PAIRED HYBRID POLICY IMPROVEMENT (PHPI): THE GEOMETRY OF DEPLOYMENT EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Paired Hybrid Policy Improvement (PHPI) asks when observations justify replacing a baseline with a frozen learned controller: which feasible worlds still require opposite decisions? Under common real-chart, coverage and attainability conditions, we characterize their separation exponent by valuation ratios and provide exact primal–dual certificates for a restricted positive-monomial class. Source-linked recurrent-control constructions show that timing supports produce exact ambiguity or first-, third- and fifth-order separation, while complementary calibrations can remove collisions at prohibitive finite evidence cost. A locked 48-world study of a 16-state AC/DC model with a frozen GRU tests numerical predictions, revealing processor-dependent rankings, errors, abstentions and matched-method ties. An OPAL-RT RT-SIL study reproduces four complete replacement directions and all 36 sign/status outcomes of a precomputed digital stress map. These exact-source, numerical and RT-SIL results support different levels of evidence; continuous physical deployment remains uncertified.

1 INTRODUCTION

A learned controller may improve its plant yet be difficult to approve from calibration data: the same record can fit another plant where it misses the gain target or violates safety. Paired Hybrid Policy Improvement (PHPI) studies this post-training replacement decision with both controllers frozen. Unlike baseline-relative policy construction (Laroche et al., 2019), the task is to justify a candidate already in hand.

Decision-region determination and task-directed exploration seek information for a downstream decision (Javdani et al., 2014; Wagenmaker et al., 2021). For controller replacement, the key objects are feasible worlds with conflicting complete execution labels. Their records may nearly coincide even when local sensitivity looks favorable: for $F(u) = (u^2, u^3)$, $F(u) - F(-u) = (0, 2u^3)$. The even term cancels; observing it alone leaves an exact collision. Complete identification is sufficient but unnecessary. What matters is excluding opposite decisions within the declared domains.

Contributions.

- We formulate frozen-controller replacement through complete labels and paired observation geometry, keeping true-world and opponent domains explicit.
- We characterize separation under geometric conditions and implement a restricted exact checker. Recurrent-source timing and pulse constructions exhibit ambiguity and matching finite-order separation.
- We test numerical evidence in a locked learned AC/DC study and reproduce its fixed directions and digital-stress categories in RT-SIL, retaining errors, unknowns and strong-method ties.

Figure 1 gives the architecture; Figure 2 develops its geometry. The supplement supplies proofs and executable contracts.

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.

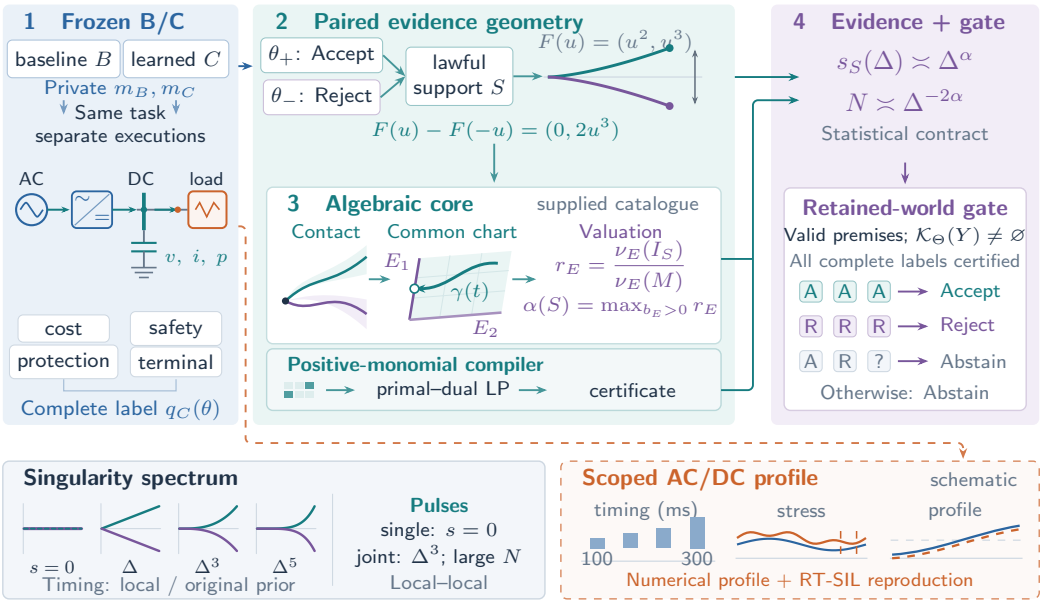


Figure 1: **From complete executions to replacement evidence.** Frozen controllers define labels; lawful observations retain possible worlds. Theorem 1 links paired differences to separation orders. The gate requires valid premises, nonempty compatibility and certified agreement of all retained labels. The positive-monomial checker is restricted; the dashed AC/DC branch uses a numerical profile. Plant, geometry and signal motifs are schematic. Figures 2 and 7 develop the geometry and report RT-SIL outcomes, respectively.

