V	Final cost/d	Final λ
Slack ($d = 100$)	20	5	0.000	—	0.00
Binding ($d = 5$)	20	5	0.832	1.12	8.10
Binding ($d = 5$)	50	10	1.000 (10/10)	4.69 ± 1.06	25.0 (10/10)
Binding ($d = 5$)	100	10	1.000 (10/10)	11.37 ± 0.82	25.0 (10/10)

Each final policy uses eight deterministic evaluation episodes. At $N = 20$, the binding-row 95% interval for final V is $[0.650, 0.957]$. High-N $\pm$ is sample SD across seeds; all high-N seeds reach the configured dual cap.

The transport test in the main text reuses the released `g3_safe_mpe_scalar_budget_raw.csv` trace of Farama MPE2 Simple Spread and applies the frozen rule (window 12, $\bar{r} \le .05$, $\bar{v} \ge .50$, $\lambda/\lambda_{\max} \ge .98$) without refitting. Channels are mapped as $\bar{r}$ to mean observed residual, $\bar{v}$ to violation rate, $\lambda/\lambda_{\max}$ to mean shield saturation, and the obstruction channel to a zero feasible-action count at window end, withheld where that channel is identically zero. Each method contributes 945 windows over five seeds.

Table 8: Packaged high-N RCBS rows recomputed from raw CSVs. Only $N = 50$ has binding evidence; $N = 100$ is slack and cannot support a binding claim.

Regime	N	d	Seeds	Mean V	Final V	Cost ratio
Binding	50	5	3	1.000	1.000	4.539
Slack	100	500	5	.011	.000	.173
Binding	100	—	0	—	—	unavailable

Table 9: Action-overwrite construct check across three MPE2 environments, two learners, and 20 model seeds (480 runs). Learned baselines train on the two environments. No T1–T3 accuracy contrast is Holm-significant.

Method	Accuracy	Macro R	Healthy R	Stall R	Sat. R	Obstr. R
Frozen T1–T3	.9875	.9875	.9500	1.000	1.000	1.000
Copied instrument labels	.9875	.9875	.9500	1.000	1.000	1.000
Depth-3 tree	.9896	.9896	.9583	1.000	1.000	1.000
Selective centroid	.9833	.9833	.9333	1.000	1.000	1.000

Paired alarm recovery	Mean difference	Model-seed bootstrap 95% CI	Exact p
Post-trigger team reward	$+190.57$	$[136.75, 239.52]$	1.91×10^{-5}
Post-trigger violation rate	$-.9538$	$[-.9775, -.9276]$	1.91×10^{-6}

Every window is UNRESOLVED and no window meets the violation gate. The largest $\bar{v}$ observed is .076667 for the budget method, .040000 for random, and .080000 for RCBS, against the required .50; the largest dual ratio is .560000, against the required .98. T2 never fires for the budget and RCBS methods even though their obstruction channel is defined, and no window fires T1 or T3. Thus the measured exception count is zero; no prospective registration claim is made. The runner is `simulation/run_mpe2_diagnostic_transfer.py`, its test is `simulation/code/tests/test_mpe2_diagnostic_transfer.py`, and the receipt is `simulation/results/mpe2_diagnostic_transfer.json`. This is a single public environment and a single released trace, and the shield-saturation proxy for $\lambda/\lambda_{\max}$ is an explicit modelling choice; the measurement therefore bounds transport of the frozen thresholds rather than establishing accuracy against natural failure labels.

F MECHANISM-LABELED MPE2 TRANSFER CONTRACT

The recorded protocol fixes MPE2 1.1.0 Simple Spread (3 agents, observation dimension 18), Simple Formation (4, 6), and Simple Line (4, 8), all with five discrete actions. For each environment and seed 2000–2019, independent REINFORCE uses one linear softmax actor per agent, 1,000 training steps, learning rate .01, and a running return baseline. The separate linear evolution-strategy implementation uses 20 antithetic iterations, 1,000 total environment steps, perturbation scale .05, and learning rate .02. Evaluation uses seed $100000+$model seed and never refits T1–T3.

