SAMESHOCK: AUDITING ADAPTIVE STRESS TESTS AT MATCHED INTERVENTION EXPOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

An adaptive stress test can increase a controller's failure rate by changing when a fixed disturbance occurs. A valid comparison must distinguish harmful state-dependent timing from changes in the distribution of intervention times and in how often an intervention reaches an active system. We formulate this comparison through the joint law of scheduled time and pre-intervention activity, and develop a finite-sample audit for a frozen library of causal stressors. Untouched nominal trajectories estimate every candidate's intervention law by integrating its stopping probabilities; separate intervention rollouts estimate failure risk. A small linear program selects an episode-level mixture that certifies both a prescribed total-variation tolerance and a lower bound on excess failure probability after mixture selection. On 80 FrozenLake controllers, matching time alone permits an exact risk-maximizing tester to increase active exposure by 8.95 percentage points. Controlled audits separate savings from path sharing, modest gains from stopping integration, and gains from a sharper reference bound. On ten proximal policy optimization controllers in a continuous-state CartPole test, SameShock exposes a mean 2.40 percentage-point failure increase at a 1% joint-law tolerance, using a strict matched transition budget. At a lower impulse strength, independent-shock failure falls to 57.15% and the audited failure increment remains 2.28 percentage points.

1 INTRODUCTION

An inventory controller can absorb a demand shock at high stock yet fail under the same shock at low stock. Likewise, a balancing controller can recover from an impulse in one state and fail under it in another. Independent scheduling estimates average vulnerability; an adaptive tester observes the state and chooses a dangerous moment. Timing attacks and failure search exploit this distinction (Lin et al., 2017; Lee et al., 2020; Uesato et al., 2019), within the broader study of simulation-based safety validation (Corso et al., 2021). The unresolved evaluation question is how to compare such testers under a declared, statistically verified intervention distribution.

Fixing a single shock's magnitude and direction isolates timing as the tester's decision, but does not fix intervention exposure. Even with the same scheduled-time distribution, the fraction of shocks that reach an active system can change. Conditioning failure rates on survival until the shock also changes the population in the denominator. We retain every initial episode, including absorbed episodes, and match the joint distribution of scheduled time and pre-shock activity.

With a known finite-state model, this target is a specialization of distribution-constrained optimal stopping (Bayraktar & Miller, 2019; Beiglböck et al., 2018). Its occupancy linear program (LP) provides an exact reference. In execution, however, raw fitted and prefix-calibrated testers show substantial joint-law violations. Independent auditing establishes which risk increases satisfy the intervention contract.

SameShock freezes the candidate library before drawing audit data, separating candidate search from the evidence used for certification. The one-intervention structure yields a shared-path identity: before the shock, every candidate follows the same nominal dynamics. One untouched nominal trajectory can therefore estimate the entire intervention law of every frozen candidate, integrating out its stopping coins. Failure risks are estimated using separate complete intervention rollouts.

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.

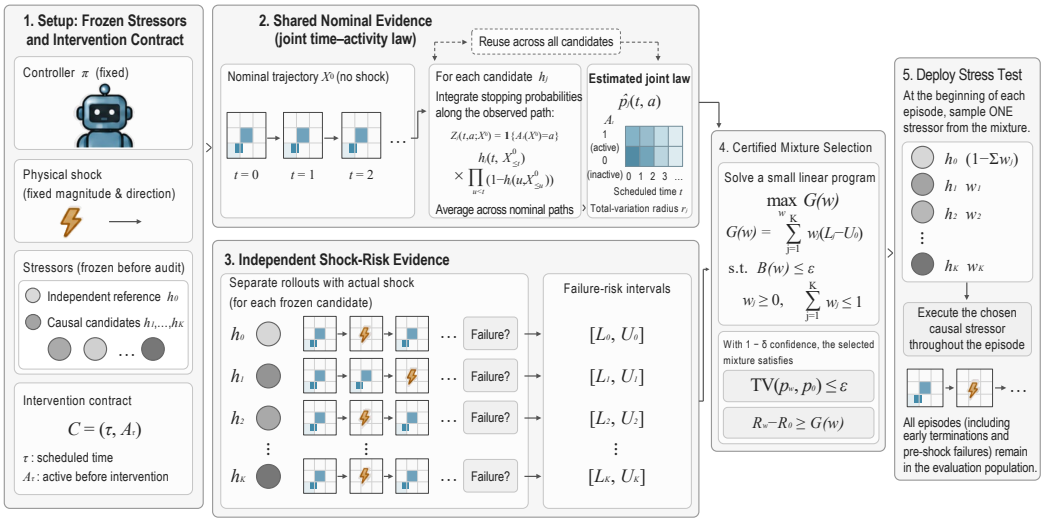

Figure 1: SameShock overview. (1) Fix the controller and physical shock, and freeze the stressor library $h_0, \ldots, h_K$. The intervention contract $C = (\tau, A_\tau)$ records scheduled time and pre-shock activity. (2) Shared nominal paths yield stopping-mass vectors Z_j; averaging gives joint-law estimates $\widehat{p}_j$ with total-variation (TV) radii r_j. (3) Separate intervention rollouts give failure-risk intervals $[L_j, U_j]$. (4) Simultaneous law and risk bounds define a linear program maximizing the failure-risk-increment lower bound $G(w)$ subject to $B(w) \le \epsilon$. The certificate holds after selection with confidence $1 - \delta$. (5) Draw one causal stressor per episode and retain all initial episodes. Complete nominal paths are used offline; deployed stressors use only past and current states. Trajectories and law cells are schematic.

Simultaneous confidence bounds then define a convex feasible set of episode mixtures. The selected tester maximizes a certified failure-risk increment within this set (Figure 1).

The contribution has three parts. First, the joint time/activity contract makes intervention exposure comparable while leaving state-dependent timing available for failure discovery. Exact stopping comparisons isolate the physical exposure change that time matching permits. Second, the shared-path identity estimates every frozen candidate law from common nominal trajectories; controlled audits quantify the roles of path reuse, stopping integration and the reference-specific confidence bound. Third, simultaneous law and risk bounds yield a selection-valid certificate and an uncertainty-dependent comparison with every feasible library mixture. Finite-state and neural controller experiments establish the resulting failure discovery power.

2 Related Work

Adversarial evaluation. Adaptive stress testing searches for likely failure trajectories (Lee et al., 2020; Koren et al., 2018), while falsification searches for specification violations (Dreossi et al., 2019; Yamagata et al., 2021). Adversarial evaluation targets difficult initial conditions (Uesato et al., 2019), attack times (Lin et al., 2017), or opponent policies (Gleave et al., 2020). Robust training addresses disturbance forces (Pinto et al., 2017), action uncertainty (Tessler et al., 2019), observation attacks (Zhang et al., 2020; 2021; Nakanishi et al., 2025), and temporally coupled perturbations (Liang et al., 2024). In large language models, SpatialJB (Mou et al., 2026b) evaluates layout-based jailbreaks, while RolePlay (Mou et al., 2026a) targets inference cost through persona conditioning. SameShock complements candidate search with a statistical audit: it certifies exposure matching and failure-risk gain for frozen causal timing rules under a fixed physical shock.

Distribution constraints. Distribution-constrained stopping fixes the stopping-time law (Bayraktar & Miller, 2019; Beiglböck et al., 2018); its finite-state occupancy specialization supplies our

exact reference. Robust decision processes address transition uncertainty (Iyengar, 2005; Nilim & El Ghaoui, 2005), while constrained policy optimization targets return and cost constraints (Achiam et al., 2017; Polosky et al., 2022). Our constrained decision is the tester's timing, with the controller held fixed.

Statistical selection. High-confidence policy improvement (Thomas et al., 2015a; Laroche et al., 2019) and Seldonian learning (Thomas et al., 2019) separate performance optimization from statistical constraint certification. Risk-controlling prediction sets (Bates et al., 2021) and Learn then Test (Angelopoulos et al., 2025) calibrate frozen models with finite-sample guarantees; conformal risk control (Angelopoulos et al., 2024) controls expected monotone loss. These works establish the selection principle. SameShock applies simultaneous law and risk bounds to certify every episode mixture, including the one chosen using the audit.

3 The Intervention Contract

Fix a controller π, an initial-state distribution, a horizon H, and a physical shock operator of declared magnitude and direction. Let $\mathcal{T} \subseteq \{0, \dots, H-1\}$ be the allowed schedule slots and let q_t be a probability law on them. The stressor observes the current state and past history. It schedules one shock at a stopping time $\tau \in \mathcal{T}$; it has no access to future environment randomness. The controller acts before the current shock decision takes physical effect. Its subsequent observations and actions follow the specified controller interface.

Let $A_t \in \{0, 1\}$ indicate that the system is active immediately before the step-t intervention. Both successful and failed terminal states are inactive. Inactivity is absorbing in our benchmarks. Terminal episodes remain in the accounting population, with their nominal state held absorbing, so every episode receives a scheduled time even if the shock has no physical effect. Let $Y \in \{0, 1\}$ denote failure by H, including failures before τ. Define $R_h = \mathbb{E}_h Y$ and

$$C = (\tau, A_\tau), \qquad p_h(t, a) = \mathbb{P}_h(C = (t, a)), \qquad \mathrm{TV}(p, p') = \tfrac{1}{2} \sum_c |p(c) - p'(c)|. \tag{1}$$

The reference h_0 samples $\tau \sim q$ independently of the nominal path. Writing $\mathbb{P}_{\mathrm{nom}}$ for the untouched pre-shock process gives $p_0(t, a) = q_t \mathbb{P}_{\mathrm{nom}}(A_t = a)$. Our exact target is $p_h = p_0$. The sample-based target is explicitly $\mathrm{TV}(p_h, p_0) \le \epsilon$ with confidence $1 - \delta$. Marginalization implies both $\mathrm{TV}(\mathcal{L}_h(\tau), q) \le \epsilon$ and $|\mathbb{P}_h(A_\tau = 1) - \mathbb{P}_0(A_\tau = 1)| \le \epsilon$.

Joint matching leaves the dependence between timing and more detailed physical state free. It controls exposure frequency, not which vulnerable states are chosen within each time/activity cell. This is the intended stress dimension. Matching time alone does not achieve the same control: a tester can move active interventions from one cell to another while offsetting their scheduled mass with inactive episodes. A constructive example is given in Appendix A.1.

Known-model reference. For a finite Markov state augmented as needed for the controller, let P_t be the pre-shock nominal kernel and $g_t(s)$ the failure probability if the shock occurs now and the controller follows its post-shock branch. Stopped and continuing masses $z_t, c_t \ge 0$ obey

$$z_0 + c_0 = \mu, \quad z_t + c_t = P_{t-1}^{\mathsf{T}} c_{t-1}, \quad \sum_{s : A(s) = a} z_t(s) = p_0(t, a). \tag{2}$$

Maximizing $\sum_{t,s} z_t(s) g_t(s)$ gives the exact joint-constrained reference. Removing the activity split gives the time-only reference. This linear program specializes classical stopping using state-action flows (Altman & Shwartz, 1991). Fitted versions generate candidates; the true model is reserved for evaluation.

4 Independent Auditing with Shared Nominal Paths

Freeze K candidate hazards $h_1, \dots, h_K$ and the reference h_0 before collecting audit data. A hazard is the conditional probability of stopping at an allowed slot given the observed pre-shock history and no previous stop. Every hazard stops at the last slot if it has not stopped earlier. The library, shock, schedule, tolerance, and confidence level are fixed before the audit.