2 Complete Replacement and Paired Separation

Let $\theta \in \Theta$ be a physical world and ω an exogenous task. Frozen controllers B, C execute separately under the same task, each with its own state, memory and events. For a fixed panel or a declared task law, define

$$H_\tau(\theta) = \mathbb{E}_\omega[(1-\tau)J_B(\theta,\omega) - J_C(\theta,\omega)]\,. \tag{1}$$

AC/DC uses $\tau = .05$ and 28 fixed tasks; the two-task mechanical source writes $H_{20} = 19\bar{J}_B - 20\bar{J}_C$. Safety is required on each complete execution, so mean gain cannot compensate for a safety failure. The complete label q_C retains Accept, Reject-Gain or Reject-Safety, with baseline-unsafe, invalid and unresolved cases kept separate (Appendix A).

A safe support S gives ideal observations $F_S(\theta)$. For a covered reading set Y, compatible worlds are $\mathcal{K}_\Theta(Y) = \{\theta \in \Theta : F_S(\theta) \in Y\}$. A definite direction requires nonempty compatibility and one common certified label, including a common rejection reason. A feasible point alone establishes neither actual plant membership nor this agreement. For a true-world power class $V \subseteq \Theta$ and fixed-normalization directional margin ρ_C, paired separation is

$$s_{S,V,\Theta}(\Delta) = \inf_{\substack{\theta\in V,\ \rho_C(\theta)\geq\Delta \\ \vartheta\in\Theta,\ q_C(\vartheta)\neq q_C(\theta)}} \|F_S(\theta) - F_S(\vartheta)\|\,. \tag{2}$$

The source-valid region U supports the label proof, V specifies where nonabstention power is required, and Θ contains the opponents. A positive-margin exact collision gives $s = 0$; an unproved lower bound is a separate outcome. Appendix A.5 fixes boundary and norm conventions.

3 Conditional Paired Geometry

The key object is support-conditioned separation between opposite-decision worlds. Figure 2 gives the picture; Theorem 1 gives the rate. Coordinates reveal vanishing orders without adding observations or removing exact collisions.

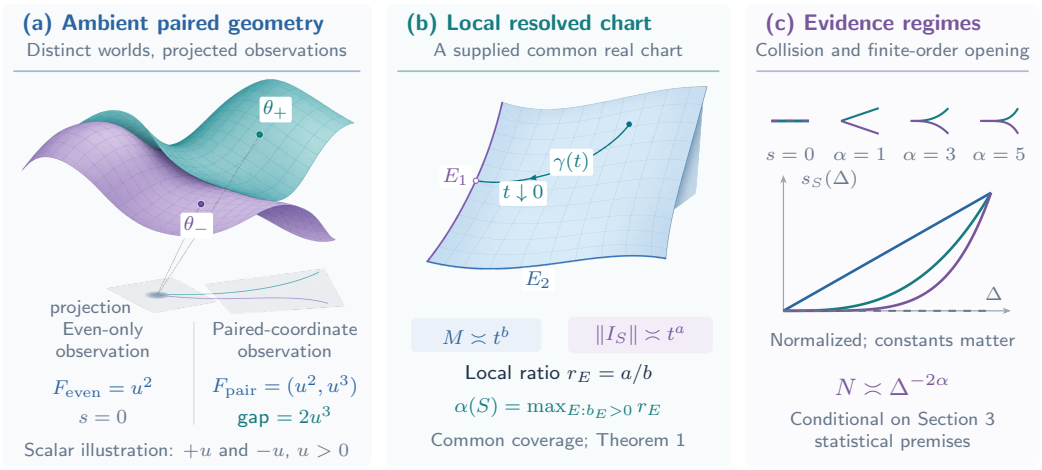


Figure 2: **Paired cancellation and vanishing orders.** (a) Opposite-label regions project to observations; the separate scalar example has gap $2u^3$. The sheets are illustrative, not a manifold assumption. (b) A feasible real arc approaches a divisor in a supplied chart; t is an approach parameter, not time. (c) Collision and finite-order opening lead to different evidence regimes under Theorem 1 and the statistical conditions below.

On the conflict relation, let M be the first world's operational margin. The ideal I_k is generated by the coordinates of $F_k(\theta) - F_k(\vartheta)$, and $I_S = \sum_{k \in S} I_k$ combines probes; its norm uses that generating vector. A common chart expresses these quantities with monomial orders. A divisor E is a boundary component and ν_E its vanishing order. The chart catalogue is supplied, with real approach arcs and coverage as specified next.

