Testing Before Trusting: Causal Skeptic Bandits under Unreliable Historical Evidence

Anonymous submission

Abstract

Warm-started contextual bandits face a trust problem: historical logs and causal graphs can accelerate learning when valid, yet cause negative transfer under confounding or graph error. We propose Causal Skeptic Bandits (CSB), which maintain naive, causally adjusted, and cold-start interpretations as online-testable candidate worlds. Online interventional prediction errors update their influence, while diagnostic actions target disagreements that ordinary reward exploration may not resolve. A causal diagnostic veto (CDV) then rescues adjusted transfer only when supported and blends toward a cold fallback when it is contradicted. Controlled experiments isolate both mechanisms: diagnostics reduce regret in probe-necessary regimes, and the veto substantially limits wrong-graph transfer, although causal and cold-start learners remain the correct endpoint oracles. On ten public IHDP semi-synthetic replications, CDV is competitive with strong warm starts and clearly improves over cold start, but ties its no-diagnostic ablation. Finally, unannounced regime switches expose path dependence in cumulative trust and rule out a stronger nonstationary-adaptation claim. The supported conclusion is therefore predeployment regime-uncertainty control, not universal or instantaneous graph validation. **Content areas:** ML, RU.

Introduction

Warm-started contextual bandits are common in recommendation, advertising, clinical decision support, and experimental design. A deployed learner rarely starts from scratch; it is initialized from logs collected by previous policies or domain experts. Such logs are valuable, but their actions are not neutral observations. A physician may prescribe aggressive treatment because a patient already appears high risk, and a recommender may expose an item because a user already appears likely to engage. Consequently, historical action-reward associations may differ from interventional effects (Rubin 1974; Pearl 2009; Dudík, Langford, and Li 2011; Bottou et al. 2013; Swaminathan and Joachims 2015; Joachims, Swaminathan, and Schnabel 2017).

Causal bandits use structural assumptions to improve exploration (Lattimore, Lattimore, and Reid 2016; Lee and Bareinboim 2018), but practical deployments may not know whether the graph or adjustment set is correct. If the learner ignores logs, it wastes useful experience; if it trusts them naively, it can inherit spurious correlations; if it trusts a wrong graph, causal transfer can become catastrophic. The key deployment question is therefore not which warm start wins after the reliability regime has been labeled. It is how to act while that regime is still uncertain. The missing object is a skeptical warm-started bandit that treats historical evidence and causal structure as hypotheses to test online.

We propose *Causal Skeptic Bandits* (CSB). CSB represents alternative interpretations of the historical logs as candidate worlds, including a naive logged-data model, a backdoor-adjusted causal model, graph-selective variants, and a cold-start fallback. Each world predicts rewards and receives trust only to the extent that it predicts online interventional observations. CSB also adds *diagnostic falsification*: actions are valuable not only when they have high reward or high statistical uncertainty, but also when their outcomes distinguish competing assumptions. CDV further turns diagnostic evidence into a policy rule: rescue causal transfer when the adjusted world is supported, and veto it when online evidence falsifies the graph.

We therefore position CSB as regime-uncertainty control for unreliable historical evidence, not as another bandit leaderboard. If the graph is known correct, pure causal learning can be best; if the graph is known wrong, cold start can be safest. The deployment problem is regime uncertainty: before interaction, the learner may not know whether logs are clean, confounded but adjustable, or generated under a wrong graph. We show that CDV occupies the broad intermediate region between causal and cold-start endpoint oracles in the evaluated deployment mixture. We separately test changes *during* deployment and find that the current cumulative trust rule is path dependent; static regime uncertainty and online nonstationarity must not be conflated.

Contributions. We formulate unreliable historical transfer as online model selection over testable interpretations, couple it to decision-targeted diagnostics and a causal-transfer veto, and separate diagnostic value, fallback value, and their costs using matched ablations and stress tests. Diagnostics help only when an available action separates decision-relevant worlds; trust need not recover after a deployment shift.