4.1 ESTIMATING ALL INTERVENTION LAWS ON COMMON PATHS

Let X^0 be a complete nominal trajectory of the controller's pre-shock branch. For $a \in \{0, 1\}$ define the path contribution

$$Z_j(t, a; X^0) = \mathbf{1}\{A_t(X^0) = a\}\, h_j(t, X^0_{\le t}) \prod_{u \in \mathcal{T}: u < t} \big(1 - h_j(u, X^0_{\le u})\big). \tag{3}$$

Proposition 1 (Shared-path representation)**.** *Each $Z_j(X^0)$ is a simplex vector on the $m = 2|\mathcal{T}|$ cells and $\mathbb{E}Z_j(X^0) = p_j$. For the independent reference, $Z_0(t, a; X^0) = q_t \mathbf{1}\{A_t(X^0) = a\}$.*

The product integrates all stopping coins conditional on a nominal path. Before the first shock, that path has the same transition law as the intervened process. Thus the offline auditor can use the unused nominal suffix without granting the online stressor future information. Averaging Z_j over n independent nominal trajectories estimates every candidate law. This removes stopping-coin noise and shares environment transitions across the library; it still requires evaluating each candidate's scores.

Write $\delta_e = \delta/[2(K + 1)]$. For any candidate simplex observation, a union of one-sided Hoeffding bounds over cell subsets gives

$$r_j = \min\left\{1, \sqrt{\frac{m \log 2 + \log(1/\delta_e)}{2n}}\right\}, \qquad \mathbb{P}\{\mathrm{TV}(\widehat{p}_j, p_j) > r_j\} \le \delta_e. \tag{4}$$

The bound applies to both integrated simplex contributions and one-hot categorical observations. The union bound permits dependence across candidates; the declared support includes zero-count cells.

The independent reference admits a sharper radius. Writing $S(t) = \mathbb{P}_{\mathrm{nom}}(A_t = 1)$ gives

$$\mathrm{TV}(\widehat{p}_0, p_0) = \sum_t q_t |\widehat{S}(t) - S(t)|.$$

Absorbing inactivity makes S a survival function. The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956; Massart, 1990) therefore gives

$$r_0 = \min\{1, \sqrt{\log(2/\delta_e)/(2n)}\}. \tag{5}$$

For a nonabsorbing activity indicator, the generic bound in Equation 4 remains available. Failure risks require complete intervention rollouts: use separate independent batches to obtain Clopper–Pearson intervals $[L_j, U_j]$ with coverage $1 - \delta_e$. The intersection of the $K + 1$ law events and $K + 1$ risk events has probability at least $1 - \delta$.

4.2 SELECTING A CERTIFIED EPISODE MIXTURE

Let $w_j \ge 0$ for $j \ge 1$, $\sum_j w_j \le 1$, and assign the remainder to the independent reference. Draw this component once per fresh episode and execute it throughout. Selecting a component once preserves its stopping rule, including any dependence on the observed state and history, while keeping the intervention law and failure risk linear in the mixture weights. Define

$$B(w) = \tfrac{1}{2} \left\| \sum_{j=1}^{K} w_j (\widehat{p}_j - \widehat{p}_0) \right\|_1 + \sum_{j=1}^{K} w_j (r_j + r_0), \tag{6}$$

$$G(w) = \sum_{j=1}^{K} w_j (L_j - U_0). \tag{7}$$

SameShock solves

$$\widehat{w} \in \arg \max_{w \ge 0,\, \mathbf{1}^\top w \le 1} G(w) \quad \text{subject to } B(w) \le \epsilon. \tag{8}$$

Absolute-value auxiliary variables make this an LP with $K + m$ variables. The baseline $w = 0$ is always feasible. Unlike summing each candidate's individual TV distance, the norm in Equation 6 retains cancellation between signed distribution deviations.

Theorem 2 (Validity after mixture selection)**.** *Conditional on the frozen controller and library, with probability at least* $1-\delta$ *over independent audit trajectories,*

$$\mathrm{TV}(p_{\widehat{w}}, p_0) \le B(\widehat{w}) \le \epsilon, \qquad R_{\widehat{w}} - R_0 \ge G(\widehat{w}) \ge 0.$$

The bounds hold simultaneously for all episode mixtures on the same audit event, so optimization using the audit does not incur an additional selection penalty.

The proof applies the triangle inequality to the law and simultaneous risk intervals to the objective (Appendix A). Switching or averaging hazards within an episode generally induces a different law; such a rule can be included as an additional candidate if fixed before auditing, with its own law and risk estimates.

Executable procedure. Freeze the controller, hazards and audit configuration; draw nominal paths and independent risk batches; compute law radii and risk intervals; solve Equation 8; then draw a component once at the start of each new episode. Return both $B(\widehat{w})$ and $G(\widehat{w})$ with the weights. Return the baseline when no positive gain is certified. Appendix A.3 explains the audit parameter choices.

Corollary 4 compares the auditor with every library mixture v satisfying $\mathrm{TV}(p_v, p_0) \le \epsilon$, including boundary optima. For $e_v = \sum_j v_j(r_j + r_0)$ and $b_v = \sum_j v_j[(U_j - L_j) + (U_0 - L_0)]$, it gives, when $\epsilon > 0$,

$$R_{\widehat{w}} - R_0 \ge \max\left\{0, \frac{\epsilon}{\epsilon + 2e_v}(R_v - R_0 - b_v)\right\}.$$

Shrinking v toward the independent reference pays for law uncertainty; b_v accounts for risk uncertainty. We evaluate this comparison against the exact same-library optimum in Appendix A.5. At zero tolerance, positive radii force the baseline; Appendix A.7 examines exact matching conditional on the calibration sample.

Certified single-candidate comparator. For each candidate let $d_j = \mathrm{TV}(\widehat{p}_j, \widehat{p}_0) + r_j + r_0$ and $\alpha_j = \min\{1, \epsilon/d_j\}$. Select the candidate with the largest nonnegative $\alpha_j(L_j - U_0)$ and mix it with the reference. This *certified single* uses the same data, bounds, and objective. It isolates the value of mixing multiple candidates. The unrestricted LP cannot have a lower certified objective, but its realized risk gain need not exceed that of the single-candidate solution.

Interpreting a risk increment. Let $m_0(c) = \mathbb{E}_0[Y \mid C = c]$, extended to $[0,1]$ on zero-probability cells. The failure change attributable to cell frequencies alone is $\sum_c (p_w(c) - p_0(c))m_0(c)$, whose magnitude is at most $\mathrm{TV}(p_w, p_0)$. Thus $G(\widehat{w}) > \epsilon$ certifies a positive residual of at least $G(\widehat{w}) - \epsilon$ after removing this frequency contribution. This outcome decomposition does not identify a conditional causal effect.

5 Experiments

We test whether activity matching changes the evaluation contract, whether shared auditing increases valid failure discovery at a common sampling budget, and whether the effect persists across controller capabilities and shock strengths. Confirmation uses training seeds distinct from method development. Component comparisons use fresh audits of these frozen controllers; contract reanalyses use existing audit data. Experimental details and additional results appear in Appendices B–F.

Environments and controllers. Two stylized finite-state tasks represent resource extraction (25 states, stock loss 6) and inventory control (52 states, extra demand 5), each with $H = 32$. FrozenLake uses the canonical slippery 8×8 map and the registered 200-step horizon, with one reversal of the commanded action. For these tasks, empirical transition models use 400 independent generative samples per state-action row for each nominal and shocked kernel. Finite-horizon backups train controllers under independent scheduling with failure penalties $\{0, 2, 8, 32\}$. We retain all four policies for each of 20 independent model seeds per environment. Exact dynamics are withheld from the auditor.

CartPole uses Gymnasium Euler dynamics (Towers et al., 2025) and an explicit 200-step test horizon. Ten new controllers are trained with proximal policy optimization (PPO) (Schulman et al.,

Table 1: Activity control on all 80 FrozenLake-200 controllers. Gain and change in active exposure are percentage points relative to independent timing. Joint total variation (TV) is in percent. Exact references impose equality over the full stopping space. Audited rules use the same frozen library, observations and confidence radii, with a 1% tolerance on their respective target. Gain intervals resample the 20 training seeds.

Contract	Risk gain	Active change	Joint TV
Exact time-only	11.75 [11.67, 11.83]	8.95	12.05
Exact joint	5.77 [5.72, 5.83]	0.00	0.00
Time-only, common bounds	5.03 [4.80, 5.28]	4.26	5.48
Audited joint	1.83 [1.65, 2.02]	0.16	0.59

2017), using Stable-Baselines3 (Raffin et al., 2021), 100,352 environment steps, and a 16-unit actor hidden layer. The shock adds $+2\,\mathrm{rad\,s^{-1}}$ to angular velocity once; its sign and size are fixed before observing any state. The test uses our intervention wrapper and a 200-step cap; the native CartPole-v1 cap is 500 steps. A supplemental $+1.75$ experiment reuses all ten controllers with fresh strength-specific score fitting, prefix calibration, audits and tests. This strength was chosen after an independent-scheduling baseline probe; its full audit design was fixed before observing method gains. The schedule is uniform over $20, 30, \ldots, 170$ for both 200-step tasks, and over steps $4, \ldots, 27$ for the stylized tasks.

Candidates and data separation. The tabular library contains independent scheduling, fitted time-only LP, fitted joint LP, fitted quota-greedy scheduling, and prefix quantile calibration. The stressor model is estimated independently of the controller model. CartPole replaces model-based candidates with four prefix-calibrated scores: quadratic logistic failure prediction, absolute pole angle, signed angular velocity, and absolute action-logit gap. The failure score uses 20,000 independent training episodes; prefix calibration uses 4,000 prefixes with no prior intervention per slot and candidate, including terminated systems. Every hazard is frozen before audit sampling. Training, prefix calibration, audit, and final test streams are disjoint.

Audits, budgets, and uncertainty. The primary tolerance is $\epsilon = 0.01$ with $\delta = 0.05$ per controller. The categorical audit uses 100,000 complete rollouts for each of five components. The shared audit uses 200,000 nominal paths and 60,000 risk rollouts per component; a reduced allocation uses 100,000 and 20,000. In finite-state experiments, counted transitions include absorbing-state steps; the shared allocations cost at most the categorical allocation. CartPole counts actual pre-termination physics transitions. If the completed categorical audit used B transitions, the shared audit allocates $\lfloor B/(2t_{\max})\rfloor$ nominal paths and $\lfloor B/(2 \cdot 5H)\rfloor$ risk episodes per component, so its maximum possible transition cost is at most B. The allocation is fixed from the earlier cost count before drawing fresh audit samples. Common candidate-generation and final-test costs are reported separately.

The controlled comparisons below count nonabsorbing transitions in both tasks. Finite-state risk and joint TV are evaluated exactly; independent rollouts also check execution. CartPole uses 200,000 independent test episodes per method; test TV is estimated from these episodes. Means and 95% percentile bootstrap intervals (Efron, 1979) use 10,000 resamples of independent training seeds; the four penalties are first averaged within each tabular seed. Audit certificates and cross-seed intervals answer different uncertainty questions.

5.1 When activity matching changes the test

The independent-scheduling active shares are 99.989%, 100.000%, 67.554%, and 99.985% in liquidity, inventory, FrozenLake and CartPole (the last is estimated). Activity matching is therefore tested most directly by FrozenLake. With the time law fixed exactly, its full-space time-only optimum increases failure by 11.75 points and active exposure by 8.95 points. Joint matching leaves 5.77 points of additional failure with unchanged exposure. Holding the sample bounds fixed, 60/80 time-only audited rules violate the 1% joint tolerance. This comparison isolates the activity constraint. The raw-candidate audit primarily diagnoses errors in fitting the population law: only 13 of its 939 joint-law violations pass the time-only tolerance.

Table 2: Finite-state confirmation: exact risk increases over independent scheduling, in percentage points, with 95% seed-bootstrap intervals. All methods use a 1% joint total-variation tolerance. Shared-single and shared-mixture use identical audit data. Raw counts show candidate laws exceeding the tolerance before correction. Each task has 20 independent seeds and four penalties per seed.