Theorem 1 (Paired separation on a common accessible chart catalogue)**.** *Suppose a finite family of compact real analytic subcharts covers every germ on the declared pair relation where M and I_S approach zero. The feasible real approach set is rectilinearized there; in each chart $M \circ \pi_\ell = u_\ell x^{b_\ell}$ and $\|I_S \circ \pi_\ell\| \asymp x^{a_{\ell,S}}$, with analytic units uniformly bounded above and away from zero. Every extremal coordinate divisor used below has a feasible transverse real arc, zero-margin-order coordinates do not independently carry positive observation order in this regime, and the compact remainder has positive separation. Assume also that no positive-margin exact collision occurs. Then, for all sufficiently small $\Delta > 0$,*

$$s_S(\Delta) \asymp \Delta^{\alpha(S)}, \qquad \alpha(S) = \max_{\ell, j: b_{\ell,j} > 0} \frac{a_{\ell,S,j}}{b_{\ell,j}}. \tag{3}$$

Bounded units reduce the lower bound to $x^a \geq (x^b)^\alpha$. A feasible transverse arc at a maximizing divisor attains the upper order; compact separation controls the remainder. Appendix B proves both bounds and states the stronger common-support catalogue formula and its weighted set-cover consequence (Chvátal, 1979). The resolution-based exponent machinery is classical (Bivià-Ausina & Encinas, 2009); real attainability and pair coverage are essential here.

An exact computable subclass. For $M = x^b$ and monomial observations $F_k = x^{a_k}$ with strictly positive exponent entries on $[0, 1]^d$, comparison with the zero-margin set gives

$$\alpha(S) = \max_{v \geq 0,\ b \cdot v = 1} \min_{k \in S} a_k \cdot v. \tag{4}$$

A rational feasible direction and dual mixture certify the optimum exactly (Appendix F). The checker verifies 205 supports, including $M = xy$, $f_1 = xy^3$, $f_2 = x^3 y$, $f_3 = x^2 y^2$. The first two each have exponent three but jointly have exponent two: a common direction must constrain both observations, and their separate worst points miss the diagonal ray.

From separation to evidence. With known fixed positive-definite Gaussian covariance, independent episodes, valid mean coverage and a complete compatibility oracle, separation $\asymp \Delta^\alpha$ gives

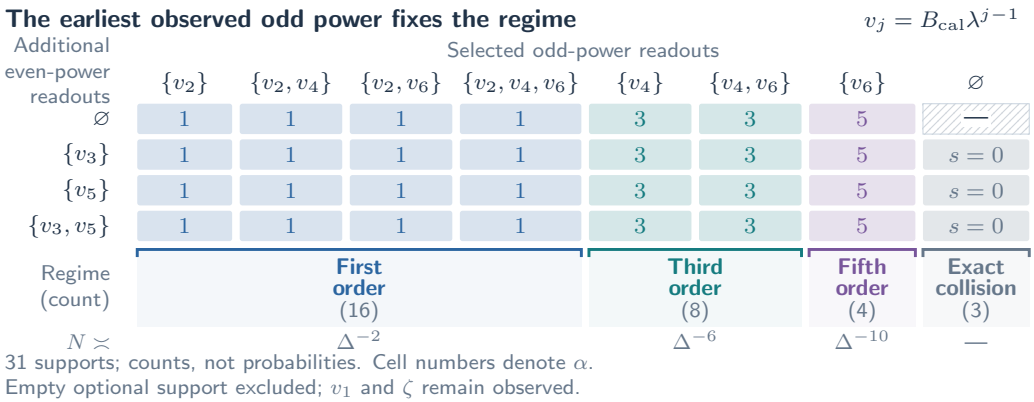


Figure 3: **Timing support fixes the evidence regime.** The 31 nonempty optional supports split 16/8/4/3 across first-/third-/fifth-order and exact collision. Rows add even-power readouts; counts enumerate supports. Theorem 2 compares $V \Subset U$ with the full prior Θ. Episode orders require Section 3's statistical conditions; equal exponents need not give equal finite budgets.

a sufficient $O(\Delta^{-2\alpha})$ episode exponent at fixed error and abstention bounds δ, β. A feasible Accept/Reject pair with $O(\Delta^{2\alpha})$ KL for every legal probe gives the matching adaptive lower exponent when $2\delta + \beta < 1$, by the stopped-transcript chain rule (Cover & Thomas, 2006). Finite constants and oracle completeness matter; Appendix C retains the full bias, covariance and repeated-look contracts, including time-uniform coverage (Howard et al., 2021).