Method

Candidate worlds and trust. At time t, the learner observes context x_t, chooses arm $a_t \in \{1, \dots, K\}$, and receives reward r_t. It also has historical data $\mathcal{D}_0$ and possibly unreliable causal assumptions. CSB builds candidate worlds $\mathcal{M}$, where each world m interprets $\mathcal{D}_0$ differently and returns a predicted reward $\hat{r}_{m,t}(a)$ and ridge uncertainty radius $b_{m,t}(a)$. Trust weights are updated from online prediction error,

$$w_{m,t} \propto \rho_m \exp(-\eta L_{m,t}) + \epsilon,$$
$$L_{m,t+1} = \gamma L_{m,t} + \ell(r_t, \hat{r}_{m,t}(a_t)).$$

Historical fit alone cannot earn trust; a world must survive online interventional tests.

Diagnostic action selection. CSB scores each arm by

$$S_t(a) = \sum_m w_{m,t}\hat{r}_{m,t}(a) + \alpha\Big(\sum_m w_{m,t} b_{m,t}(a)^2\Big)^{1/2} + \lambda_t D_t(a).$$

The first two terms are reward and UCB value. The diagnostic term $D_t(a)$ measures whether pulling a can distinguish worlds. In CDV we use a decision-rescue target between the naive and adjusted worlds,

$$D_t(a) = \sqrt{w_{N,t} w_{C,t}}\,[\hat{r}_{C,t}(a) - \hat{r}_{N,t}(a)]_+ Q_t(a),$$

where $Q_t(a)$ downweights arms far from the decision boundary. The principal entropy gate decays as trust concentrates; optional no-harm gates also suppress high-opportunity-cost diagnostics.

Causal diagnostic veto. Diagnostic evidence is useful only if it changes transfer decisions. CDV uses the adjusted world for causal rescue when its trust is high, but blends toward a cold-start fallback when the adjusted world is falsified. Thus the same online evidence can both rescue useful causal transfer and veto unsafe graph transfer. In the principal setting, the fallback blend activates only when adjusted-world weight is at most 0.30 and cold-world weight is at least 0.10; CDV-NC omits this blend. Algorithm 1 summarizes the resulting loop.

Why falsification helps. If two worlds agree on the exploitative arm but differ by gap Δ on a probe arm, assume its reward is σ^2-sub-Gaussian under either world. A midpoint empirical-mean test after n probe pulls has error at most $2\exp[-n\Delta^2/(8\sigma^2)]$. Thus

$$n \geq \frac{8\sigma^2}{\Delta^2} \log \frac{2}{\delta}$$

suffices to reject the wrong prediction with error at most δ. This is an identification cost, not a full bandit-regret theorem: if a non-diagnostic policy never selects the probe, its observed history can leave the worlds indistinguishable.

Regime uncertainty. Let p be the probability that the graph is wrong. For any fixed policy, mixture regret is affine:

$$R_\pi(p) = pR_{\pi,\text{wrong}} + (1-p)\bar{R}_{\pi,\text{valid}}.$$

CDV should win neither endpoint. It is useful where wrong-graph excess risk of transfer-heavy policies outweighs CDV's valid-regime opportunity cost, but before the opportunity cost of cold start is outweighed by its safety.

Algorithm 1 Causal Skeptic Bandit with diagnostic veto

1: Fit candidate worlds $\mathcal{M}$ from $\mathcal{D}_0$; initialize $L_{m,1} = 0$ and trust priors ρ_m.
2: **for** $t = 1, \dots, T$ **do**
3: Observe x_t and obtain $\hat{r}_{m,t}(a), b_{m,t}(a)$ for all m, a.
4: Normalize $w_{m,t} \propto \rho_m \exp(-\eta L_{m,t}) + \epsilon$.
5: Apply trust-triggered rescue or veto blends, then add gated decision-relevant scores $D_t(a)$.
6: Choose $a_t \in \arg\max_a S_t(a)$ and observe interventional reward r_t.
7: Update every world's loss and predictor on the same observation: $L_{m,t+1} = \gamma L_{m,t} + \ell(r_t, \hat{r}_{m,t}(a_t))$.
8: **end for**

This interval has a direct characterization. Let A, D, C denote causal, CDV, and cold policies. Suppose A is better in valid regimes by $c_v = R_{D,v} - R_{A,v} > 0$, while D is better under graph failure by $c_w = R_{A,w} - R_{D,w} > 0$. Then D beats A exactly when

$$p > p_{DA} = \frac{c_v}{c_v + c_w}.$$

Likewise, if D beats cold start in valid regimes by $d_v = R_{C,v} - R_{D,v} > 0$, but cold start wins under failure by $d_w = R_{D,w} - R_{C,w} > 0$, then D beats C exactly when

$$p < p_{DC} = \frac{d_v}{d_v + d_w}.$$

Both statements follow by expanding the affine mixture regret. A nonempty CDV interval exists iff $p_{DA} < p_{DC}$; the empirical mixture curve estimates these thresholds rather than assuming them.