Task	Raw invalid	Categorical mixture	Shared single	Shared mixture
Liquidity	317/320	0.46 [0.37, 0.58]	0.62 [0.49, 0.78]	0.73 [0.56, 0.92]
Inventory	309/320	1.80 [1.75, 1.85]	2.50 [2.38, 2.64]	2.80 [2.68, 2.92]
FrozenLake-200	313/320	1.27 [1.20, 1.34]	1.74 [1.57, 1.93]	1.83 [1.65, 2.02]

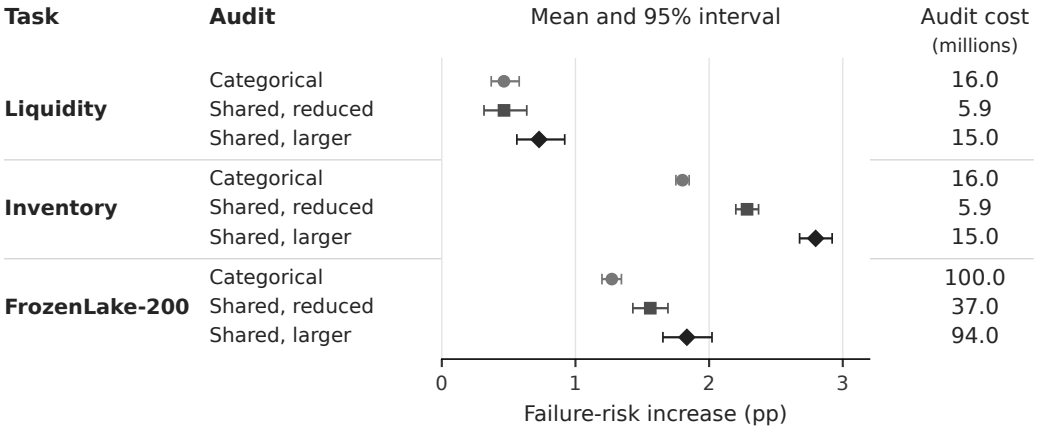

Figure 2: Failure-probability increase over independent scheduling, in percentage points (pp), with audit costs in millions of transitions (right column). Points are means; horizontal bars are 95% percentile bootstrap intervals from 10,000 resamples of 20 independent training seeds, after averaging four penalties within each seed. Shared-path audits use the frozen reduced or larger allocations; categorical audits use separate component rollouts. Every method certifies the same 1% joint total-variation tolerance. These original finite-state costs include absorbing-state steps and exclude the common candidate-construction cost. FrozenLake uses 200 steps.

A fixed late-window probe of the same ten CartPole controllers found 2 with at least a five-point increase in inactive exposure. Using refitted scores, the same library, fresh shared data and identical generic bounds, time-only auditing yields 6.50% and 8.48% held-out joint TV on those controllers; both 95% lower bounds exceed the 1% joint tolerance. Joint auditing achieves 4.56 [3.27, 5.72] points of mean failure gain across all ten. Appendix E gives exposure profiles, all controller results and a certificate-projected time-only comparison.

5.2 Validity and failure-risk increases

Across the three main finite-state tasks, 939 of 960 raw candidate laws exceed the 1% joint-TV tolerance. All primary corrected laws satisfy the tolerance under exact evaluation. Shared-path auditing increases mean failure probability by 0.73, 2.80, and 1.83 percentage points in liquidity, inventory, and FrozenLake-200, respectively. It improves on the categorical mixture while using 6.25%, 6.25%, and 6% fewer counted audit transitions. Figure 2 shows the cost–power comparison for both shared-audit allocations. The exact joint-constrained optimum measures remaining timing vulnerability at zero tolerance. Table 2 reports raw-candidate violations alongside the performance of methods certified at the 1% tolerance.

5.3 Task capability and failure discovery

Table 3 separates task performance from stress-test sensitivity. Penalties 0, 2 and 8 retain 82.36–89.04% nominal goal completion and expose 2.28–2.52 percentage points of additional failure.

Table 3: FrozenLake task capability and audited failure discovery, grouped by training penalty. Completion and independent-shock failure are percentages; shared-audit gains are percentage points with 95% training-seed bootstrap intervals. Each row contains all 20 controllers at that penalty from the 80-controller confirmation.

Penalty	Nominal completion	Independent failure	Audited gain
0	89.04	8.36	2.52 [2.29, 2.74]
2	87.53	2.57	2.47 [2.17, 2.79]
8	82.36	1.88	2.28 [2.05, 2.52]
32	16.95	0.03	0.07 [0.06, 0.08]

Table 4: CartPole failure discovery on the same ten proximal policy optimization controllers. The +2 experiment is the primary confirmation; the supplemental +1.75 experiment follows a baseline-only probe and uses newly fitted stressors and independent audits/tests. Gains and mean certified lower confidence bounds (LCBs) are percentage points; brackets give 95% seed-bootstrap intervals. Each strength uses its own strict matched audit-transition budget.

Audit / selection	+2 gain	LCB	+1.75 gain	LCB
Categorical single	1.74 [0.61, 3.06]	1.67	1.51 [0.67, 2.43]	1.37
Categorical mixture	1.73 [0.62, 3.01]	1.68	1.48 [0.63, 2.39]	1.38
Shared single	2.35 [0.80, 4.20]	2.18	2.17 [0.89, 3.57]	1.92
Shared mixture	2.40 [0.84, 4.27]	2.19	2.28 [1.00, 3.65]	1.96

Penalty 32 lowers completion to 16.95% and yields 0.07 points. A post hoc average of penalties 0/2/8 is 2.42 [2.18, 2.67] points; the primary all-penalty estimate is 1.83 points.

5.4 Continuous-state confirmation and strength sensitivity

The ten controllers have 94.50% mean nominal completion. At +2, independent-shock failure is 84.66% and the shared mixture adds 2.40 [0.84, 4.27] percentage points. At +1.75, baseline failure falls to 57.15%, while the audited gain is 2.28 [1.00, 3.65] points. The shared mixture exceeds the categorical mixture by 0.80 [0.36, 1.27] points at +1.75 and uses 9.96 [7.13, 12.92]% fewer audit transitions. Its mixture-minus-single increment is 0.105 [0.005, 0.220] points (0.050 [-0.037, 0.132] at +2). At +1.75, the shared mixture and certified single select the same episode-level rule for 8 of 10 controllers. Mixing selects a different rule for the remaining 2; its incremental gain is small relative to the improvement from shared auditing.

Null results. On three development PPO seeds, a $+1$ impulse produces zero or nearly zero failure, whereas a $+3$ impulse saturates failure at one. Within the original four-penalty policy sets, selecting by joint-oracle or audit-based risk under a 95% nominal-utility eligibility filter produces zero improvement over independent-risk selection in all 60 seed-environment groups. The utility filter leaves one eligible controller in each of the 20 original 48-step FrozenLake groups; in the 40 resource-task groups, all three criteria select the same controller. The complete strength, task-utility and earlier 48-step FrozenLake results are retained in Appendix D.

5.5 What the audit components contribute

Table 5 isolates path sharing, stopping integration and the reference bound using fresh audits of the same controllers. Both prespecified allocations split the budget equally between law and risk estimation. At the larger allocation, integration changes gain by 0.05 [0.02, 0.08] and 0.01 [-0.02, 0.03] percentage points in FrozenLake and CartPole; the generic confidence radii are identical. CartPole candidate observations coincide under integration and stop draws; only reference-law integration changes this comparison. The DKW reference bound adds 0.23 [0.19, 0.28] and 0.34 [0.12, 0.59] points. At equal law sample size, separate candidate paths use 4.15 and 3.23 times the nominal-path transitions, respectively. The corresponding risk gains are similar. Equal-cap comparisons share a maximum budget; realized costs vary with episode length. Sharing therefore increases the law sample available within the declared cap; integration primarily changes the empirical discrepancy, while

Table 5: Controlled audits with fresh draws: 80 FrozenLake-200 and ten CartPole +2 controllers, 100,000 categorical episodes per component. Gains are percentage points over independent timing (95% seed-bootstrap intervals); costs are millions of nonabsorbing transitions. Shared variants use identical paths and risk samples. DKW: Dvoretzky–Kiefer–Wolfowitz reference bound. Separate paths use integrated observations and generic bounds; equal n uses a larger budget.

Audit	Lake gain	Cost	CartPole gain	Cost
Categorical	1.27 [1.19, 1.34]	60.60	1.76 [0.62, 3.11]	64.15
Shared, one stop draw	1.44 [1.34, 1.54]	42.21	2.19 [0.78, 3.87]	52.88
Shared, integrated	1.49 [1.37, 1.61]	42.21	2.20 [0.78, 3.88]	52.88
Shared, integrated + DKW	1.73 [1.56, 1.89]	42.21	2.54 [0.91, 4.47]	52.88
Separate, equal cap	0.93 [0.88, 0.97]	38.04	1.35 [0.45, 2.44]	41.54
Separate, equal n	1.50 [1.37, 1.62]	112.68	2.21 [0.77, 3.94]	124.50

DKW directly reduces the reference radius. Appendix D.2 reports both budgets, paired contrasts, variance and measured audit-stage runtimes.

What remains within the contract. At the 1% joint-law tolerance, the larger shared audits certify $G > \epsilon$ for 5/80 liquidity, 76/80 inventory, 60/80 FrozenLake-200 configurations, and the strict-budget CartPole +2 audit does so for 5/10 controllers. Each qualifying certificate establishes a positive outcome residual of at least $G - \epsilon$ after removing the cell-frequency contribution. Exact finite-state decompositions confirm that within-cell outcome differences dominate these gains (Table 14). Same-library, same-tolerance exact optima exceed the audited gains (Table 13), separating statistical conservatism from library capacity.

6 Discussion

SameShock measures failure risk under a certified intervention-exposure law. The experiments show that adaptive timing exposes additional failures while respecting this contract. Shared nominal paths improve sampling efficiency; estimating laws and risks dominates the sampling cost. In CartPole +2, averaging the per-controller saving ratios gives 17.94% for auditing alone and 10.11% when prefix construction is included. Controller training, score training and final testing are outside these counts. Path sharing still requires evaluating each candidate on the sampled states. Certifying tight exposure tolerances requires substantial simulation; Equation 4 and Figure 3 show the trade-off between tolerance and sampling budget.

The representation and certificate assume a single fixed shock and independent simulator episodes, with the controller and finite candidate library frozen before auditing. Multiple interacting shocks require additional counterfactual structure. The experiments cover two stylized resource tasks, FrozenLake and CartPole; field-calibrated applications and additional continuous-control systems provide concrete directions for broader evaluation.

7 Conclusion

Matching disturbance strength and count is insufficient to compare adaptive stress tests. SameShock controls the joint scheduled-time/activity law, reuses nominal paths to estimate that law for a frozen library, and selects a tester with a simultaneous feasibility and risk-gain certificate. Exact finite-state comparisons and a fixed late-window neural-control test show why matching time alone changes joint exposure. Controlled audits attribute most sampling gains to path reuse and the reference bound; stopping integration and multi-candidate mixing have smaller, task-dependent effects. Continuous-state confirmation retains additional failure discovery under the joint constraint, with task capability and shock strength reported alongside the gain.

REPRODUCIBILITY STATEMENT

Appendix A gives complete proofs. Appendices B and C specify the environments, controller training, stressor construction and audit budgets; Appendices D and E report additional results and per-controller continuous-control comparisons. Appendix F lists software versions and training seeds. The accelerated implementations were checked against Gymnasium dynamics and Stable-Baselines3 actions. Computation used an Apple M4 CPU. Upon publication, we will release the implementation and scripts for reproducing the experiments under an open-source license.

AI USAGE STATEMENT

Generative AI tools assisted with retrieving relevant literature, revising the manuscript for clarity and consistency, refining experimental protocols, implementing and analysing simulation experiments, and developing mathematical derivations and proofs. Experimental results were obtained from executable simulations and numerical analysis. Verification included source checks, simulator and controller cross-checks, and computational checks of the reported probability laws and risk bounds. The authors take responsibility for the final content.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Eitan Altman and Adam Shwartz. Markov decision problems and state-action frequencies. *SIAM Journal on Control and Optimization*, 29(4):786–809, 1991. doi: 10.1137/0329043. URL https://doi.org/10.1137/0329043.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-aoas1998. URL https://doi.org/10.1214/24-aoas1998.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Erhan Bayraktar and Christopher W. Miller. Distribution-constrained optimal stopping. *Mathematical Finance*, 29(1):368–406, 2019. doi: 10.1111/mafi.12171. URL https://doi.org/10.1111/mafi.12171.