4 Exact source-linked mechanisms

Safe velocity-feedback calibration: timing changes evidence order. This construction compares local true worlds V with all opponents in the original prior Θ. The mechanical source is $p_{t+1} = p_t + v_t/2$, $v_{t+1} = av_t + bu_t/2$, with $a \in [1/2, 1]$, $b \in [1/2, 2]$ and offset $\zeta \in [-1/32, 1/32]$. Appendix A specifies the two six-step reset tasks, clipped baseline, three frozen recurrent candidates and complete costs. A separate prospective calibration applies $u_t^{\text{cal}} = \text{clip}_{[-1,1]}(-\kappa v_t + A\mathbf{1}_{t=0})$, with $\kappa \in \{3/2, 7/4, 2\}$ and $A \in \{1/32, 1/16\}$, giving $v_j = B_{\text{cal}}\lambda^{j-1}$, $B_{\text{cal}} = Ab/2$ and $\lambda = a - \kappa b/2$. Appendix D certifies strict no-clip, speed, position and protection bounds throughout the prior; these do not cover physical reset or noisy feedback.

For each κ, a source-certified zero-pole neighborhood has $H_{20} > 0$ for all three candidates. Its sole active constraint is the second-step slack $S_{\text{safe}}(a, b) = 3/4 - b(a + 1)/2$: the complete label is Accept at $S_{\text{safe}} \geq 0$ and Reject-Safety otherwise. The first two saturated actions determine this boundary; later recurrent execution remains in the cost and safety proof. This is not a naturally memory-induced high-order zero, and Δ is safety slack, not H_{20}. Let $V \Subset U$ contain the center. Supports always include $v_1 = B_{\text{cal}}$ and legal offset ζ, plus v_j for $j \in J \subseteq \{2, \ldots, 6\}$; J denotes optional times and B_{cal} is distinct from baseline B.

Theorem 2 (Timing-support separation)**.** *Let r_* be the least odd exponent among the observed $j - 1$. If such an exponent exists, then, for true worlds in V and oppositely labelled alternatives throughout the original prior Θ, $s_{J,V,\Theta}(\Delta) \asymp \Delta^{r_*}$. If no odd exponent is observed, actual opposite-label worlds have identical observations arbitrarily near the boundary.*

The odd root of v_j/B_{cal} gives a globally defined inverse with $1/r_*$ Hölder modulus. A positive gap separates V from opponents outside U; the pair $(a, b, \zeta) = (a_\kappa \pm u, b_\kappa, 0)$ attains the upper rate. Full inverse constants and complete-source certification are in Appendix D. Figure 3 partitions all 31 nonempty timing supports into 16 first-, 8 third-, 4 fifth-order, and 3 exact-collision supports. A method allowed to read v_2 is not subject to the fifth-order support's lower bound.

Complementary pulses: the alternative domain matters. A common pilot and two fixed dyadic pulses after separate resets give a second mechanism. Both worlds now lie in the certified three-

dimensional U_{pulse}: each pulse alone has a feasible opposite-label collision, while their joint record excludes every conflicting pair in that box. The terminal-velocity difference is $\eta h(b,t)$ with $\partial_b h = A(t)(b-\beta(t))^2 + R(t)$, $A(t) \geq A_* > 0$ and $R(t) \geq 0$. Integrating the square and controlling transverse pilot variation yields $s_{12,U,U}(\Delta) \geq c\Delta^3$; a feasible two-sided arc matches this rate. Appendix E supplies the arbitrary-pair bound, roots and source certificates. Unlike timing support, this comparison is local–local.

Adding the legal offset coordinate gives a local linear inverse and exponent one; whether it can be obtained in the same episode is a separate acquisition question. A known sufficiently small positive change to one pulse's final input also restores eventual exponent one, but with constants deteriorating as the change vanishes. An unknown small input error instead admits an exact same-reading, opposite-label source pair. The cubic mechanism is therefore informative but not a recommended efficient calibration: at one retained margin the hypothetical two-simple-world Gaussian reference requires roughly 1.98×10^{24} repeats, versus about 726 for offset. These are reference budgets, not executed episodes.

5 From structure to an evidence interface

Offline specification. Freeze B, C, the complete target, domains, lawful menu and backend. Mechanical backends use certified parameter and margin bounds; the AC/DC backend uses continuous numerical triangles with separate empirical observation and target buffers (Appendices E.7, F.4 and H).

Online decision. Retain every world compatible with permitted readings and declared uncertainty. Return a direction only if the nonempty set has one certified complete label; otherwise abstain with a reason. True parameters and source labels are for scoring only. Appendix A.4 gives the five-step routine; numerical gain proposals are reported separately.