Experiments

We evaluate synthetic structural bandits, a healthcare-style simulator, and ten public IHDP semi-synthetic replications (Hill 2011; Louizos et al. 2017). Historical logs span clean, confounded, diagnostic-confounded, and wrong-graph settings; deployment may be misspecified or switching. Baselines include naive and causal warm starts, cold start, loss-weighted ensembles, no-diagnostic CSB, target-aware diagnostic ensembles (LWD+Diag), and CDV variants. Principal stationary runs use $T = 18000$ and 1000–2700 paired seeds; sensitivity cells use 1500. Stress tests add 5376 misspecification, 5760 diagnostic-cost, 8960 IHDP, and 4608 unannounced-switch runs; IHDP uncertainty is clustered over ten public replications. All $\pm$ and HW values are 95% confidence half-widths; the supplement gives full protocols, seed ranges, and negative results.

Results

Diagnostic falsification buys information. Table 1 and Figure 1 show that no-diagnostic CSB has regret 152.91 ± 5.59 in a diagnostic-confounding setting, while CDV diagnostic exploration reduces regret to 26.53 ± 0.39. This

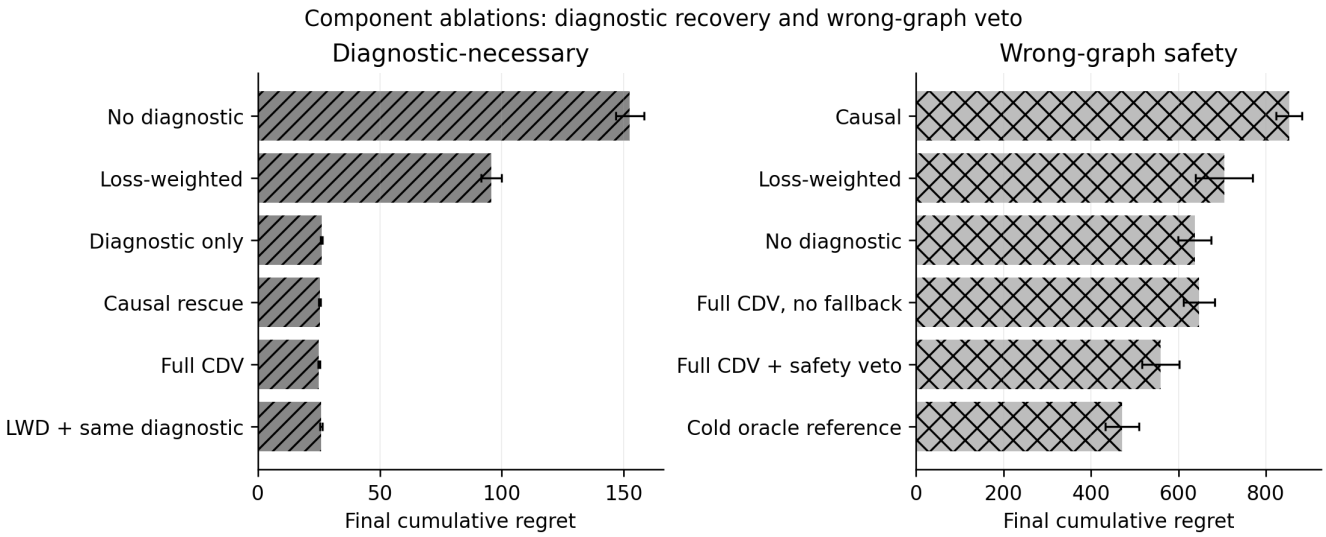

Figure 1: Two 500-seed, $T = 18000$ component audits. Left: the controlled diagnostic-necessary benchmark. Right: the healthcare wrong-graph benchmark. They complement, rather than duplicate, the matched cells in Table 1.