Mathias Beiglböck, Manu Eder, Christiane Elgert, and Uwe Schmock. Geometry of distribution-constrained optimal stopping problems. *Probability Theory and Related Fields*, 172(1–2):71–101, 2018. doi: 10.1007/s00440-017-0805-x. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6191205/.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 1947. doi: 10.1214/aoms/1177730497. URL https://doi.org/10.1214/aoms/1177730497.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Anthony Corso, Robert Moss, Mark Koren, Ritchie Lee, and Mykel Kochenderfer. A survey of algorithms for black-box safety validation of cyber-physical systems. *Journal of Artificial Intelligence Research*, 72:377–428, 2021. doi: 10.1613/jair.1.12716. URL https://doi.org/10.1613/jair.1.12716.

Tommaso Dreossi, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A. Seshia. VerifAI: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *Computer Aided Verification*, pp. 432–442, 2019. doi: 10.1007/978-3-030-25540-4_25. URL https://doi.org/10.1007/978-3-030-25540-4_25.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. doi: 10.1214/14-sts500. URL https://doi.org/10.1214/14-sts500.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27 (3):642–669, 1956. doi: 10.1214/aoms/1177728174. URL https://doi.org/10.1214/aoms/1177728174.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884. URL https://doi.org/10.1287/mnsc.38.6.884.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgEMpVFwB.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, et al. Array programming with NumPy. *Nature*, 585 (7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

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, pp. 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://doi.org/10.1609/aaai.v32i1.11694.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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. URL https://doi.org/10.1214/20-aos1991.

Q. Huangfu and J. A. J. Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5. URL https://doi.org/10.1007/s12532-017-0130-5.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129. URL https://doi.org/10.1287/moor.1040.0129.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Mark Koren, Saud Alsaif, Ritchie Lee, and Mykel J. Kochenderfer. Adaptive stress testing for autonomous vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7, 2018. doi: 10.1109/ivs.2018.8500400. URL https://doi.org/10.1109/ivs.2018.8500400.

Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015. doi: 10.1145/2833157.2833162. URL https://doi.org/10.1145/2833157.2833162.

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

Ritchie Lee, Ole J. Mengshoel, Anshu Saksena, Ryan W. Gardner, Daniel Genin, Joshua Silbermann, Michael Owen, and Mykel J. Kochenderfer. Adaptive stress testing: Finding likely failure events with reinforcement learning. *Journal of Artificial Intelligence Research*, 69:1165–1201, 2020. doi: 10.1613/jair.1.12190. URL https://doi.org/10.1613/jair.1.12190.

Yongyuan Liang, Yanchao Sun, Ruijie Zheng, Xiangyu Liu, Benjamin Eysenbach, Tuomas Sandholm, Furong Huang, and Stephen McAleer. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c8b2f897e45770595656a79a9ad91e89-Abstract-Conference.html.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *International Joint Conference on Artificial Intelligence*, pp. 3756–3762, 2017. doi: 10.24963/ijcai.2017/525. URL https://www.ijcai.org/Proceedings/2017/525.

P. Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Zhiyi Mou, Wangze Ni, Tianfang Xiao, Haoyang Li, Chen Jason Zhang, Hanzhi Ma, Yang Bai, Zhibo Wang, and Kui Ren. From role prompt to infinite thinking: Exploiting persona conditioning for inference cost attacks in LLMs. *arXiv preprint arXiv:2607.25936*, 2026a. URL https://arxiv.org/abs/2607.25936.

Zhiyi Mou, Jingyuan Yang, Zeheng Qian, Wangze Ni, Tianfang Xiao, Ning Liu, Chen Zhang, Zhan Qin, and Kui Ren. SpatialJB: How text distribution art becomes the "jailbreak key" for LLM guardrails. In *International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/65200.

Kosuke Nakanishi, Akihiro Kubo, Yuji Yasui, and Shin Ishii. Off-policy actor-critic for adversarial observation robustness: Virtual alternative training via symmetric policy evaluation. In *International Conference on Machine Learning*, pp. 45624–45662, 2025. URL https://proceedings.mlr.press/v267/nakanishi25a.html.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216. URL https://doi.org/10.1287/opre.1050.0216.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826, 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Nicholas Polosky, Bruno C. Da Silva, Madalina Fiterau, and Jithin Jagannath. Constrained offline policy optimization. In *International Conference on Machine Learning*, pp. 17801–17810, 2022. URL https://proceedings.mlr.press/v162/polosky22a.html.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL https://jmlr.org/papers/v22/20-1364.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224, 2019. URL https://proceedings.mlr.press/v97/tessler19a.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388, 2015a. URL https://proceedings.mlr.press/v37/thomas15.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 3000–3006, 2015b. doi: 10.1609/aaai.v29i1.9541. URL https://doi.org/10.1609/aaai.v29i1.9541.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468): 999–1004, 2019. doi: 10.1126/science.aag3311. URL https://doi.org/10.1126/science.aag3311.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4916. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d7ff1795e8527f6443371c3933bdb52b-Abstract-Datasets_and_Benchmarks_Track.html.

Jonathan Uesato, Ananya Kumar, Csaba Szepesvári, Tom Erez, Avraham Ruderman, Keith Anderson, Krishnamurthy Dvijotham, Nicolas Heess, and Pushmeet Kohli. Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xhQhRcK7.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2. URL https://doi.org/10.1038/s41592-019-0686-2.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003. URL https://shiftleft.com/mirrors/www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf.

Yoriyuki Yamagata, Shuang Liu, Takumi Akazaki, Yihai Duan, and Jianye Hao. Falsification of cyber-physical systems using deep reinforcement learning. *IEEE Transactions on Software Engineering*, 47(12):2823–2840, 2021. doi: 10.1109/tse.2020.2969178. URL https://doi.org/10.1109/tse.2020.2969178.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21024–21037, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f0eb6568ea114ba6e293f903c34d7488-Abstract.html.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=sCZbhBvqQaU.

Siyu Zhou, Patrick Vossler, Venkatesh Sivaraman, Yifan Mai, and Jean Feng. Adaptive auditing of AI systems with anytime-valid guarantees. *arXiv preprint arXiv:2605.07002*, 2026. URL https://arxiv.org/abs/2605.07002.

A Complete proofs and a contract counterexample

A.1 Why the time marginal is insufficient

Consider two equally likely, observed initial types, E and L, both active at time zero. Without a shock, type E terminates successfully before time one, whereas type L remains active. A fixed shock causes failure whenever applied to an active system and has no effect after termination. Let $q_0 = q_1 = 1/2$. Independent scheduling has joint law

$$(p_0(0,1), p_0(0,0), p_0(1,1), p_0(1,0)) = (1/2, 0, 1/4, 1/4)$$

and failure probability $3/4$. A causal stressor that stops at zero for type E and at one for type L preserves the time marginal exactly, but has law $(1/2, 0, 1/2, 0)$ and failure probability one. Its apparent gain comes entirely from the additional $1/4$ physical exposure; its joint TV distance is $1/4$. The example fixes magnitude, count, and scheduled-time distribution. It shows why scheduled mass and physically effective interventions must be distinguished.

A.2 Proof of Proposition 1

For a fixed nominal trajectory, write $a_t = h_j(t, X^0_{\le t})$ and $s_t = \prod_{u<t}(1 - a_u)$ over allowed slots. Since $s_t a_t$ is the probability that independent stopping coins first succeed at t, these nonnegative quantities sum to one when $a_{t_{\max}} = 1$. Each stopped mass is assigned to exactly one activity cell; hence Z_j is a simplex vector.

To establish its expectation, couple the intervened trajectory and the nominal trajectory with the same initial state and nominal transition randomness up to the first shock. The controller's pre-shock branch is the same in both processes. Conditional on the nominal path up to t, the probability of no earlier stop and a stop at t is $s_t a_t$. Consequently, for each cell (t, a), integration over the nominal path gives

$$\mathbb{P}_j(\tau = t, A_\tau = a) = \mathbb{E}\left[\mathbf{1}\{A_t(X^0) = a\} s_t a_t\right].$$

Post-shock dynamics are irrelevant to this identity, because activity is measured immediately before the shock. History-dependent hazards are allowed when their internal state is a deterministic function of the observed prefix; any additional internal randomness must be included in the integrated law or sampled independently as part of the nominal audit. The implemented libraries use state/time scores and the independent stopping coins explicitly described in Equation 3.

For independent scheduling, $a_t = q_t / \sum_{u \ge t} q_u$. Telescoping gives $s_t = \sum_{u \ge t} q_u$ and $s_t a_t = q_t$, proving the reference formula. This construction is a conditional-expectation variance reduction (Blackwell, 1947): integrating the stopping coins cannot increase the variance of any scalar projection of the law observation. □

Relation to simulation reuse. Conditional averaging (Blackwell, 1947) and common random numbers (Glasserman & Yao, 1992) are established variance-reduction tools. Off-policy evaluation reuses behavior-policy data to estimate another policy's value (Dudík et al., 2014; Jiang & Li, 2016; Thomas et al., 2015b). Here the shared pre-shock path identity estimates every frozen candidate's intervention law without likelihood ratios; failure risks use separate intervention rollouts. The proof above identifies the common nominal dynamics that permit this reuse.

A.3 Confidence bounds

For any cell subset D and an independent simplex observation Z_i, $Z_i(D) \in [0, 1]$. Hoeffding's inequality (Hoeffding, 1963) implies

$$\mathbb{P}\left\{ n^{-1} \sum_i Z_i(D) - \mathbb{E} Z_i(D) > r \right\} \le \exp(-2nr^2).$$

Because probability vectors have equal total mass, $\mathrm{TV}(\widehat{p}, p) = \max_D(\widehat{p}(D) - p(D))$. Taking a union over at most 2^m subsets proves Equation 4 for the fractional simplex observations in Equation 3. Sharper bounds are available for categorical observations (Weissman et al., 2003). The support is fixed before sampling and includes unobserved cells.

For the reference, let T denote the nominal time of first inactivity, with $T = \infty$ if inactivity does not occur within the relevant horizon. Absorption gives $A_t = \mathbf{1}\{T > t\}$ under a consistent time convention. The empirical survival function obeys the DKW inequality, including atoms and mass at infinity. For each t, the two activity cells have opposite estimation errors, so

$$\mathrm{TV}(\widehat{p}_0, p_0) = \sum_t q_t |\widehat{S}(t) - S(t)| \le \sup_t |\widehat{S}(t) - S(t)|.$$

The bound $2\exp(-2nr^2)$ yields Equation 5.

For k failures in N independent intervention episodes, the implemented two-sided interval is

$$L = \begin{cases} 0 & k = 0, \\ F^{-1}_{\mathrm{Beta}(k,N-k+1)}(\delta_e/2) & k > 0, \end{cases} \quad U = \begin{cases} 1 & k = N, \\ F^{-1}_{\mathrm{Beta}(k+1,N-k)}(1-\delta_e/2) & k < N. \end{cases}$$

These are Clopper–Pearson limits (Clopper & Pearson, 1934). All $2(K+1)$ law and risk bounds are combined by a union bound. Cross-candidate path sharing is permitted; trajectories within each audit sample must be independent conditional on the frozen library. Audit sample sizes are fixed; time-uniform confidence sequences (Howard et al., 2021) and adaptive auditing (Zhou et al., 2026) address sequential audit sampling and stopping, distinct from within-episode intervention timing. The strict CartPole allocation is fixed from an earlier audit's cost before the fresh shared-audit samples are drawn; the guarantee holds conditionally on that allocation.