For one fixed safe support with valid separation lower bound $L_S(\Delta)$, Gaussian mean radius a_N, and a uniformly proved deterministic observation-model discrepancy η_S, the sufficient diameter test is $2(a_N + \eta_S) < L_S(\Delta)$. For $a_N = 4\sigma_S/\lfloor\sqrt{N}\rfloor$ its least sufficient square budget is

$$N_S = \left(\left\lfloor \frac{8\sigma_S}{L_S(\Delta) - 2\eta_S} \right\rfloor + 1 \right)^2, \quad L_S(\Delta) > 2\eta_S. \tag{5}$$

Equation (5) applies to one support and lower certificate. Declared setup/episode prices permit a separate finite-menu comparison, not unrestricted sample-optimal design. Failure to clear $2\eta_S$ means no *certified* gap, not no physical gap. Fixed bias, covariance, adaptive selection and multiple looks retain their own contracts (Appendix C).

6 Learned AC/DC control: a locked numerical study

Figure 4 follows the paired replacement question into learned AC/DC control: timing and approximation determine which observations are useful, while complete source executions score the numerical proposals. This case tests finite-scale evidence rather than transferring the exact separation exponents.

Plant and frozen candidate. The 16-state averaged model runs 20 μs RK4 with a 2 ms control clock. A frozen eight-state GRU adds residual outputs over baseline $B1$, which has secondary-control memory. Uncertainty concerns $T_{\mathrm{DC},2}$ and C_2. The final checkpoint comes from hand-written-teacher action imitation, independently of confirmation; Appendix G retains architecture and training. The 28 tasks were nominal-baseline-qualified before this study. Complete costs and per-task safety define

$$H(\theta) = \frac{1}{28} \sum_{i=1}^{28} [0.95 J_i^{B1}(\theta) - J_i^{C}(\theta)]. \tag{6}$$

At $T_{\mathrm{DC},2} = 24/56$ ms, complete source evaluations are safe and reduce raw mean cost by 5.2758% and 2.7984%, respectively. Only the first clears the 5% target: REJECT-GAIN means insufficient improvement, not necessarily higher raw cost. Separately, an intervention after an unchanged 60 ms

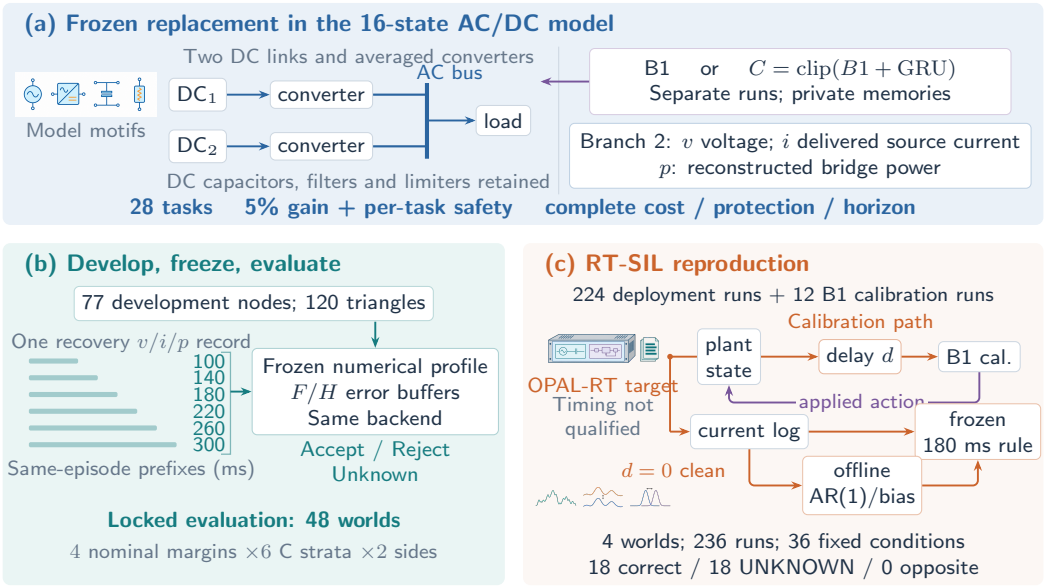


Figure 4: **One AC/DC case across Python and RT-SIL.** (a) Frozen B1/GRU alternatives share the task and complete replacement target. (b) A frozen numerical profile is evaluated on a locked 48-world panel. (c) The target runs plant and controllers internally: 224 deployment and 12 B1 calibration runs. Clean delay-zero logs pass through offline AR(1)/bias to the frozen 180 ms rule; current-time delay-condition logs use that same rule. All component and signal illustrations are schematic.

prefix changes the 24 ms *12-task* gain from 6.7214% to 4.6039%. This development diagnostic establishes later memory relevance on those tasks, not the *28-task* confirmation result or the mechanical early-saturation mechanism. Neither establishes a natural high-order singularity in this plant.

Locked panel and forecasts. The profile uses 77 development vertices and 120 continuous triangles on $C_2 \in [0.0479, 0.0481]$, with $T_{\mathrm{DC},2}$ within 10^{-4} s of the interpolated boundary. Observation and target errors have separate empirical buffers, not uniform source bounds; true parameters and labels serve only independent scoring (Appendix H.1).