Method	Diagnostic	Wrong graph
Pure causal	26.64 ± 0.41	1585.42 ± 43.00
Loss-weighted	106.58 ± 9.20	126.05 ± 5.85
No diagnostic CSB	152.91 ± 5.59	124.11 ± 5.56
CDV diagnostic	26.53 ± 0.39	–
CDV + veto	–	127.59 ± 7.50

Table 1: Matched 500-seed, $T = 18000$ mechanism cells. Diagnostic falsification is needed in the diagnostic setting; vetoed transfer is needed when the graph is wrong.

isolates the role of choosing interventions that test assumptions, not merely maintaining multiple models. In spurious-diagnostic controls with 1000 seeds, causal-rescue diagnostics reduce regret from 215.23 for no-diagnostic CSB to 58.92. No-harm gates reduce unnecessary diagnostic overhead in clean/confounded/mixed regimes while preserving most diagnostic-needed gains.

Public semi-synthetic evidence. The public IHDP/CEVAE benchmark supplies real covariates and standardized potential outcomes. Its original outcome construction is policy-degenerate: treatment 1 is optimal for 93.5% of held-out units. We retain it as a negative control and report a fixed policy-stress transformation that centers the treatment effect using only the historical split; the optimal treatment rate is then 49.4%. Each cell contains 32 paired splits in each of ten public replications, and uncertainty is clustered by replication. Table 2 shows that CDV is competitive with warm-start methods and substantially better than cold start, but it is statistically tied with no-diagnostic CSB. With two actions and no dedicated probe, IHDP validates robust warm starting but does not isolate a diagnostic-action advantage.

Method	Full adjust.	Partial adjust.
CDV-Veto	122.42	122.33
No diagnostic	121.95	122.04
LWD+Diag	127.30	128.28
Pure causal	135.10	136.71
Cold start	302.85	302.85

Table 2: Mean regret on the balanced IHDP policy stress test (10 replications, 32 splits each, $T = 3000$). CDV–cold paired gains are 180.43 and 180.52; replication-clustered 95% intervals exclude zero. CDV–no-diagnostic intervals include zero.

Veto prevents wrong-graph transfer. Pure causal transfer fails under wrong graphs with regret 1585.42 ± 43.00. CDV-Veto reduces this to 127.59 ± 7.50, comparable to non-causal fallback baselines. Implementation audits show that the exact trust loss is not the main source of the gain; the stable mechanism is the inclusion of diagnostic evidence and a cold fallback world.

Healthcare warm starts. Figure 2 and Table 3 show that, in the diagnostic healthcare regime, CDV-NC has regret 17.48 ± 1.18, improving over LWD+Diag (20.46 ± 1.04) and no-diagnostic CSB (44.27 ± 1.49); paired improvements are 2.99 ± 1.52 and 26.79 ± 1.86, respectively. In the wrong-graph regime, CDV-Veto reduces pure causal regret from 854.29 ± 12.00 to 547.81 ± 17.17, and improves over loss-weighted and no-diagnostic warm starts. Cold start remains best when wrong-graph failure is known in advance.

Misspecification and priced diagnostics. Across in-set, unseen-linear, and unseen-nonlinear wrong-graph deployments, CDV is much safer than pure causal transfer (128 paired seeds), with mean gains 215.04, 505.65, and 345.87.