Choosing the audit configuration. Choose ϵ from the acceptable deviation in time/activity exposure and choose δ as the allowed certificate failure probability before auditing. We use $\epsilon = 0.01$ and $\delta = 0.05$ per frozen configuration. In Equation 4, the numerator is $m \log 2 + \log[2(K+1)/\delta]$: $m = 2|\mathcal{T}|$ counts exposure cells, K nonreference candidates, and n nominal paths. Thus K enters logarithmically and the number of exposure cells m linearly. More candidates also require additional score evaluations and per-component risk samples. Develop complementary scores using training/prefix data, then freeze the library. Figure 3 reports the original $\epsilon = 0.5\%, 1\%, 2\%$ and audit-size grid at $\delta = 0.05$, including the 48-step FrozenLake setting.

A.4 Proof of Theorem 2

Let $w_0 = 1 - \sum_{j\ge1} w_j$. Randomly selecting a component once per episode gives $p_w = \sum_{j=0}^K w_j p_j$ and $R_w = \sum_{j=0}^K w_j R_j$. On the simultaneous confidence event,

$$\begin{aligned} \mathrm{TV}(p_w, p_0) &\le \tfrac{1}{2} \left\| \sum_{j\ge1} w_j(\widehat{p}_j - \widehat{p}_0) \right\|_1 + \sum_{j\ge1} w_j \left[\mathrm{TV}(p_j, \widehat{p}_j) + \mathrm{TV}(p_0, \widehat{p}_0)\right] \\ &\le B(w), \\ R_w - R_0 &= \sum_{j\ge1} w_j(R_j - R_0) \ge \sum_{j\ge1} w_j(L_j - U_0) = G(w). \end{aligned}$$

The event does not depend on w, so the inequalities hold for all weights simultaneously. They therefore hold for the audit-selected $\widehat{w}$. Feasibility yields $B(\widehat{w}) \le \epsilon$, and the feasible baseline has objective zero, yielding $G(\widehat{w}) \ge 0$. Marginal-law and total active-exposure claims follow by projecting the joint law. □

The implementation solves the LP with nonnegative auxiliary variables u_c satisfying $-u \le Dw \le u$, where column j of D is $\widehat{p}_j - \widehat{p}_0$, and enforces $\frac{1}{2}\mathbf{1}^\mathsf{T} u + e^\mathsf{T} w \le \epsilon$. A fresh component is drawn independently at the start of every episode.

A.5 Comparison within the frozen library

Proposition 3 (Comparison within the frozen library)**.** *On the event of Theorem 2, let* $e_j = r_j + r_0$ *and* $b_j = (U_j - L_j) + (U_0 - L_0)$. *For any mixture* v *satisfying* $\mathrm{TV}(p_v, p_0) + 2\sum_j v_j e_j \le \epsilon$,

$$R_{\widehat{w}} - R_0 \ge (R_v - R_0) - \sum_j v_j b_j.$$

Table 6: Application of Corollary 4 to the exact optimum over the same five-component library and 1% tolerance. Values are mean percentage points over all 80 controllers per task. The comparison expression uses exact evaluator risks for diagnosis; the auditor uses its sampled lower confidence bound (LCB). The guarantee holds on the simultaneous event of Theorem 2.

Task	Library optimum	Comparison bound	Audit LCB	Positive bounds
Liquidity	1.67	0.13	0.46	42/80
Inventory	5.76	1.46	2.50	80/80
FrozenLake-200	3.16	1.10	1.69	74/80

Proof. The triangle inequality gives

$$\tfrac{1}{2}\left\|\sum_j v_j(\widehat{p}_j - \widehat{p}_0)\right\|_1 \le \mathrm{TV}(p_v, p_0) + \sum_j v_j(r_j + r_0).$$

Thus $B(v) \le \mathrm{TV}(p_v, p_0) + 2\sum_j v_j e_j \le \epsilon$. The comparator v is feasible for the audit LP, so $G(\widehat{w}) \ge G(v)$. Moreover,

$$(R_j - R_0) - (L_j - U_0) = (R_j - L_j) + (U_0 - R_0) \le (U_j - L_j) + (U_0 - L_0) = b_j.$$

Combining these inequalities with Theorem 2 proves the claim. The single-candidate solutions are a subset of the full feasible set; hence the corresponding comparison of certified objectives is deterministic for the realized audit. It is not an ordering of realized true risks. □

Corollary 4 (Comparison with every feasible library mixture)**.** *On the event of Theorem 2, fix $\epsilon > 0$ and any library mixture v with $\mathrm{TV}(p_v, p_0) \le \epsilon$. Write $e_v = \sum_{j=1}^K v_j(r_j + r_0)$, $b_v = \sum_{j=1}^K v_j[(U_j - L_j) + (U_0 - L_0)]$, and $\alpha = \epsilon/(\epsilon + 2e_v)$. Then*

$$R_{\widehat{w}} - R_0 \ge G(\widehat{w}) \ge \max\{0, \alpha[(R_v - R_0) - b_v]\}.$$

Proof. The proof of Proposition 3 gives $B(v) \le \mathrm{TV}(p_v, p_0) + 2e_v \le \epsilon + 2e_v$ and $G(v) \ge (R_v - R_0) - b_v$. Scale each nonreference weight to $u_j = \alpha v_j$, assigning the remaining probability to h_0. Positive homogeneity implies $B(u) = \alpha B(v) \le \epsilon$ and $G(u) = \alpha G(v)$. The audit LP can select u or the baseline, so its objective is at least $\max\{0, G(u)\}$. Theorem 2 then yields the claim simultaneously for all such v, including an exact library optimizer on the tolerance boundary. □

A.6 The frequency-only contribution

Write $m_w(c) = \mathbb{E}_w[Y \mid C = c]$ where defined. Then

$$R_w - R_0 = \sum_c p_w(c)(m_w(c) - m_0(c)) + \sum_c (p_w(c) - p_0(c))m_0(c).$$

Since $m_0 \in [0, 1]$, the second summand lies in $[-\mathrm{TV}(p_w, p_0), \mathrm{TV}(p_w, p_0)]$ by the variational characterization of TV. The residual is consequently at least $G(w) - \epsilon$ for certified weights. On cells with $p_0(c) = 0$, any fixed extension of m_0 into $[0, 1]$ maintains the bound. The two terms separate within-cell outcome differences from changes in cell frequencies.

A.7 Conditional exact matching with an unknown law

Consider two allowed times and an observed initial binary state $X \sim \mathrm{Bernoulli}(\theta)$ with unknown $\theta \in (0, 1)$. A frozen hazard chosen from a finite calibration sample D stops at time zero with probabilities $h_D(0)$ and $h_D(1)$. Suppose that, for every θ and every D with positive probability under that θ, the deployed rule must have exact marginal stopping probability $q \in (0, 1)$. It must satisfy

$$(1 - \theta)h_D(0) + \theta h_D(1) = q.$$

Every finite binary sample has positive probability for every $\theta \in (0, 1)$. For each fixed D, the left side is affine in θ and constant only if $h_D(0) = h_D(1) = q$. Thus this universal conditional exactness requirement forbids state dependence already in an elementary problem. Exactness here is required for every realized calibration sample under the unknown law. The audit instead allows a prescribed tolerance with confidence $1 - \delta$.

B ENVIRONMENT AND LEARNING SPECIFICATIONS

Liquidity. The resource stock $x \in \{0, \dots, 24\}$ is absorbing at zero. Actions $a \in \{0, 1, 2\}$ extract resources with reward $a + 0.2$ while $x > 0$. The independent increment takes values $(-2, -1, 1, 2, 3)$ with probabilities $(0.08, 0.12, 0.25, 0.30, 0.25)$. With shock indicator b, the next stock is $\max\{0, \min\{24, x + \text{increment} - a - 6b\}\}$ unless already absorbed. Initial stock is uniform on $\{10, \dots, 14\}$; zero stock is failure. This stylized resource-control environment has horizon 32.

Inventory. State records stock $x \in \{0, \dots, 16\}$ and an arriving order in $\{0, 2, 4\}$, with one additional absorbing failure state. The chosen order $a \in \{0, 2, 4\}$ arrives next period. Independent demand is $0, 1, 2, 3, 4$ with probabilities $(0.08, 0.20, 0.34, 0.25, 0.13)$. After capping replenished stock at 16, subtract demand and an additional 5 on the shocked step. Negative remaining stock causes failure; otherwise the remaining stock and current order form the next state. Reward is $2.5 - 0.35a - 0.08x$ in active states and zero after failure. Initial stock is 8 with arriving order 2. The horizon is 32.

FrozenLake. The canonical 8×8 map is

```
SFFFFFFF
FFFFFFFF
FFFHFFFF
FFFFFHFF
FFFHFFFF
FHHFFFHF
FHFFHFHF
FFFHFFFG
```

At each nonterminal state, nominal motion selects the commanded direction or either adjacent direction with probability $1/3$ each, respecting grid boundaries. Holes are absorbing failures; the goal is absorbing success and pays reward one on arrival. Timeout remains a separate active-at-horizon outcome, not a hole failure. The shock replaces the command a by $(a+2) \bmod 4$ once. The 48-step original condition uses schedule slots $4, \dots, 35$; the 200-step extension uses $20, 30, \dots, 170$. We checked all 256 state-action rows against Gymnasium 1.3.0. The maximum transition-probability discrepancy was 1.11×10^{-16}. Our wrapper adds the single action-reversal intervention to these dynamics.

Tabular learner. For each nominal and shocked state-action row, 400 independent simulator samples estimate transition probabilities. Backward finite-horizon backups maximize reward minus λ times new failure probability, with $\lambda \in \{0, 2, 8, 32\}$. The pre-shock backup uses the independent hazard $q_t / \sum_{u \geq t} q_u$; the post-shock backup uses nominal dynamics. The controller observes time, physical state, and whether a previous shock occurred. It does not observe the current shock decision before choosing its current action. FrozenLake arrival reward is computed for the correct nominal or reversed-action kernel. No environment-specific hyperparameter search is performed on confirmation seeds.

CartPole. The simulator implements Gymnasium's Euler CartPole equations with gravity 9.8, cart mass 1, pole mass 0.1, half-pole length 0.5, force magnitude 10, and step duration 0.02. Initial coordinates are independently uniform on $[-0.05, 0.05]$. Termination occurs when $|x| > 2.4$ or $|\theta| > 12\pi/180$. The action is selected from the current state, then the shock adds the declared positive impulse to angular velocity before the physics update. Once terminated, physical updates stop while schedule accounting continues. We test $H = 200$, compared with the native CartPole-v1 horizon of 500. Return is the number of active physics steps; completion means avoiding termination through the horizon.

PPO uses four environments, rollout length 512, batch size 256, ten epochs, learning rate 3×10^{-4}, a tanh actor hidden layer of width 16, and a critic with two hidden layers of width 32. Training runs for 100,352 steps using CPU evaluation and deterministic argmax deployment. Remaining optimizer settings use the Stable-Baselines3 2.9.0 defaults. The exported float32 actor was compared with the

saved Stable-Baselines3 model on 20,000 random states per controller with zero action discrepancies. On 10,000 states per controller, physics with impulses in $\{0, 1, 2, 3\}$ matched Gymnasium to less than 10^{-10} with zero termination discrepancies.

C CANDIDATE CONSTRUCTION, BUDGETS, AND EXECUTION

Fitted tabular candidates. A second empirical model, independent of the controller's training model, constructs the fitted time and joint LP candidates. The fitted greedy rule allocates each time quota to states ranked by predicted shock-now failure. The prefix-calibrated candidate ranks states using fitted shock-now minus continuation values. At each supported time, it samples fresh nominal prefixes conditional on no previous scheduled shock and selects a threshold with randomized boundary to attain the empirical conditional stopping quota. This conditioning is on prior schedule decisions, not physical survival; absorbed states remain eligible for scheduling. The independent audit measures the resulting population-law discrepancy.