Quiet/short/recovery last 300/100/300 ms. Records use DC voltage, delivered current and reconstructed converter power with Gaussian scales $\sigma_0 = (10^{-5}, 10^{-4}, 10^{-4})$ in source units, not measured hardware or feedback noise. Software-full adds 16 simulated state coordinates from the same recovery episode; these privileged readouts are not validated sensors. Full-record methods may ignore extras and use any subrecord.

The lock fixes 48 worlds (four nominal margins × six capacitance strata × two signs), eight synthetic histories per world, four noise scales and eight looks ending at $N = 128$. All panels were source-safe, with 24 ACCEPT and 24 REJECT-GAIN labels. The 384 histories per condition reuse 48 calibration means: N counts readout repeats, not extra trajectories or plants. Appendices G–H retain reconstruction, run accounting and buffer values.

At original noise, the development occupancy ordering carries over to the new panel: the 180 ms rule gives 380/384 correct endpoint recommendations, versus 288/384 for 300 ms; the respective old occupancies were 32/32 and 24/32. At $1000\sigma_0$, the new endpoint counts are instead 167 and 194. The ranking thus depends on the noise and processing contract. Occupancy of 32 exposed noiseless means is a development diagnostic, not the same probability object as a noisy recommendation rate on the new panel.

Errors and processor comparisons. Fixed endpoints and interim decisions differ. At original noise, quiet has three wrong $N = 128$ directions and the 100 ms recovery processor has one. At $1000\sigma_0$, recovery 180 ms has two wrong directions at $N = 2$ and one at $N = 8$, despite none at $N = 128$; its

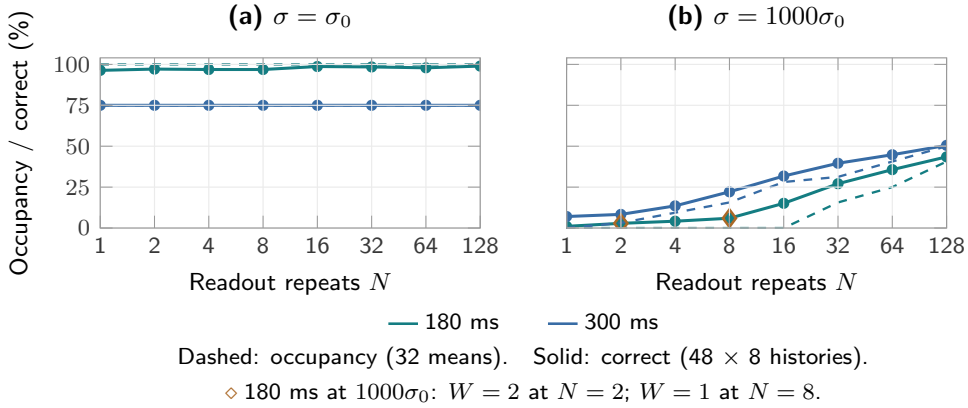


Figure 5: **Development predictions and new-panel outcomes.** Dashed curves show occupancy of 32 old means; solid curves score 384 histories from 48 new worlds. Diamonds retain the wrong high-noise 180 ms outputs (2 at $N = 2$, 1 at $N = 8$). Occupancy and recommendation rates measure different quantities. Appendix H includes all scales, errors and unknowns.

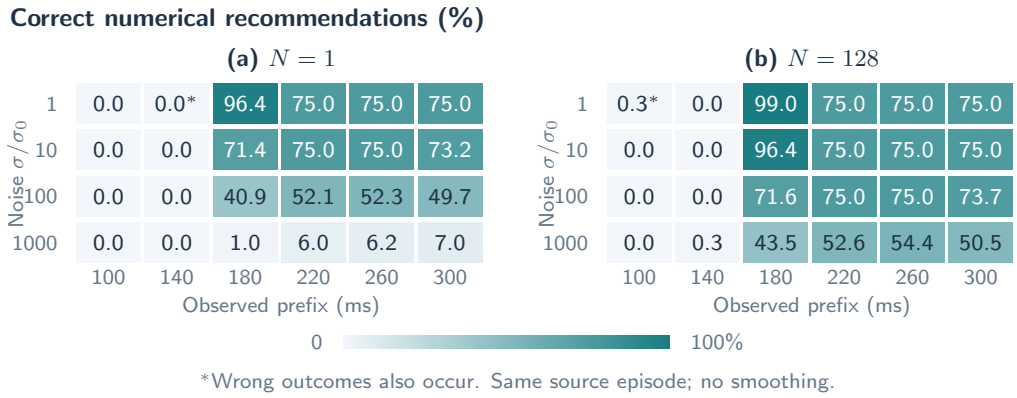


Figure 6: **Timing and readout budget change numerical usefulness.** Correct percentages among 384 histories, with stars marking wrong outputs, are shown at four noise scales without smoothing. All prefixes use one 300 ms episode; full records can apply any prefix rule. Appendix H.3–H.4 retains the fixed-budget dashboard and every correct/wrong/unknown count.

first-look audit reports three wrong first directions and 29 withdrawals. These overlapping summaries are not new trials. Endpoint zero error establishes neither error-free monitoring nor anytime-valid stopping (Appendix H.4).