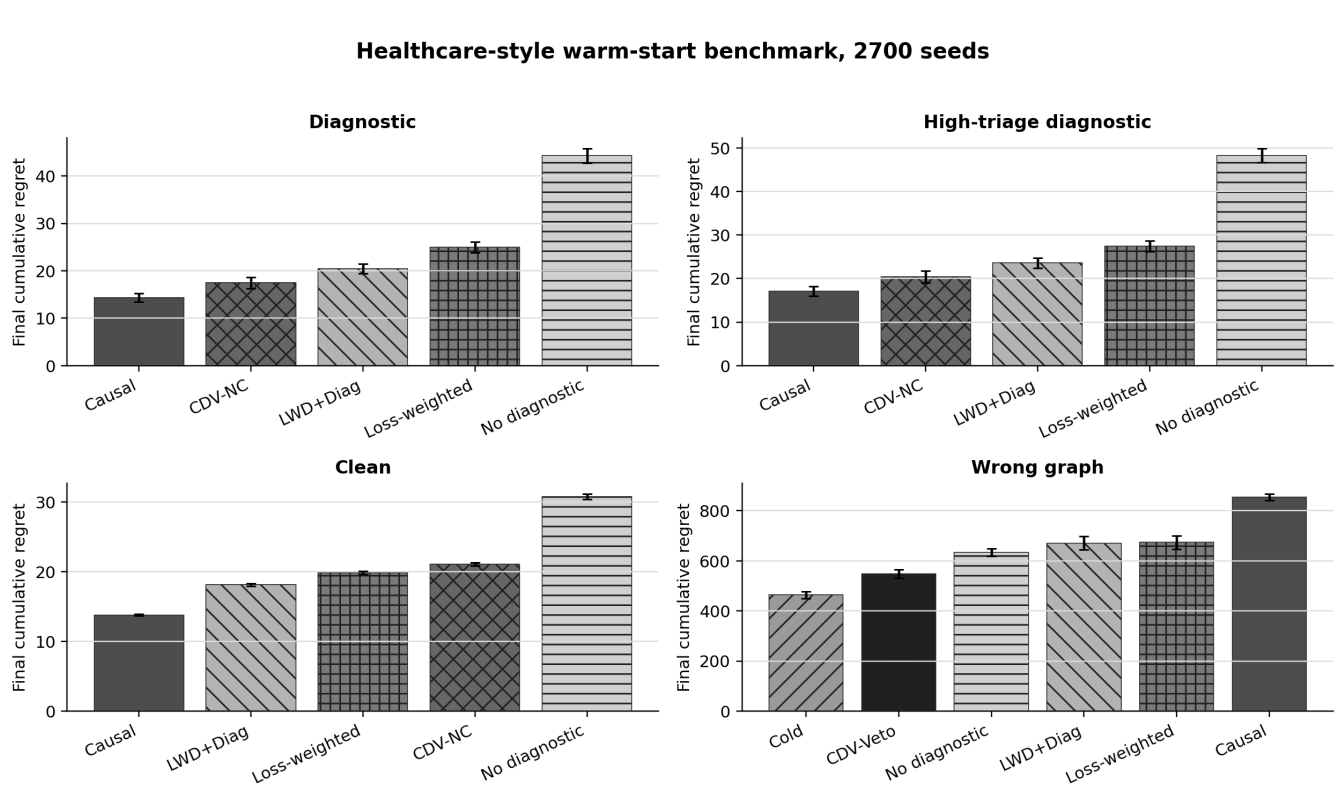

Figure 2: Healthcare-style warm-start benchmark with 2700 seeds. CDV-NC improves on composite and no-diagnostic warm starts when causal rescue is useful; CDV-Veto reduces wrong-graph transfer. Causal and cold start remain endpoint oracles.

Cold start nevertheless wins the in-set and nonlinear endpoints. In a separate abrupt-drift condition, CDV does not beat pure causal (647.99 versus 632.02). Pricing probe actions from 0 to 1 causes CDV's probe rate to fall from 0.760 to 0.317 in the correctly specified diagnostic world and from 0.063 to 0.003 under an in-set wrong graph. This is a cost–robustness curve, not evidence of uniform dominance.

Unannounced regime switches: a negative result. We switch valid-to-wrong and wrong-to-valid mechanisms after 25%, 50%, or 75% of $T = 9000$, using 128 paired seeds per cell. The current $\gamma = 1$ cumulative-loss rule does not reliably reverse trust: after valid-to-wrong switches CDV still assigns 0.930–0.953 mean final weight to the adjusted world, whereas after wrong-to-valid switches its final adjusted-world weight is only 0.172–0.221. CDV can still reduce total regret relative to an endpoint baseline in several cells because its online predictors continue updating, but it does not significantly beat no-diagnostic skepticism consistently. An independently seeded holdout shows that fixed forgetting $\gamma = 0.999$ helps an early wrong-to-valid switch by 840.56 (95% CI [469.61, 1211.51]) yet significantly harms all three valid-to-wrong switches. Thus a single discount is not a symmetric nonstationarity solution. A final windowed detector fires after 94–100% of valid-to-wrong changes, yet reset variants increase regret; detecting a shift does not make a global trust reset safe.