Each 120-step evaluation has 24 warmup and 96 intervention steps. Healthy executes the learned action; stall replaces all actions by no-op; obstruction exposes an empty candidate set and executes no-op; saturation replaces all actions by one seed-fixed nonzero action. The label is stored outside the extractor. Residual is the fraction of executed actions changed from the prior step; violation is one when all actions are no-op or the same nonzero action; dual ratio is the saturation-replacement share; feasibility is zero only for the empty candidate set. Frozen thresholds and precedence are $.05/.50/.98/12$ and T2>T3>T1.

The diagnostic grid has $3 \times 2 \times 20 \times 4 = 480$ runs. Depth-3 CART and the selective centroid train on the other two environments; the stored direct control copies instrument labels rather than independently classifying the four fields. Metrics aggregate within model seed, then across environment–learner cells; model-seed bootstrap uses 10,000 replicates. For each nonhealthy run, a paired replay either keeps the intervention or removes it after the first correct full-window alarm, yielding 360 recovery pairs. Complete rows are in `simulation/results/f42_labeled_mpe2_transfer_v1.json`; the verifier recomputes every aggregate. These intervention labels are not natural causal labels.

Table 10: Untouched-seed V2 transfer for independently scheduled controller faults (three MPE2 environments, two learners, 20 model seeds; 480 runs). Tree and centroid train leave-one-environment-out. Direct telemetry is the fixed no-fit severity argmax specified in Appendix G.

Method	Accuracy	Macro R	Healthy R	Stall R	Sat. R	Obstr. R
Frozen T1–T3	.7667	.7667	.7333	.7250	.6083	1.000
Direct telemetry	.5646	.5646	.7333	.7250	.6083	.1917
Depth-3 tree	.7417	.7417	.3083	.9000	.7583	1.000
Selective centroid	.7396	.7396	.7083	.7250	.6833	.8417

Contrast (T1–T3 minus baseline)	Mean	Seed-bootstrap 95% CI	Holm p
Direct telemetry	$+.2021$	$[.1833, .2208]$	5.72×10^{-6}
Depth-3 tree	$+.0250$	$[-.0021, .0500]$	.2098
Selective centroid	$+.0271$	$[.0000, .0563]$	.2098

Paired recovery (descriptive)	Triggered / pairs	Mean among triggered
Alarm-triggered fault removal	$304/360$	reward $+74.02$; violation $-.5378$

G Independently Scheduled Controller-Fault Transfer

A preliminary smoke check exposed a scheduler defect (the nominal stall continued dual updates), wrote no production artifact, and supports no claim. The completed protocol corrects that fault mechanism without changing T1–T3 and freezes untouched seeds 4000–4019 before outcomes. For each pretrained controller, a common post-step-40 candidate-cost shift is applied. Healthy control continues both updates; stall freezes safety-bias and dual updates; saturation keeps both laws but lowers the hidden cap from 5 to .05; obstruction lowers budget below the latent sum of candidate minima.

The extractor signature has no label, scheduler, recovery, or expected-verdict argument. It computes normalized parameter-update residual, realized noisy cost violation, dual/cap, and feasibility from noisy observed candidate costs. Three environments, two learners, 20 seeds, and four labels yield 480 diagnostic runs; each nonhealthy cell also has persistent/recovery replays. Tree and centroid train only on the other two environments. The direct comparator uses no fitting. Writing $x = (r, v, \lambda/\lambda_{\max}, o)$, it returns healthy if $v < .5$ and $o < .5$; otherwise it takes the deterministic argmax over

$$\big(0,\ v\max\{0, 1 - r/.05\},\ v\lambda/\lambda_{\max},\ o\big)$$

in the order healthy, stall, saturation, obstruction. It is a transparent severity baseline rather than a strong classifier: its recall differs from T1–T3 only on obstruction, so the paper does not interpret their significant contrast as broad diagnostic superiority. All rows, model weight hashes, and the recomputed analysis are in `simulation/results/f42_independent_fault_mpe2_v2.json`; the independent verifier closes 120 models, 480 diagnostic cells, and 360 recovery pairs. Recovery means in Table 10 use only the 304 triggered pairs and are descriptive. The accuracies in that table and in Table 9 score UNRESOLVED as HEALTHY (`VERDICT_TO_LABEL["UNRESOLVED"]=0`); the selective-verdict recomputation in Table 6 is generated by `simulation/selective_verdict_reaggregation.py` from the same stored runs and its naive column matches these tables to machine precision (`simulation/results/f42_selective_verdict_reaggregation_v1.json`). In the labeled study the stored control, labeled "Copied instrument labels" in the tables, is a copy of the instrument's verdicts. Its equality with T1–T3 under the stored healthy-label mapping is not an independent baseline comparison.