Figure 6 separates readout and processor limitations: at original noise the 300 ms rule stays at 288/384 correct from $N = 1$ to 128; at $1000\sigma_0$ it rises from 27 to 194. Appendix H.3–H.4 retains the fixed-budget dashboard, all counts and residuals against both the quadratic design proxy and the piecewise-linear inference target.

The task-c, parameter-E, direct-pair, PHPI-pair and engineering-standard selectors all chose 300 ms under the same data, backend and stopping rule and tied at 288/384 at original noise. Random selection gave 81/1/302 correct/wrong/unknown. The frozen software-full processor also gave 288/384, but its raw record may run the 180 ms rule and match 380/384. Thus neither the strong-method tie nor the fixed-processor advantage of 180 ms is a selection-algorithm victory or information-dominance result.

Appendices G–I retain the task, warm-prefix, nuisance and residual analyses. RT-SIL extends this same case to a second execution platform.

6.1 REAL-TIME TARGET REPRODUCTION UNDER CONTROLLED DIGITAL STRESS

After developing the stress panel, we froze four medium/high-margin worlds, the 28-task target, weights, 5% threshold, 180 ms profile, buffers and predictions. The OPAL-RT target ran the same plant and controllers internally: 224 deployment runs and twelve B1 recovery calibrations, all complete and safe without trips. The four deployment directions matched Python (Appendix J).

AR(1)/bias perturb acquired traces offline. Model-side delay ($0/20/50\,\mu$s) affects B1 calibration observations while the logger retains current-time outputs (Figure 4c). The frozen $N = 16$ inputs are digital replay/averaging, not sixteen independent target acquisitions. Figure 7 reports all 36 conditions; the strongest tested stress in each family yields only UNKNOWN. Appendix J.1–J.3 retains levels, trace reuse and the full accounting.

(a) Python prediction: frozen before target execution

	AR(1) correlation ρ			Fixed bias / σ_0			Calibration-path delay (μs)		
	0.5	0.9	0.99	0.11	0.12	0.13	0	20	50
w10p	A	U_M	U_M	A	U_M	U_M	A	A	U_M
w10n	R	U_M	U_M	R	U_M	U_M	R	R	U_M
w33p	A	A	U_O	A	A	U_M	A	A	U_M
w33n	R	U_M	U_M	R	U_M	U_M	R	R	U_M

(b) OPAL-RT RT-SIL: retained target-trace outcomes

	AR(1) correlation ρ			Fixed bias / σ_0			Calibration-path delay (μs)		
	0.5	0.9	0.99	0.11	0.12	0.13	0	20	50
w10p	A	U_M	U_M	A	U_M	U_M	A	A	U_M
w10n	R	U_M	U_M	R	U_M	U_M	R	R	U_M
w33p	A	A	U_O	A	A	U_M	A	A	U_M
w33n	R	U_M	U_M	R	U_M	U_M	R	R	U_M

A: positive definite R: negative definite U_M, U_O: UNKNOWN

U_M: no compatible numerical cell; U_O: orthogonal model mismatch.

Figure 7: **A frozen stress map transfers to the real-time target.** All 36 sign/status pairs match: 18 correct definite directions, 18 UNKNOWN and no opposite direction. A/R denote numerical gain proposals; U_M/U_O distinguish no compatible numerical cell from orthogonal model mismatch. The two panels compare pre-target Python predictions with retained target-trace outcomes.

Evidence scope. These are finite numerical and fixed-panel RT-SIL results. The profile's empirical buffers leave uniform source error, continuous 28-task/safety coverage, model membership and NumPy/IEEE/ODE transfer open (Appendices G–I). Physical I/O and an independent external controller (CHIL) were not tested; digital stress is not measured sensor noise or communication latency. Controller computations were below 9 μs against a 2 ms budget, with zero native overruns, but 143 subsequent intervals exceeded 2.1 ms in two B1 cases. Their cause is unresolved and scheduler timing remains `NOT_QUALIFIED`. The fixed stress map establishes neither a critical threshold nor population reliability; Appendix J retains timing anomalies and unavailable-data boundaries.