CDV wins the uncertain middle. Figure 3 evaluates a deployment mixture over clean, diagnostic-confounded, high-triage diagnostic, and wrong-graph regimes. Pure causal learning is best at $p_{\text{wrong}} = 0$ and 0.03. CDV-Veto becomes best around $p_{\text{wrong}} = 0.05$ and remains best through $p_{\text{wrong}} = 0.775$; cold start takes over at 0.80. At $p_{\text{wrong}} = 0.50$, CDV-Veto has expected regret 286.35, compared with 337.34 for no-diagnostic CSB, 345.80 for LWD+Diag, 349.05 for loss-weighted warm starts, 403.42 for cold start, and 434.68 for causal learning. At $p_{\text{wrong}} = 0.75$, CDV-Veto remains best at 417.08, ahead of cold start at 433.48. A 1500-seed sensitivity grid confirms that CDV-NC beats both LWD+Diag and no-diagnostic CSB in all 24 diagnostic cells, while CDV-Veto beats the best warm-start baseline in 18/24 wrong-graph cells; all losses occur at the lowest hidden-risk level.

Mechanism and Robustness Controls

The headline comparisons are useful only if they separate diagnostic value from extra exploration, and safety from indiscriminate conservatism. We therefore organize the controls around three questions: whether the diagnostic target is specific, whether the diagnostic cost is bounded when no diagnosis is needed, and whether the result depends on one trust-loss implementation.

Diagnostic specificity and no-harm calibration. In the spurious-diagnostic control, historically attractive evidence

Regime	Method	Regret	95% HW
Diagnostic	Causal	14.32	0.92
Diagnostic	CDV-NC	17.48	1.18
Diagnostic	LWD+Diag	20.46	1.04
Diagnostic	No diagnostic	44.27	1.49
Clean	Causal	13.79	0.09
Clean	LWD+Diag	18.10	0.22
Clean	CDV-NC	21.04	0.23
Wrong graph	Cold	463.54	14.51
Wrong graph	CDV-Veto	547.81	17.17
Wrong graph	No diagnostic	633.59	14.67
Wrong graph	Loss-weighted	674.04	26.99
Wrong graph	Causal	854.29	12.00

Table 3: Healthcare summary. CDV is strongest as a compromise: pure causal learning is best when the graph is correct, and cold start is safest when the graph is known wrong.