Continuous candidates. The logistic score is trained on observed shock states and binary final failures from 20,000 independent, randomly scheduled training episodes. Features comprise four state coordinates, t/H, and all degree-two monomials of these five inputs (20 features without an intercept feature). Features are standardized on active training states. Logistic regression uses regularization parameter $C = 1$ and at most 1,000 iterations. The intercept is irrelevant for ranking; when all active training labels coincide, the score is constant. The other scores are $|\theta|$, signed angular velocity, and absolute actor-logit gap. Inactive states receive a fixed low priority. Fresh prefix samples calibrate all four candidates to the declared conditional time quotas. The final supported slot always stops with probability one. The same strength-specific library is used by every categorical/shared and single/mixture comparison.

Evaluation and execution. The auditor uses sampled law contributions and failure counts. A separate exact finite-state evaluator propagates occupancies and computes post-shock values for evaluation. Transition tables implement the simulator's sampling interface; they are not supplied to the auditor. CartPole is evaluated through continuous-state rollouts. Nominal paths run to the last supported shock time; risk episodes run to failure or the horizon.

The solver enforces zero hazards outside the declared schedule and a final-slot hazard of one. If presolve reports infeasibility, the full system is retried with dual simplex under the same constraints. Structural-zero enforcement removes numerical residues before auditing.

Cost accounting. Categorical audit counts in finite-state tasks are exactly $5NH$, including absorbing-state transitions. Shared counts are $nt_{\max} + 5N_R H$. The larger allocation uses 15 million versus 16 million transitions for each stylized task, and 94 million versus 100 million for the 200-step lake. The reduced allocation uses 5.9 million and 37 million respectively. These are audit-specific transition counts; simulator time, score evaluation, controller learning, candidate construction, and independent tests are separate computational quantities. For continuous dynamics, absorbing steps are not counted as physics transitions. Strict budget allocation as described in the main text avoids equating nominal and failing-episode lengths.

Table 7: Sampling costs. Controller training counts are transitions per model seed (shared across four penalties) for tabular tasks, and per controller for CartPole. Prefix and audit columns are mean millions of transitions per controller. A separate tabular stressor-model fit costs the same number of transitions as its controller-model fit per seed. CartPole additionally uses 20,000 score-training episodes (at most four million physics steps). Candidate construction is identical between compared audits.

Task	Controller training	Prefix	Categorical audit	Shared audit
Liquidity	60,000	3.98	16.00	15.00
Inventory	124,800	3.98	16.00	15.00
FrozenLake-200	204,800	13.53	100.00	94.00
CartPole-200 (+2)	100,352	48.45	64.15	52.87

D ADDITIONAL RESULTS AND AUDIT ABLATIONS

The following post hoc mechanism ablation holds the shared trajectories, risk episodes, library, and LP fixed and replaces only the reference-specific Dvoretzky–Kiefer–Wolfowitz (DKW) radius with the generic simplex radius. For a zero-distance comparator, Proposition 3 permits total nonbaseline weight at most 36.52% in the two resource tasks and 41.12% in FrozenLake at the original larger shared allocation. This slack-dependent comparison is complemented by Corollary 4, which covers boundary mixtures by shrinking their nonreference weights.

Table 8: Exact failure-risk increase in percentage points with 95% seed-bootstrap intervals under identical shared-audit data. The ablation measures the contribution of the reference radius, with all methods retaining a valid 1% joint total-variation certificate.

Task	Generic reference bound	DKW reference bound
Liquidity	0.59 [0.46, 0.74]	0.73 [0.56, 0.92]
Inventory	2.29 [2.21, 2.38]	2.80 [2.68, 2.92]
FrozenLake-200	1.60 [1.46, 1.74]	1.83 [1.65, 2.02]

Table 9: Task utility for all tabular penalty settings. Nominal utility is expected unpenalized return (goal-completion probability for FrozenLake). Success and timeout are percentages under the categorical-audited mixture's independent test episodes. In resource tasks, success denotes survival through the horizon.

Task	Penalty	Nominal utility	Test success (%)	Timeout (%)
Liquidity	0	52.406	63.93	0.00
Liquidity	2	51.979	71.03	0.00
Liquidity	8	51.149	88.20	0.00
Liquidity	32	49.980	96.85	0.00
Inventory	0	39.224	41.31	0.00
Inventory	2	39.033	76.61	0.00
Inventory	8	38.340	91.14	0.00
Inventory	32	37.486	96.06	0.00
FrozenLake-200	0	0.890	86.92	3.04
FrozenLake-200	2	0.875	85.49	10.26
FrozenLake-200	8	0.824	83.35	13.19
FrozenLake-200	32	0.170	57.21	42.71

Table 10: All three development CartPole seeds at each prespecified impulse size. This sensitivity table is developmental, separate from the ten-seed confirmation at strength 2. Audit uses the original categorical allocation.

Impulse	Independent failure (%)	Audited increment (pp)
1	0.033	0.028
2	82.929	1.834
3	100.000	0.000

D.1 SUPPLEMENTAL CARTPOLE STRENGTH EXPERIMENT

The +1.75 experiment uses every controller in the original +2 confirmation. An exploratory baseline-only probe of 20,000 episodes per controller motivated this operating point. The new design fixes 20,000 score-training episodes, 4,000 prefix samples per slot and candidate, 100,000 categorical audit episodes per component, and 200,000 final-test episodes per method. Shared allocations follow the same strict physical-budget rule as the original confirmation. All candidate fitting, audit and test streams are separated from the original experiment and the baseline probe. Controller weights are reused unchanged. The score and prefix thresholds are fitted anew at +1.75 before auditing. Table 17 reports the results for every controller.

Table 11: Raw candidate comparisons in the three main finite-state tasks. Values are means across all 80 controllers per task. Risk increases are in percentage points; joint total variation (TV) is expressed as a percentage. Invalid counts refer to the 1% joint-TV threshold.

Task	Candidate	Risk increase	Joint TV	Invalid
Liquidity	fitted greedy	1.46	2.55	78/80
Liquidity	fitted joint lp	1.37	2.60	80/80
Liquidity	fitted time lp	2.02	2.54	79/80
Liquidity	prefix calibration	2.01	1.89	80/80
Inventory	fitted greedy	7.65	2.51	77/80
Inventory	fitted joint lp	7.67	2.43	76/80
Inventory	fitted time lp	7.67	2.42	76/80
Inventory	prefix calibration	7.69	2.04	80/80
FrozenLake-200	fitted greedy	9.83	23.62	79/80
FrozenLake-200	fitted joint lp	5.70	2.79	75/80
FrozenLake-200	fitted time lp	11.71	13.38	79/80
FrozenLake-200	prefix calibration	11.71	12.58	80/80

Table 12: Independent risk, exact zero-tolerance joint-oracle risk increment, and categorical certified-single increment. Independent risk is expressed as a percentage; risk increments are in percentage points. The oracle uses zero tolerance. The certified single and the main-table categorical mixture use the same audit data and 1% tolerance.

Task	Independent risk	Exact joint increment	Categorical single
Liquidity	19.54	1.40	0.42 [0.34, 0.52]
Inventory	21.90	7.69	1.66 [1.61, 1.72]
FrozenLake-200	3.21	5.77	1.20 [1.13, 1.28]

D.2 Controlled sampling and exposure comparisons

The controlled sampling experiments reuse the previously evaluated frozen controller and stressor libraries from the main confirmation. The two categorical allocations, 50,000 and 100,000 episodes per component, the equal law/risk budget split, metrics and sampling streams were fixed before drawing the new audit batches. Both allocations use the full controller set with unchanged training and prefix calibration.

For each completed categorical audit, let B be its nonabsorbing transition count. Shared auditing allocates $n = \lfloor B/(2t_{\max}) \rfloor$ nominal paths and $N_R = \lfloor B/(2 \cdot 5H) \rfloor$ risk episodes per component. The independent-path audit uses $\lfloor n/5 \rfloor$ law paths per component under the same maximum budget; an additional equal-n comparator measures its larger actual cost. Independent paths end when their stopping mass is exhausted. These allocations compare declared caps, while the tables also report realized costs; different methods need not exhaust their cap.

Each shared path produces both its integrated law contribution and one sampled stopping cell per candidate. Independent stopping uniforms sample the latter conditional on the nominal path. All shared-law variants reuse identical risk batches. Generic radii remain unchanged between one-hot and integrated observations at equal n; only empirical laws change. The DKW variant replaces the reference radius using its integrated survival estimate. Independent-path low and equal-n allocations use nested samples for each candidate. Sample trace variances, score evaluations, simulation counts, law-pass time, risk-pass time and solver time were measured. Table 23 sums these recorded durations on the Apple M4 CPU after kernel warm-up. Categorical rollouts estimate laws and risks together; the other methods add a nominal-law pass to their intervention rollouts. Law-pass timings include both paired estimators. Shared variants reuse the same measured sampling passes and differ in their solver calls. Times exclude controller training, candidate construction and final testing.

FrozenLake laws and risks are evaluated exactly. CartPole evaluates every frozen component on 200,000 fresh episodes, then uses mixture linearity for the risk estimates of all methods. These test batches are common across methods and excluded from selection. Continuous joint-TV values

Table 13: Original shared audit and exact mixture optimum over the same frozen library at the same 1% total-variation tolerance. Gains are percentage points with 95% seed-bootstrap intervals. The last column counts controllers whose certified gain exceeds the full 1% frequency allowance.

Task	Shared audit gain	Exact library gain	$G > \epsilon$
Liquidity	0.73 [0.56, 0.92]	1.67 [1.34, 2.07]	5/80
Inventory	2.80 [2.68, 2.92]	5.76 [5.31, 6.24]	76/80
FrozenLake-200	1.83 [1.65, 2.02]	3.16 [2.64, 3.73]	60/80

Table 14: Post hoc paired mixture-minus-single increments, cell-frequency contributions and within-cell outcome residuals from the original larger shared audits, in percentage points. Finite-state decompositions are exact; intervals resample training seeds. The residual is the algebraic decomposition defined in the main text, not a conditional causal effect. Five of ten original strict-budget CartPole +2 certificates satisfy $G > \epsilon$.

Task	Mixture minus single	Frequency	Within-cell residual
Liquidity	0.104 [0.069, 0.146]	-0.0027 [-0.0041, -0.0013]	0.730 [0.566, 0.922]
Inventory	0.295 [0.231, 0.358]	0.0011 [-0.0016, 0.0038]	2.796 [2.675, 2.922]
FrozenLake-200	0.093 [0.052, 0.139]	0.0106 [0.0064, 0.0151]	1.823 [1.644, 2.009]
CartPole-200	0.050 [-0.037, 0.132]	not exact	not exact

remain empirical. All cross-controller intervals resample training seeds; four penalties are averaged within each FrozenLake seed.

The sample trace variances of candidate law observations (averaged over nonbaseline candidates and controllers) are 0.6200 versus 0.9486 for integrated versus one-hot observations in FrozenLake, and 0.9374 versus 0.9374 in CartPole at the larger allocation. The CartPole reference variances are 0.0000182 and 0.9375, respectively. All sampled CartPole candidate contributions are one-hot, so integrating their stopping rule removes no additional randomness on these paths. Reference integration accounts for the remaining effect. The confidence radii are computed independently of these empirical variance estimates. All 960 finite-state method/allocation rows satisfy the exact law and risk lower bounds in evaluation.

The exposure analysis reuses the original larger shared audits and compares three finite-library certificates. Marginalizing their joint observations with the same law radii isolates the constraint geometry. A stronger time-only certificate uses the known reference q, radius zero for that reference, and the generic radius with $|\mathcal{T}|$ cells for the other candidates. Both time-only certificates control the time marginal. Exact library optima use the same 1% tolerance; the full stopping-space references instead use exact equality.

The full-space references eliminate slots with $q_t = 0$: initial mass is propagated to the first allowed slot, nominal kernels between successive allowed slots are multiplied, and the shock-now values retain the full 200-step outcome horizon. This eliminates forced-continuation variables while preserving the state space. All compressed solutions were checked by executing their hazards in the full-horizon evaluator.

Table 15: CartPole strength-specific operating regimes and mean physical transition costs. Failure and active exposure are percentages under independent scheduling; costs are millions of nonabsorbing transitions per controller. Prefix construction is shared by the two audits. These costs exclude controller training, score-training episodes and final evaluation.