H Calibration-Targeted Selective-Risk Reanalysis

This analysis is post hoc and uses only the fixed 480-run synthetic scheduled-fault record, all of which had been inspected before the analysis. For each leave-one-environment-out fold, seeds 4000–4014 in the other two environments fit the tree and centroid, and seeds 4015–4019 calibrate confidence cutoffs and deterministic boundary-tie probabilities. The resulting selectors are applied unchanged to the held-out environment; neither held-out labels nor held-out score quantiles set them, and exact held-out coverage is not forced. Seed identifiers recur across environments, although the

Table 11: Post-hoc calibration-targeted selective risk on the fixed 480-run scheduled-fault record. Positive Δ risk is comparator minus frozen. Intervals are paired model-seed-cluster bootstrap 95% intervals conditional on the fitted selectors, fixed synthetic data, and three observed environments. The foldwise $\pm.05$ coverage gate determines the final column.

Method	resolved/n	coverage	risk [95%]	Δ risk [95%]	Δ coverage	all folds matched?
Frozen T1–T3	322/480	0.6708	0.1304 [0.1129, 0.1477]	–	–	reference
Direct telemetry	332/480	0.6917	0.3645 [0.3168, 0.4134]	+0.2340 [+0.1909, +0.2756]	+0.0208	no
Depth-3 tree	315/480	0.6562	0.1365 [0.1139, 0.1587]	+0.0061 [-0.0199, +0.0313]	-0.0146	yes
Centroid	302/480	0.6292	0.1490 [0.1125, 0.1853]	+0.0186 [-0.0185, +0.0561]	-0.0417	no

Table 12: Held-out-environment fold summaries. Each fold has $n = 160$. C is coverage, R is selective risk, and differences are comparator minus frozen. The tree has adverse risk differences on Formation and Line; no favorable fold is omitted. Direct and centroid columns show why their pooled coverage gaps do not pass the every-fold gate.

Held-out env.	frozen res.	frozen C	frozen R	tree res.	tree C	tree R	tree ΔR	ΔC tree/direct/centroid
Formation	94	0.5875	0.1170	94	0.5875	0.1596	+0.0426	+0.0000/+0.2125/-0.0438
Line	112	0.7000	0.0893	107	0.6687	0.1121	+0.0229	-0.0312/-0.1000/+0.0125
Spread	116	0.7250	0.1810	114	0.7125	0.1404	-0.0407	-0.0125/-0.0500/-0.0938

runs and trained models are distinct. Clustering rows with the same identifier prevents a row-level leak, but this is an environment holdout rather than an unseen-ID test.

The $\pm.05$ gate is evaluated in every fold, not from pooled coverage alone. Thus only the tree is called coverage-matched: its foldwise coverage differences are $0, -.0313, -.0125$. Direct telemetry $(+.2125, -.1000, -.0500)$ and centroid $(-.0438, +.0125, -.0938)$ fail the gate. The tree's foldwise risk differences are adverse on Formation $(+.0426)$ and Line $(+.0229)$, and favorable on Spread $(-.0407)$; Table 12 retains all three. The aggregate $+.0061$ effect and $[-.0199, .0313]$ interval therefore neither show superiority nor establish equivalence.

The paired model-seed-cluster bootstrap holds trained models, calibration rows, cutoffs, and tie probabilities fixed; there is no refitting inside a replicate. Its interval is conditional on these fitted selectors, this fixed synthetic dataset, and the three observed environments. It includes neither training, recalibration, new-seed, nor external-population uncertainty. Fully enumerated paired sign-swap diagnostics are secondary: their interpretation requires within-seed method-label exchangeability, or the corresponding sign symmetry under the sharp null; enumeration does not make them unconditional exact significance tests. The anonymous supplement retains the protocol, post-analysis inference amendment, input, code, tests, full forced-scoring and native-centroid results, class summaries, and every prespecified risk–coverage grid point. From its root, `python-msimulation.matched_coverage_selective_risk` reproduces the JSON and Markdown, and `python-msimulation.render_matched_coverage_tables--check` verifies these tables against that JSON.