7 RELATED WORK

Policy improvement and experimental design. Baseline bootstrapping (Laroche et al., 2019), local Monte Carlo search (Castellini et al., 2023) and factored multi-agent methods (Bianchi et al., 2024) construct policies relative to a baseline. PHPI fixes the candidate and asks which observations exclude conflicting replacement labels. Decision-region acquisition (Javdani et al., 2014), task-optimal exploration (Wagenmaker et al., 2021) and learned Bayesian design (Foster et al., 2021; Blau et al., 2022) optimize information for a task. Our finite support menu instead changes paired

cancellation; nonregular information (Lin et al., 2019), resolution exponents (Bivià-Ausina & Encinas, 2009) and set cover (Chvátal, 1979) supply classical ingredients.

Statistical and verification guarantees. Confidence sequences (Howard et al., 2021), sequential multiple testing (Malek et al., 2017) and learned betting tests (Pandeva et al., 2024) address valid stopping. Geometric power additionally depends on compatible opposite labels. Runtime validation (Mitsch & Platzer, 2016) and neural closed-loop reachability (Huang et al., 2022) address assurance obligations left open here. These references provide context, not additional experimental baselines.

8 Discussion and conclusion

Complete executions define replacement; paired geometry determines which conflicting worlds the data can exclude. Under the stated conditions, relative vanishing orders explain separation, while the exact constructions show how support choice changes the order or preserves collision. The AC/DC study exposes finite-scale effects of processing and approximation, including strong-method ties. RT-SIL reproduces a frozen numerical stress map across platforms. Together, these results connect a replacement target to the evidence needed for a decision, with guarantees set by the source, statistical and execution contracts.

Reproducibility statement

Appendices A–K contain proofs, contracts, full counts and retained-data provenance. The anonymous Python supplement supplies exact examples and frozen numerical outputs. This manuscript's `build.py` rebuilds figures and both PDFs without scientific reruns; Appendix K identifies entries and unavailable target data.

AI use statement

AI tools assisted with literature search and review, manuscript drafting, and language editing. They also contributed to writing portions of the program code, as well as organizing the code and improving its readability. For the mathematics, AI assistance was limited to constructing the counterexamples in Appendix E and checking the proofs of the remaining mathematical theory. The authors remain responsible for the accuracy of all arguments, results, and submitted content.

REFERENCES

Federico Bianchi, Edoardo Zorzi, Alberto Castellini, Thiago D. Simão, Matthijs T. J. Spaan, and Alessandro Farinelli. Scalable safe policy improvement for factored multi-agent MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3952–3973. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24b.html.

Carles Bivià-Ausina and Santiago Encinas. Łojasiewicz exponents and resolution of singularities. *Archiv der Mathematik*, 93:225–234, 2009. doi: 10.1007/s00013-009-0032-5.

Tom Blau, Edwin V. Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2107–2128. PMLR, 2022. URL https://proceedings.mlr.press/v162/blau22a.html.

Alberto Castellini, Federico Bianchi, Edoardo Zorzi, Thiago D. Simão, Alessandro Farinelli, and Matthijs T. J. Spaan. Scalable safe policy improvement via Monte Carlo tree search. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3732–3756. PMLR, 2023. URL https://proceedings.mlr.press/v202/castellini23a.html.

Vašek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395. PMLR, 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Chao Huang, Jiameng Fan, Xin Chen, Wenchao Li, and Qi Zhu. POLAR: A polynomial arithmetic framework for verifying neural-network controlled systems. In *Automated Technology for Verification and Analysis*, volume 13505 of *Lecture Notes in Computer Science*, pp. 414–430. Springer, 2022. doi: 10.1007/978-3-031-19992-9_27.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, Drew Bagnell, and Siddhartha Srinivasa. Near optimal bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438, 2014. URL https://proceedings.mlr.press/v33/javdani14.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Yi Lin, Ryan Martin, and Min Yang. On optimal designs for nonregular models. *The Annals of Statistics*, 47(6):3335–3359, 2019. doi: 10.1214/18-AOS1780.

Alan Malek, Sumeet Katariya, Yinlam Chow, and Mohammad Ghavamzadeh. Sequential multiple hypothesis testing with Type I error control. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1468–1476. PMLR, 2017. URL https://proceedings.mlr.press/v54/malek17a.html.

Stefan Mitsch and André Platzer. ModelPlex: Verified runtime validation of verified cyber-physical system models. *Formal Methods in System Design*, 49:33–74, 2016. doi: 10.1007/s10703-016-0241-z.

Teodora Pandeva, Patrick Forré, Aaditya Ramdas, and Shubhanshu Shekhar. Deep anytime-valid hypothesis testing. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 622–630. PMLR, 2024. URL https://proceedings.mlr.press/v238/pandeva24a.html.

Andrew J. Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10641–10652, 2021. URL https://proceedings.mlr.press/v139/wagenmaker21a.html.