Impulse	Failure	Active	Categorical	Shared	Prefix
+2	84.66	99.985	64.15	52.87	48.45
+1.75	57.15	99.985	80.08	72.57	48.62

Table 16: All strength and selection comparisons. Gains and mean lower confidence bounds (LCBs) are percentage points; empirical joint total variation (TV) is a percentage. Test TV has sampling uncertainty. Fallback counts denote exact selection of the independent reference, out of ten controllers; independent final-test noise can give a nonzero estimated gain on a fallback.

Impulse	Selection	Gain [95% interval]	LCB	Test TV	Fallback
+2	Categorical single	1.74 [0.61, 3.06]	1.67	0.63	2
+2	Categorical mixture	1.73 [0.62, 3.01]	1.68	0.60	2
+2	Shared single	2.35 [0.80, 4.20]	2.18	0.68	2
+2	Shared mixture	2.40 [0.84, 4.27]	2.19	0.73	2
+1.75	Categorical single	1.51 [0.67, 2.43]	1.37	0.63	3
+1.75	Categorical mixture	1.48 [0.63, 2.39]	1.38	0.57	3
+1.75	Shared single	2.17 [0.89, 3.57]	1.92	0.69	3
+1.75	Shared mixture	2.28 [1.00, 3.65]	1.96	0.69	3

Table 17: Every controller in the supplemental +1.75 experiment. Nominal completion and independent-shock failure are percentages; remaining columns are held-out failure-risk increments in percentage points. Cat. denotes categorical auditing.

Seed	Completion	Failure	Cat. mix	Shared single	Shared mix
151000	70.66	23.61	0.683	0.804	0.785
151001	100.00	74.35	0.625	1.055	1.060
151002	100.00	31.98	4.021	6.389	6.253
151003	100.00	35.82	2.298	3.347	3.350
151004	100.00	36.39	2.598	4.009	4.019
151005	100.00	87.53	-0.075	-0.096	0.076
151006	74.26	90.75	0.054	-0.195	-0.040
151007	100.00	65.49	1.287	1.681	2.137
151008	99.76	79.47	-0.223	-0.101	-0.071
151009	100.00	46.09	3.552	4.853	5.224

Table 18: FrozenLake-200 controlled audits at both categorical allocations. Gains are percentage points with seed-bootstrap intervals; LCB is the mean certified lower confidence bound, cost is millions of nonabsorbing transitions, and fallback counts refer to controllers. The equal-n separate-path audit is outside the common maximum budget.

Categorical N	Audit	Gain	LCB	Cost	Fallback
50000	Categorical	1.03 [0.98, 1.07]	0.94	30.30	0
50000	Shared, one stop draw	1.19 [1.12, 1.25]	0.99	21.10	0
50000	Shared, integrated	1.23 [1.15, 1.31]	1.03	21.10	0
50000	Shared, integrated + DKW	1.46 [1.34, 1.57]	1.21	21.10	0
50000	Separate, equal cap	0.71 [0.69, 0.74]	0.60	19.02	0
50000	Separate, equal n	1.23 [1.15, 1.31]	1.02	56.34	0
100000	Categorical	1.27 [1.19, 1.34]	1.19	60.60	0
100000	Shared, one stop draw	1.44 [1.34, 1.54]	1.27	42.21	0
100000	Shared, integrated	1.49 [1.37, 1.61]	1.31	42.21	0
100000	Shared, integrated + DKW	1.73 [1.56, 1.89]	1.51	42.21	0
100000	Separate, equal cap	0.93 [0.88, 0.97]	0.82	38.04	0
100000	Separate, equal n	1.50 [1.37, 1.62]	1.31	112.68	0

Table 19: Paired component effects for FrozenLake-200. Units are percentage points. Path-sharing contrasts use generic reference bounds in both methods; integration and DKW contrasts hold the indicated samples fixed.

Categorical N	Contrast	Gain difference
50000	Integration minus stop draw	0.044 [0.023, 0.067]
50000	DKW minus generic reference	0.225 [0.194, 0.259]
50000	Shared minus separate, equal cap	0.518 [0.454, 0.581]
50000	Shared minus separate, equal n	0.003 [-0.010, 0.015]
100000	Integration minus stop draw	0.052 [0.022, 0.084]
100000	DKW minus generic reference	0.233 [0.192, 0.276]
100000	Shared minus separate, equal cap	0.567 [0.487, 0.651]
100000	Shared minus separate, equal n	-0.002 [-0.012, 0.006]

Table 20: CartPole-200 controlled audits at both categorical allocations. Gains are percentage points with seed-bootstrap intervals; LCB is the mean certified lower confidence bound, cost is millions of nonabsorbing transitions, and fallback counts refer to controllers. The equal-n separate-path audit is outside the common maximum budget.

Categorical N	Audit	Gain	LCB	Cost	Fallback
50000	Categorical	1.38 [0.50, 2.40]	1.26	32.08	2
50000	Shared, one stop draw	1.70 [0.59, 2.99]	1.44	26.44	3
50000	Shared, integrated	1.80 [0.61, 3.22]	1.53	26.44	3
50000	Shared, integrated + DKW	2.13 [0.73, 3.80]	1.81	26.44	3
50000	Separate, equal cap	1.05 [0.35, 1.88]	0.90	20.76	3
50000	Separate, equal n	1.83 [0.62, 3.26]	1.55	62.24	3
100000	Categorical	1.76 [0.62, 3.11]	1.65	64.15	2
100000	Shared, one stop draw	2.19 [0.78, 3.87]	1.96	52.88	2
100000	Shared, integrated	2.20 [0.78, 3.88]	1.97	52.88	2
100000	Shared, integrated + DKW	2.54 [0.91, 4.47]	2.26	52.88	2
100000	Separate, equal cap	1.35 [0.45, 2.44]	1.22	41.54	2
100000	Separate, equal n	2.21 [0.77, 3.94]	1.98	124.50	2

Table 21: Paired component effects for CartPole-200. Units are percentage points. Path-sharing contrasts use generic reference bounds in both methods; integration and DKW contrasts hold the indicated samples fixed.

Categorical N	Contrast	Gain difference
50000	Integration minus stop draw	0.107 [0.014, 0.234]
50000	DKW minus generic reference	0.328 [0.113, 0.582]
50000	Shared minus separate, equal cap	0.752 [0.256, 1.351]
50000	Shared minus separate, equal n	-0.024 [-0.041, -0.007]
100000	Integration minus stop draw	0.007 [-0.015, 0.034]
100000	DKW minus generic reference	0.336 [0.124, 0.585]
100000	Shared minus separate, equal cap	0.850 [0.322, 1.461]
100000	Shared minus separate, equal n	-0.006 [-0.096, 0.068]

Table 22: Full post hoc FrozenLake contract comparisons. Gain, active-exposure change and frequency contribution are percentage points; total variation (TV) is in percent. Exact global references use zero tolerance, while library optima and audited rules use 1% tolerance on their respective target. Common-bound auditing changes only the norm after marginalization; known-q auditing also uses the sharper time-only confidence information.

Method	Gain	Time TV	Joint TV	Active change	Frequency
Independent	0.00	0.00	0.00	0.00	0.000
Exact time-only	11.75	0.00	12.05	8.95	0.916
Exact joint	5.77	0.00	0.00	0.00	0.000
Library time-only	10.22	0.99	10.73	8.09	0.798
Library joint	3.16	0.92	1.00	0.22	0.014
Audited joint	1.83	0.54	0.59	0.16	0.011
Time-only, common bounds	5.03	0.47	5.48	4.26	0.416
Time-only, known q	6.89	0.61	7.32	5.73	0.553

Table 23: Recorded audit-stage runtimes and failure-risk gains for the controlled experiments. Times sum law-pass, intervention-rollout and solver durations after kernel warm-up on an Apple M4 CPU. Values are means over 20 FrozenLake training seeds (four penalties averaged within seed) and ten CartPole controllers. Gain is the increase over independent timing in percentage points (pp). N is the categorical episode count per component. Law-pass times include both paired estimators; shared variants reuse the same measured sampling passes. DKW: Dvoretzky–Kiefer–Wolfowitz reference bound. Separate equal-n paths use a larger budget.

		FrozenLake-200		CartPole +2	
Categorical N	Audit	Time (s)	Gain (pp)	Time (s)	Gain (pp)
50,000	Categorical	0.69	1.03	3.87	1.38
	Shared, one stop draw	0.45	1.19	3.37	1.70
	Shared, integrated	0.45	1.23	3.37	1.80
	Shared, integrated + DKW	0.45	1.46	3.37	2.13
	Separate, equal cap	0.39	0.71	2.48	1.05
	Separate, equal n	1.07	1.23	7.39	1.83
100,000	Categorical	1.38	1.27	7.72	1.76
	Shared, one stop draw	0.90	1.44	6.76	2.19
	Shared, integrated	0.90	1.49	6.76	2.20
	Shared, integrated + DKW	0.90	1.73	6.76	2.54
	Separate, equal cap	0.77	0.93	4.96	1.35
	Separate, equal n	2.11	1.50	14.78	2.21

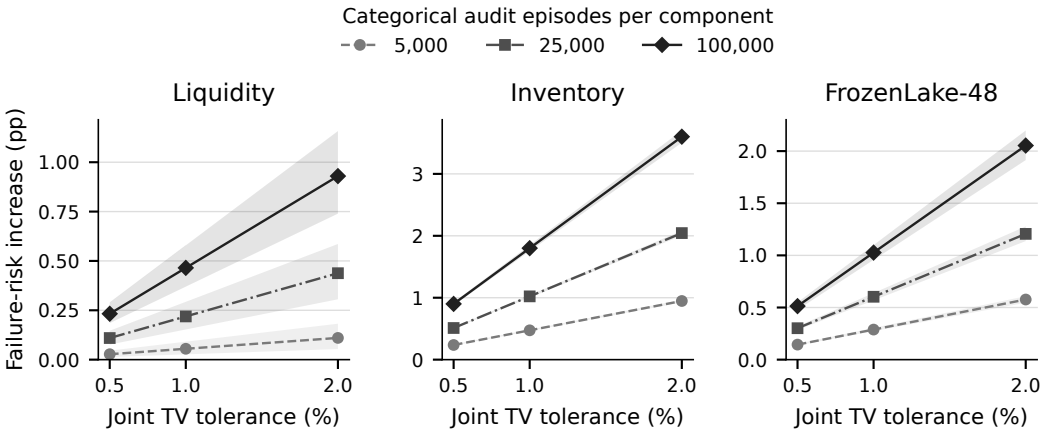

Figure 3: Audit-size and tolerance sensitivity in the finite-state tasks, with a 48-step FrozenLake horizon. Lines show exact failure-probability increases over independent scheduling, in percentage points (pp). Shading gives 95% percentile bootstrap intervals from 10,000 resamples of 20 independent training seeds; four penalties are averaged within each seed. Marker shapes and line styles distinguish audit sizes across the three joint total-variation (TV) tolerances. Audit size is the number of categorical intervention episodes per component.