I Recorded Budget Sweep Results

All recorded arms are tabulated regardless of outcome, although the submission does not independently authenticate the record's chronology. The completed corpus contains all 20 MAPPO and 4 RCBS arms with zero recorded driver failures; `simulation/budget_sweep_corpus.py` scores those complete arms.

The matched persistence comparator uses the same rows, current and future window lengths, and future-failure labels. It has no fitted parameter: the current-window mean cost/budget is compared with the physical boundary one. Its complete eight-arm confusion counts and all 24 source hashes are emitted by `simulation/natural_outcome_direct_baseline.py`; the checked receipt is `simulation/results/natural_outcome_direct_baseline_20260905.json`.

Table 13: Recorded budget sweep with T1–T3 fixed during evaluation (window 12, label: next disjoint 12-unit mean cost/budget > 1). Counts are overlapping windows. Brackets are window-based Wilson ranges at nominal 95%, not seed-cluster confidence intervals; no 95% coverage is asserted under this dependence. Lead is the current-window endpoint of the first future-failure label minus the endpoint of the first alert, in logged step/episode units (negative = late relative to that label). This is not the time of the first physical budget crossing. Seeds are the clustering unit; inference is descriptive. Artifact `simulation/results/budget_sweep_corpus_20260903.json`.

Method	limit	seeds	over/under	TP/FP/TN/FN	sensitivity	specificity	precision	alert	lead
MAPPO-Lag.	3	5	858/7	14/6/1/844	0.02 [0.01,0.03]	0.14 [0.03,0.51]	0.70 [0.48,0.85]	0.023	-163
MAPPO-Lag.	5	5	490/375	0/23/352/490	0.00 [0.00,0.01]	0.94 [0.91,0.96]	0.00 [0.00,0.14]	0.027	-101
MAPPO-Lag.	10	5	165/700	5/37/663/160	0.03 [0.01,0.07]	0.95 [0.93,0.96]	0.12 [0.05,0.25]	0.049	-56
MAPPO-Lag.	20	5	18/847	10/12/835/8	0.56 [0.34,0.75]	0.99 [0.98,0.99]	0.45 [0.27,0.65]	0.025	-8
RCBS	5	10	277/123	0/0/123/277	0.00 [0.00,0.01]	1.00 [0.97,1.00]	–	0.000	–
RCBS	10	10	15/385	0/0/385/15	0.00 [0.00,0.20]	1.00 [0.99,1.00]	–	0.000	–
RCBS	20	10	0/400	0/0/400/0	–	1.00 [0.99,1.00]	–	0.000	–
RCBS	40	10	0/400	0/0/400/0	–	1.00 [0.99,1.00]	–	0.000	–

J External Protocol Contracts

The release audit recomputes protocol facts from packaged raw CSV rows where those rows are packaged; rows whose raw artifacts are not packaged remain descriptive boundaries, as stated below. The packaged $N = 20$ RCBS active-budget file uses `ScalableMARLEnv`, $N = 20$, horizon 80, $d = 5$, seeds 0–9, 5,000 online transitions/seed, ADMM $\rho = .45$, and dual step .02. The packaged high-N RCBS files use horizon 120 and 5,000 online transitions: $N = 50, d = 5$, seeds 0–2 is binding; $N = 100, d = 500$, seeds 0–4 is slack. Thus no packaged $N = 100$ binding RCBS result exists.

The packaged MAPPO-Lagrangian CSV uses the same scalable environment interface, $d = 5$, $N \in \{50, 100\}$, seeds 0–9, 200,000 training steps, and eight final-policy evaluation episodes; candidate minima are absent. The MPE2 v3 receipt uses Simple Spread, paired seeds 11001–11010, 200,000 steps, adapter budget .600, and the adapter-defined violation endpoint; its MAPPO-minus-RCBS interval remains $[-.06619, .01719]$. Table 5 supplies descriptive protocol fields. External training dependency closures are incomplete, so those legacy rows are archival boundaries rather than independently reproducible release claims.