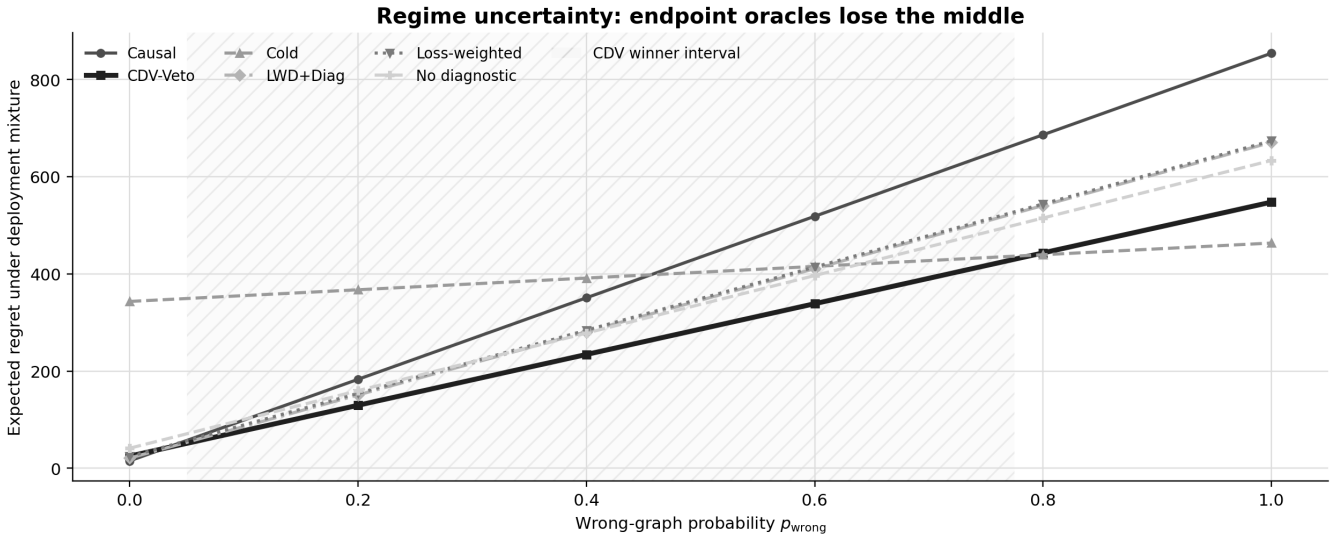

Figure 3: Regime uncertainty in the evaluated deployment mixture. Causal learning wins when graph failure is very unlikely, cold start wins when it is almost certain, and CDV-Veto minimizes expected regret in between.

points to the wrong action and the useful online probe is not selected by ordinary exploitation. Causal-rescue targeting lowers regret from 215.23 for no-diagnostic CSB to 58.92, while all-disagreement targeting gives 56.92. Pure causal learning reaches 44.92 because this control assumes a valid graph; it is an endpoint oracle rather than a policy for unknown graph reliability. This ordering matters: it shows that diagnostic actions repair a particular information failure, not that they dominate a learner that is told the correct regime.

Diagnostics can also waste reward when candidate worlds are already decision-equivalent. Table 4 reports a fresh 1200-seed calibration. Unconstrained entropy probing obtains the lowest diagnostic regret but adds 3.05 regret on average across clean, confounded, and mixed non-diagnostic regimes. A low-threshold gate cuts this overhead to 1.00 but gives up part of the diagnostic gain; a high-threshold gate nearly preserves the gain but not the overhead. The gate is therefore an explicit operating point rather than a universally optimal constant.

Method	Diagnostic	Mean ovhd.	Wrong graph
Entropy diag.	26.12	3.05	127.71
No diagnostic	149.25	0.00	125.85
Loss-weighted	95.20	−2.69	127.07
Decay gate	33.15	1.52	126.10
Low gate	55.81	1.00	126.92
High gate	26.74	3.51	128.91

Table 4: Fresh no-harm calibration with 1200 seeds. Overhead is measured relative to no-diagnostic CSB over clean, confounded, and mixed regimes.

Sensitivity and paired uncertainty. Table 5 summarizes a healthcare grid varying triage strength and hidden risk over 24 cells per reliability condition, using 1500 paired seeds per cell. CDV-NC beats LWD+Diag in every diagnostic cell, with mean gain 3.41 and cellwise gains from 0.99 to 6.46. In wrong-graph conditions, CDV-Veto beats the best warm-

Paired comparison	Gain	95% HW	Cells
CDV-NC vs. LWD+Diag, diagnostic	3.41	0.41	24/24
CDV-NC vs. no diagnostic	25.61	0.50	24/24
CDV-Veto vs. pure causal	274.55	7.55	24/24
CDV-Veto vs. no diagnostic	118.26	8.93	24/24
CDV-Veto vs. cold start	−118.01	7.97	0/24

Table 5: Paired all-cell healthcare sensitivity results. Positive gain means lower regret for the named CDV variant.

start comparator in 18 of 24 cells and beats pure causal transfer in all 24. The six warm-start losses occur at the lowest hidden-risk level, where graph error causes less damage and aggressive reuse can be slightly cheaper. Against cold start, CDV-Veto has a paired disadvantage of 118.01 ± 7.97 in these known-wrong cells. We retain this negative result because it identifies the boundary between adaptive deployment and an oracle that knows in advance to discard the history.

Implementation audit. The mechanism does not reduce to one robust-loss formula. Removing the cold world raises wrong-graph regret to 543.10, whereas removing the graph-selective world gives 175.32. Conversely, simplified no-Huber and no-discount trust updates outperform the default update in selected confounded or wrong-graph checks. These results rule out the stronger claim that the precise trust loss is theoretically privileged. What survives across audits is the structural combination: online interventions test candidate interpretations, diagnostic evidence changes which interpretation controls the action, and an explicit cold fallback limits negative transfer when the adjusted world is falsified.