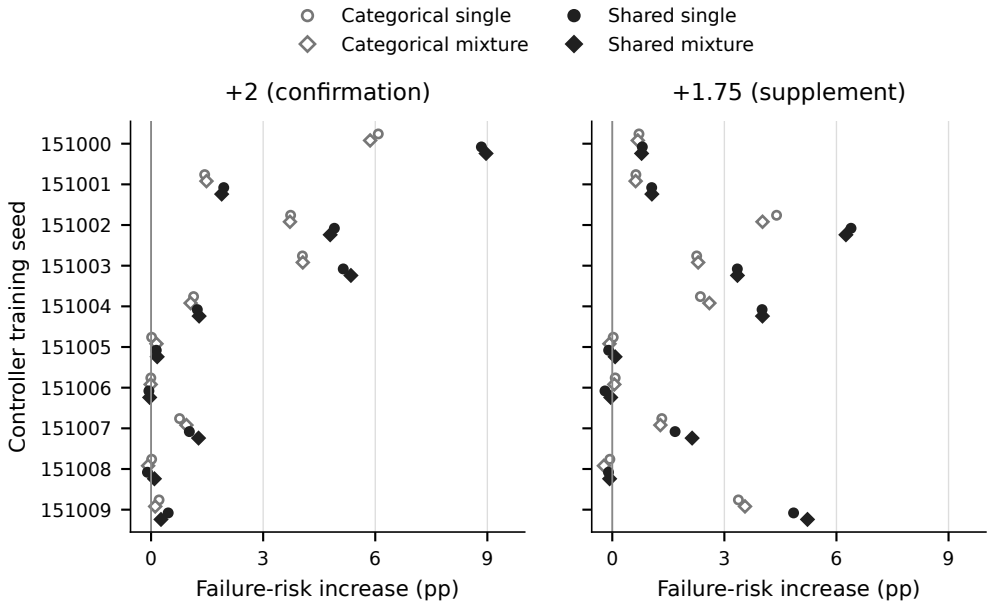

Figure 4: Per-controller failure discovery at two fixed angular-velocity impulses, $+2$ and $+1.75\,\mathrm{rad\,s^{-1}}$. Both panels retain all ten controllers trained with proximal policy optimization (PPO), on the same axis scale. Points are independent-test failure-probability increases over the strength-specific independent-scheduling reference, in percentage points (pp). Circles denote certified-single selection and diamonds mixture selection; open and filled symbols denote categorical and shared audits. Methods are offset within each seed for visibility. Shared single and mixture select identical weights for nine controllers at $+2$ and eight at $+1.75$; their independent test estimates can differ. The supplemental strength uses fresh fitting, audit, and test samples. Negative test differences remain visible.

Table 24: Fixed late-window CartPole comparison on all ten existing controllers with a $+2\,\mathrm{rad\,s^{-1}}$ angular-velocity impulse. Failure is in percent; gain and active-exposure change are percentage points relative to independent scheduling. Brackets are 95% intervals from 10,000 training-seed bootstrap draws. Joint total variation (TV) is the mean empirical final-test distance.

Rule	Failure (%)	Gain	Active change	Joint TV (%)
Independent	41.49 [36.87, 48.42]	0.00 [0.00, 0.00]	0.000	0.00
Time-only audit	47.47 [42.51, 55.14]	5.98 [5.17, 6.91]	0.063	2.03
Joint audit	46.05 [41.78, 52.17]	4.56 [3.27, 5.72]	-0.033	0.75

E FIXED LATE-WINDOW COMPARISON IN CONTINUOUS CONTROL

The original CartPole schedule places almost all interventions before natural termination. To test the activity constraint within the same continuous-state task, we fixed the later support $\{180, 185, 190, 195, 199\}$, retained the 200-step horizon and the $+2$ impulse, and kept all ten frozen controllers. Before sampling new shock outcomes, a nominal-only gate required at least two controllers to increase inactive exposure by five percentage points relative to the old schedule. Each controller supplied 20,000 fresh nominal paths; activity was recorded immediately before every step, with terminated episodes retained in the denominator. Seeds 151000 and 151006 passed this gate, with increases of 17.36 and 15.31 points. The average active exposure over all ten controllers changed from 99.987% to 96.715%.

Each controller's quadratic logistic failure score was refitted using 20,000 new independent-shock training episodes in the late window. This places the time feature inside its training support. The other three score families were unchanged. Each nonfinal candidate slot used 4,000 prefixes with no prior intervention, including physically inactive systems. The five-component library was frozen before observing the audit. Score fitting, prefix construction, audit, and final tests used disjoint random streams.

Both contracts used the same 100,000 shared nominal paths and 20,000 risk episodes per component. All five law radii used the generic ten-cell bound with event failure probability $0.05/10$. The time-only optimization summed the same joint observations over activity and kept these radii and risk intervals unchanged. Thus the comparison changes the law constraint while holding the evidence and confidence information fixed. Both tolerances were 1%; each selected rule draws its component once per episode.

We then generated 50,000 independent test episodes for each of the independent, time-only, and joint rules on every controller. The time-only rules for the two gate-positive controllers have empirical joint total variation (TV) of 6.50% and 8.48%, with conservative 95% lower bounds of 4.40% and 6.38%. Their aggregate active shares change by only $+1.58$ and -0.94 points: Figure 5 shows substantial redistribution between the time/activity cells despite this smaller aggregate change. The corresponding joint-audited empirical test distances are 1.51% and 1.02%. All joint certificates satisfy $B \leq 1\%$ and retain positive certified gains. Seven controller pairs select identical weights under the two contracts. Tables 24 and 25 retain every controller and report both comparisons.

Certificate-projected time-only comparator. Let v be the time-only optimum's nonreference weights and $\widehat{w}_J$ the joint optimum. Evaluate $B_J(v)$ using the same joint observations, generic radii and risk intervals. Project toward the independent reference by taking $u = \alpha v$, with $\alpha = \min\{1, \epsilon/B_J(v)\}$ (or one when $B_J(v) = 0$), and assign the remaining mass to the reference. Positive homogeneity gives

$$B_J(u) = \alpha B_J(v) \leq \epsilon, \qquad G(u) = \alpha G(v) \leq G(\widehat{w}_J).$$

The last inequality follows from feasibility in the joint optimization and compares certified objectives. The empirical weight comparison is reported below.

The projected and joint rules select identical weights for 10 of 10 controllers. These identical rules share the reported joint-rule test results.

Table 25: All ten late-window controllers. Nominal active exposure (Act.) is from the independent 20,000-path gate. T and J denote time-only and joint auditing; gains are percentage points and total variation (TV) is in percent. T lower is the conservative 95% lower bound on held-out joint TV. G_J is the joint audit's certified risk-gain lower bound. Every selected rule has target-TV upper bound $B = 1\%$; there are no reference fallbacks.

Seed	Act.	Gain T	Gain J	TV T	T lower	TV J	G_J
151000	82.56	5.94	1.14	6.50	4.40	1.51	0.79
151001	100.00	5.63	5.67	0.81	0.00	0.76	4.10
151002	100.00	5.93	5.61	0.62	0.00	0.32	4.75
151003	100.00	6.22	6.51	0.48	0.00	0.43	4.81
151004	100.00	3.49	2.90	0.64	0.00	0.70	2.88
151005	100.00	6.71	6.72	1.02	0.00	0.68	6.14
151006	84.63	9.24	1.09	8.48	6.38	1.02	0.97
151007	100.00	6.07	5.94	0.57	0.00	1.04	5.24
151008	99.96	4.99	4.95	0.49	0.00	0.56	3.91
151009	100.00	5.61	5.13	0.69	0.00	0.51	4.43

F Computational setup and random seeds

The experiment environment uses Python 3.12, NumPy 2.5.2 (Harris et al., 2020), SciPy 1.18.1 (Virtanen et al., 2020), Numba 0.67.0 (Lam et al., 2015), Gymnasium 1.3.0, Stable-Baselines3 2.9.0, PyTorch 2.14.0 (Paszke et al., 2019) and scikit-learn 1.9.0 (Pedregosa et al., 2011). Linear programs use SciPy's HiGHS interface (Huangfu & Hall, 2018). Computation runs on a CPU, with deterministic argmax controller deployment and disjoint controller-training, candidate-fitting, audit and final-test random streams. Freezing candidates before the audit avoids reusing holdout data for candidate construction, a source of selection bias in adaptive data analysis (Dwork et al., 2015).

The liquidity and inventory confirmation and the 48-step FrozenLake experiment use controller-model seeds 101000–101019. The 200-step FrozenLake confirmation uses seeds 202000–202019; its results are not pooled with the 48-step setting. The tabular stressor-model seed is the controller-model seed plus 10,000. CartPole confirmation uses PPO training seeds 151000–151009. The +1.75 and late-window comparisons reuse these ten controllers, with fresh candidate fits, audits and tests as specified in Appendices D and E. The three development controllers use seeds 51000–51002 and remain separate from confirmation.

Intervals across independent training runs quantify training-seed variation, following recommendations for reliable reinforcement-learning evaluation (Henderson et al., 2018; Agarwal et al., 2021).

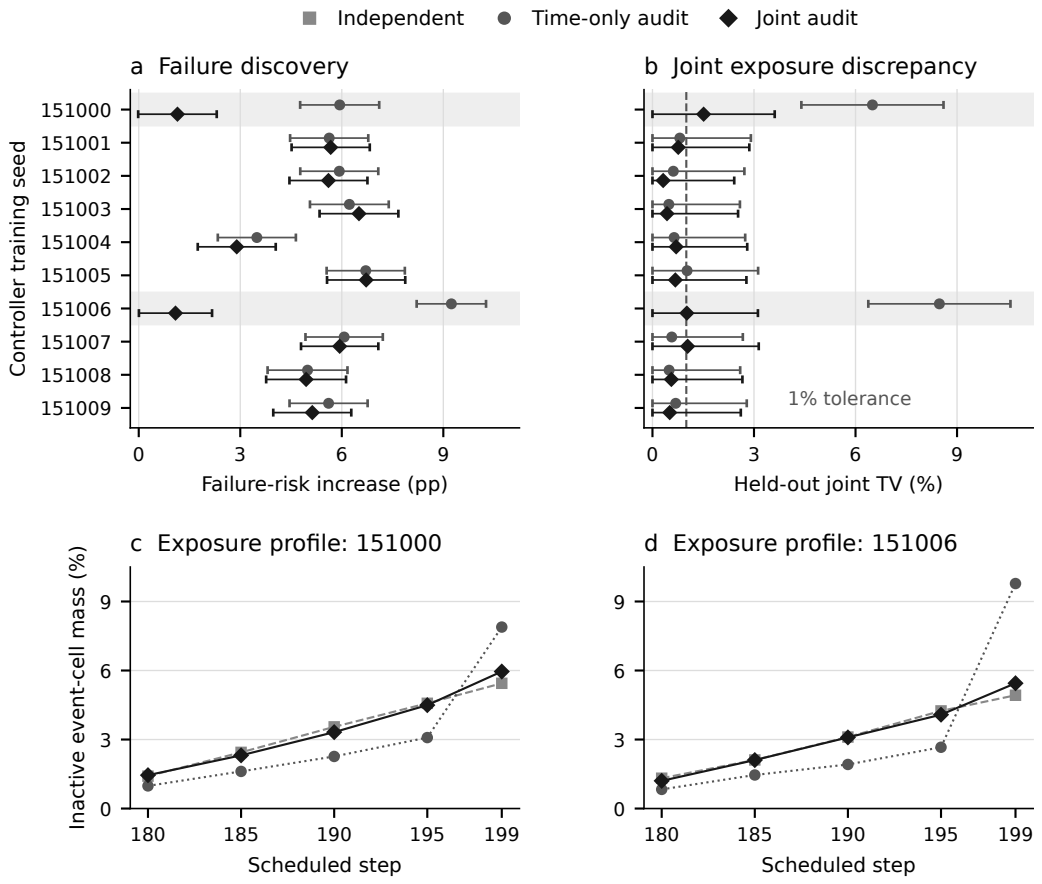

Figure 5: Fixed late-window CartPole comparison at a $+2\,\mathrm{rad\,s^{-1}}$ angular-velocity impulse. Upper panels retain all ten controllers and show held-out failure-probability increases in percentage points (pp) and joint total variation (TV) in percent. Bars are conservative within-controller 95% intervals from 50,000 independent episodes per rule: binomial bounds for risk differences and generic law bounds for TV. The dashed line marks the 1% joint tolerance. Shaded rows identify the two controllers that passed the prior nominal-only gate. Lower panels show their held-out inactive event-cell masses, $\widehat{\Pr}(\tau = t, A_\tau = 0)$, with all initial episodes in the denominator. Points are measured scheduled slots; connecting segments guide the eye. Time-only auditing redistributes inactive events across slots. Joint auditing controls the combined time/activity law. All methods share the candidate library, audit observations and generic radii.