K Scope, Limitations, and Reproducibility

The planted suite tests deliberately separated conditions in one synthetic dynamics family. Perfect construct consistency does not imply natural-failure accuracy. T1's .05 threshold is not justified by the heterogeneous grid residual; the local calibration instead selects .05/.40/.94/24 on training seeds and remains generator-specific. T1/T3 are late and non-identifying. T2 concerns estimated enumerated candidates, with measured degradation under noise. The matched screen improves synthetic safety only by sacrificing reward.

The grid protocols remain heterogeneous where Table 5 says so. Active-budget results are adverse. MPE2 uses an adapter-defined endpoint. We claim no safety certificate, optimizer dominance, high-N safety, causal scaling, official MPE2 validity, or deployment readiness.

For an individual online run, the monitor computes only the available raw flags and display map: T1/T3 wait for a complete causal-past window, T2 reads current pre-action candidate costs, and missing required telemetry or an undeclared interface returns UNRESOLVED. Artifact hashes, calibration separation, and adverse-control outcomes are release-level validation checks, not quantities re-tested on every run. In continuous action spaces, Proposition 1 applies only if candidate minima have certified bounds; approximate unbounded minimization remains UNRESOLVED. Without labeled target failures there is no justified threshold-selection rule: the frozen values may be audited as research defaults but not claimed to transport.

Natural causal mechanism labels remain unavailable; the independent scheduler supplies experimental labels only. The outcome-only future-window corpus is regenerated create-only by `simulation/natural_failure_corpus.py` from the two packaged telemetry CSVs into

the frozen natural-failure corpus record (2026-09-02); overlapping windows are dependent and are not treated as inferential units. We do not use an unsubmitted release manifest or its suite/build status as paper evidence.

L TRAJECTORY-LEVEL CALIBRATION AND ITS RESOLUTION COST

Pointwise coverage does not control repeated monitoring. The extension changes the calibration unit to a complete trajectory; it does not change the candidate-set target into natural causal diagnosis or realized-action safety.

Setup. Fix a horizon H, a controller, and an estimator trained on separate data before calibration. For trajectory j, agent $i \in \{1, \dots, N\}$, time $t \le H$, and action a in a finite nonempty candidate set $\mathcal{A}_{j,i,t}$, let $c_{j,i,t,a}$ be the reference candidate cost. Define

$$S_{j,t} = \sum_{i=1}^{N} \max\{0, \min_{a \in \mathcal{A}_{j,i,t}} c_{j,i,t,a}\}, \qquad E_j = \max_{t \le H} |\widehat{S}_{j,t} - S_{j,t}|.$$

Here $\widehat{S}_{j,t}$ is an observable pre-action estimate, and $d_{j,t} \ge 0$ is the available budget. The target is obstruction when $S_{j,t} > d_{j,t}$ and candidate feasibility otherwise. Assume the n calibration errors and one test error are exchangeable; within-trajectory independence is unnecessary. Test reference costs belong only to evaluation after prediction, not to the monitor.

Conditional proposition. For $0 < \alpha < 1$, set $k = \lceil (n+1)(1-\alpha) \rceil$. Let q be the kth smallest calibration error, or $+\infty$ if $k > n$. Freeze q before testing. Report obstruction if $\widehat{S}_t - q > d_t$, candidate feasibility if $\widehat{S}_t + q \le d_t$, and abstain otherwise. Then

$$\Pr(\text{any wrong resolved verdict by } H) \le \frac{n+1-k}{n+1} \le \alpha.$$

Proof. On $E_{\text{test}} \le q$, every reference target lies in its reported interval. Either resolution inequality therefore implies its corresponding true target, including equality at the feasibility boundary. Any wrong resolution requires $E_{\text{test}} > q$. Exchangeability bounds this strict exceedance by $(n+1-k)/(n+1)$ through the test error's rank; ties are conservative. For $k = n+1$, the infinite radius prevents resolution. The same event contains errors at a first-alarm stopping time within H. Neither an unbounded horizon nor controller changes that destroy exchangeability inherit this result.

Repeated-testing counterexample. Consider one agent, one zero-cost candidate, zero budget, and estimates independently uniform on $[-1, 1]$. The ideal pointwise radius $1 - \alpha$ gives coverage $1 - \alpha$, but false obstruction probability $\alpha/2$ each step. Independence gives run-level probability $1 - (1 - \alpha/2)^H$, equal to 0.9977032815 for $\alpha = .05$ and $H = 240$. This is a mathematical construction, not an observed deployment frequency.

Verified finite-horizon results. The local trajectory artifact records 500 calibration trajectories and 2,000 evaluation trajectories per condition, each of length 240. Each table entry gives the number of trajectories with any wrong resolution, followed by the percentage of resolved time points among $480{,}000$ points; percentages are rounded.

Condition	Raw	Pointwise	Trajectory
Independent	$2000; 100$	$684; 25.17$	$1; 1.47$
Correlated	$2000; 100$	$668; 26.78$	$3; 2.40$
Persistent	$1996; 100$	$152; 25.35$	$16; 8.56$
Shift	$2000; 100$	$2000; 84.97$	$2000; 51.90$

The first three conditions match calibration and evaluation laws. Their reduced trajectory errors accompany severe abstention, not demonstrated warning utility. Shift breaks the matching assumption and defeats every method. Empirical rates do not prove exchangeability or the proposition; time points are not independent inferential units. Ground truth is simulated candidate-set truth from separate brute-force enumeration, not natural causal fault labels.

Retained boundaries. Original construct and transport failures remain, including zero saturation recall at reduced contrast and zero stall recall on the separate simulator. The present oracle enumerates candidate choices from simulated reference costs, independently of the monitor's minima calculation. It supplies neither natural causal labels nor the T2 telemetry missing from historical outcome traces. Logical observation–prediction–action timestamps do not measure physical latency. Exchangeability after a monitor changes control actions, conditional precision, and field safety remain unestablished.

Reproduction record. The entry point is `simulation/trajectory_conformal_t2.py`. JSON result paths in this appendix are relative to the project evidence bundle (frozen 2026-09-08) at the repository root. Source and configuration bindings accompany the local records; this does not establish independent external replication.

M Seed-Heldout Natural-Outcome Ablation

This new retrospective comparison leaves the canonical frozen evaluator unchanged. MAPPO budgets $3, 5, 10, 20$ use training seeds 100–101 and test seeds 102–104; RCBS budgets $5, 10, 20, 40$ use training seeds 100–104 and test seeds 105–109. A seed stays in the same split across every budget. All 24 raw files and all 60 method–budget–seed strata are required. The original 12-unit current and next disjoint 12-unit windows are reused; neighboring windows still overlap. There are 2,876 heldout windows, not 2,876 independent replicates. All mapped T1/T3 proxy channels are present; their canonical mechanistic semantics are not thereby established. T2 remains unavailable.

Training and gate ablation. Separate future-outcome predictors use the shapes $\bar{r} \le a \wedge \bar{v} \ge b$ and $\bar{d} \ge a \wedge \bar{v} \ge b$, not causal T1/T3 labels. Each grid uses the original threshold, zero, one for unit-interval channels, and training quantiles $0, .1, .25, .5, .75, .9, 1$. We maximize class-balanced correctness after giving each training seed equal total mass; ties prefer squared distance to the original thresholds, then the numeric tuple. No test label enters fitting, grids or tie-breaking. Dropping either the residual gate or the dual gate from the frozen union $(R \wedge V) \vee (D \wedge V)$ gives V; these are identical ablations, not independent baselines.

	MAPPO (3 test seeds)		RCBS (5 test seeds)	
Rule	Error	Balanced acc.	Error	Balanced acc.
Frozen T1	.4538	.4943	.2125	.5000
Frozen T3	.4446	.5000	.2125	.5000
Frozen union	.4538	.4943	.2125	.5000
Calibrated T1 shape	.0409	.9601	.0538	.9379
Calibrated T3 shape	.0222	.9774	.0538	.9379
Calibrated union	.0222	.9774	.0538	.9379
Violation only, .5	.0409	.9607	.0538	.9379
Calibrated violation only	.0222	.9774	.0538	.9379
Cost persistence	.0405	.9613	.0538	.9367

Table 14: Macro averages of per-test-seed metrics, each pooling that seed's four budgets. These are binary alarm forecasts, not healthy-mechanism verdicts. No alarm does not resolve the original instrument's mechanistic abstention.

The selected MAPPO T1 parameters are $(a, b) = (3.2880960452, 1)$ and T3 parameters are $(0, 1)$; thus the T3 branch discards the dual gate. RCBS selects $(.5976072917, .5)$ and $(.10453125, .5)$. The calibrated union and calibrated violation-only rule have exactly identical predictions on every heldout window in both methods. Calibration therefore improves this forecast without establishing extra value from the residual or dual conjunction. This is a threshold/interface-specific finding, not a universal structural impossibility for telemetry prediction.

Uncertainty and reproduction. Paired calibrated-union-minus-persistence error differences are $-.01830$ with approximate seed-t 95% interval $[-.04525, .00864]$ for MAPPO, and approximately

zero with interval $[-.02195, .02195]$ for RCBS. Neither excludes zero. The implementation also reports bounded Hoeffding intervals over independent heldout seed metrics, conditional on the trained rules; for these paired contrasts both intervals are $[-1, 1]$. Only three and five heldout seeds support inference. Undefined class denominators stay undefined, and a zero-variance t interval is unassessed rather than a point interval. These are marginal intervals, not a simultaneous family guarantee.

`simulation/seed_heldout_natural_ablation.py` produces the local seed-heldout analysis record, with complete grids, selected parameters, per-seed confusion counts, gate-pass/block frequencies, source hashes and all test predictions. Seven regressions test split isolation, heldout-label invariance, window timing, confusion counts, bounded seed uncertainty, actual gate-drop equivalence and training-budget completeness. Raw sources are checked before and after analysis; no controller is retrained. This study uses mapped proxies and independent outcome labels, not independently identified natural causal mechanisms.

M.1 Prospective twenty-seed extension of the frozen rules

The rules fitted above stay frozen from 2026-09-08 (hash-bound): on 2026-09-11, 20 fresh heldout seeds per method (200–219) completed the grid declared in the paper-local seed-extension plan (dated 2026-09-11; internal, not externally registered), and the frozen parameters of this appendix were evaluated on them with no refit and no reselection. The anchor recomputation on the original heldout seeds matches every retained metric exactly (1,440 comparisons, zero mismatches). The extension adds 17,040 heldout windows. The 09-11 plan explicitly names these new seeds and frozen rules. The earlier 09-02 sweep plan names seeds 100–104/109, not this extension. Both are internal plans, not externally registered protocols.

Budget	MAPPO calibrated union minus cost persistence [t95]	RCBS same
3 / 5	$-.0046$ $[-.0113, .0021]$	$-.0012$ $[-.0381, .0356]$
5 / 10	$-.0408$ $[-.0457, -.0358]$	$-.0025$ $[-.0077, .0027]$
10 / 20	$-.0197$ $[-.0302, -.0091]$	degenerate; unassessed
20 / 40	$.0029$ $[-.0024, .0082]$	degenerate; unassessed

Table 15: Paired per-seed error differences (negative favors the calibrated union) on 20 prospective heldout seeds per method, with approximate seed-t 95% intervals. The frozen rules themselves remain noncompetitive (MAPPO $+.009$ to $+.98$ against either baseline at assessed budgets; RCBS $+.03$ to $+.69$ at budgets 5–10, with budgets 20/40 unassessed).

The earlier three-/five-seed intervals were inconclusive, not evidence of a tie. In the extension, MAPPO's calibrated-union-minus-persistence pointwise intervals exclude zero at budgets 5 and 10, but not 3 and 20. The assessed RCBS intervals contain zero; the budget-20 and budget-40 strata are outcome-degenerate and remain unassessed. These are marginal intervals without simultaneous coverage; predeclaring contrasts alone does not adjust for multiple comparisons.

The primary pooled-budget differences are $-.01553$ $[-.01959, -.01148]$ for MAPPO and $-.00094$ $[-.01025, .00837]$ for RCBS. A 2026-09-12 audit compares individual binary predictions rather than equal error totals: the calibrated union and calibrated violation-only rule differ on zero of 17,040 extension windows. Thus the observed improvement over persistence is available from the calibrated violation threshold alone, not additional residual or dual conjunctions. Conversely, a cost-persistence interval containing zero does not establish redundancy. These findings are specific to the frozen rules, sampled seeds, mapped proxies and window target.

`simulation/seed_heldout_prospective_extension.py` produces the seed-extension audit record (2026-09-12), retaining every per-seed metric, paired contrast and prediction-change rate, with the frozen rules bound by hash to the earlier analysis and the new raw files hashed before and after. The audit exactly reproduces the original per-seed metrics, macro summaries and paired contrasts, and adds mismatch counts and rates for calibrated union versus calibrated violation. Eleven related regression tests pass, including a fixture with equal error rates but opposite predictions. The original 09-11 analysis is preserved; this audit does not retrain controllers or refit rules.

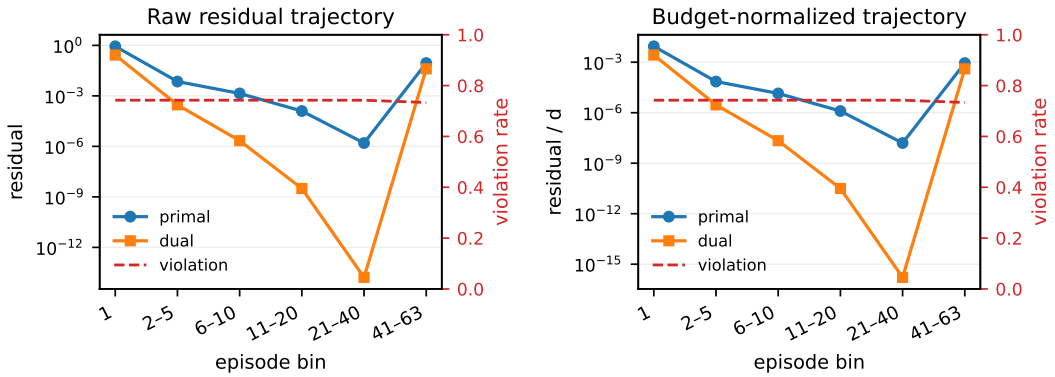

Figure 2: Raw (left) and budget-normalized (right) primal/dual residual trajectories from the same verified $N = 20$ trace. Log scaling exposes the decrease; the red dashed violation rate remains near .74. The last bin contains partial terminal episodes and shows a modest residual rebound without a safety transition.

Table 16: Exact trajectory aggregates underlying Figure 2.

Episodes	raw primal	primal/d	dual/d	violation
1	.8802	8.80×10^{-3}	2.66×10^{-3}	.7425
2–5	.00710	7.10×10^{-5}	2.92×10^{-6}	.7425
6–10	.00140	1.40×10^{-5}	2.18×10^{-8}	.7425
11–20	.000126	1.26×10^{-6}	3.15×10^{-11}	.7425
21–40	.00000158	1.58×10^{-8}	1.65×10^{-16}	.7425
41–63	.0891	8.91×10^{-4}	4.04×10^{-4}	.7335

N GRID RESIDUAL TRAJECTORY DETAILS

The retained grid trace uses $N = 20$, $d = 100$, and ten seeds. Its logged ADMM primal residual is not the simulator quantity $\|q_t - R_d(q_t)\|_2$ used by T1; the implementations expose different optimization interfaces. Agents are averaged within episode, then episodes and seeds within each displayed bin. Both residual plots use exactly the same rows; the right panel divides by d. The violation overlay is unchanged by normalization.

The mean raw primal residual over the full trace is .047, or 4.7×10^{-4} after budget normalization. More importantly, the trajectory falls by over four orders of magnitude through episodes 21–40 while violation stays .7425 and cost/budget stays about 1.282. Because the grid ADMM residual and the simulator radial-normalization residual are not the same quantity, the grid scale cannot justify T1's .05 threshold. We withdraw that threshold argument. The grid trace supports only the qualitative warning that its own agreement metric can decrease while violation persists; T1's numerical discrimination is confined to the planted simulator and requires prospective threshold sensitivity and transport testing.