Reproducibility protocol. Synthetic intervals use independently seeded environment draws, with method comparisons paired wherever the same draw is available; IHDP uncertainty is clustered by public replication. The code-and-data package contains the candidate-world implementation, complete experiment runners, unit tests, analysis scripts, and compact aggregate tables used to regenerate the paper figures and appendix tables. The supplement records the seed accounting for the staged campaign, including the 1000-seed diagnostic controls, 2700-seed healthcare summaries, 1500-seed sensitivity cells, 1200-seed out-of-sample defenses, and the independent 64-seed replication. Large per-round scheduler traces are not necessary to recompute the reported aggregate comparisons and are excluded from the anonymous upload.

Decision scope and computational cost. CSB adds computation across candidate worlds, not across an unbounded graph space. With $M = |\mathcal{M}|$ fitted worlds and K arms, prediction and diagnostic scoring require $O(MK)$ world–arm evaluations per round; the trust update is $O(M)$ after observing the selected reward. The experiments use a small fixed world set, so this overhead is minor relative to simulation and seed replication. The method does not perform posterior inference over all causal graphs, and the reported runtime should not be interpreted as evidence that such inference is free. Candidate construction remains a modeling step and is a central target for future work.

Threat model and falsifiable non-claims. The deployment threat is a warm start whose reliability is unknown before online interaction. It includes confounded logging policies, an adjustment model that may be correct or wrong, and diagnostic actions whose short-term cost must be traded against future transfer risk. It does not include an adversary that changes the graph after every action, arbitrary unmeasured nonstationarity, or a clinician-facing safety guarantee. The healthcare-style environment is a structured stress test, not clinical evidence.

Several observations would falsify the intended mechanism claim. First, if a non-diagnostic learner matched CDV in the diagnostic-confounding controls, the benefit could not be attributed to active assumption testing. Second, if CDV without a cold fallback remained safe across wrong-graph replications, the veto would be unnecessary. Third, if sensitivity gains appeared only in one hand-picked cell or vanished under paired seeds, the regime-uncertainty story would be unsupported. The experiments show the opposite pattern while also retaining the inconvenient endpoint results: causal learning wins when graph validity is known, and cold start wins when graph failure is known. Table 6 summarizes these claim boundaries.

Interpretation of trust. The weights are predictive reliability scores over a finite hypothesis set, not calibrated probabilities that a causal graph is true. A world earns relative influence by predicting online interventional rewards, and it can lose influence when those predictions fail in a stationary deployment. Under unannounced switches, accumulated evidence can make this influence difficult to reverse. This distinction prevents a semantic overreach: CDV tests whether an interpretation is useful for the observed decision process, not whether it has identified the unique data-generating causal structure. Likewise, diagnostic success means that an action separates candidate predictions sufficiently to improve control; it does not by itself establish causal discovery.

Failure analysis. The remaining errors separate into information, model-set, and policy failures. An information failure occurs when available actions do not separate candidate worlds quickly enough at tolerable cost. A model-set failure occurs when no candidate captures the online regime, so relative trust cannot recover the missing interpretation. A policy failure occurs when the learner has evidence against transfer but lacks or delays a safe fallback. CDV directly addresses the third failure and reduces the first through targeted probes; it cannot guarantee recovery from an incomplete world set. This taxonomy clarifies why automatic candidate construction is a substantive open problem rather than an implementation detail.

Limitations and Related Work

The evidence supports a mechanism-level claim, not a leaderboard claim. CDV does not beat oracles that know the reliability regime: causal learning is best when the graph is

Claim	Required control	Observed boundary
World testing helps	No-diagnostic CSB	Gain appears only when the probe is informative.
Diagnosis is specific	Spurious and all-disagreement targets	Broad probing can add avoidable regret.
Veto improves safety	No-fallback and wrong-graph variants	Removing the cold world is unsafe.
Reuse is worthwhile	Clean and valid-graph endpoints	Pure causal transfer can remain best.
Static regime hedge	Cold-start endpoint	Cold start remains best when failure is known.
Mixture gain is stable	Paired grid and fresh seeds	CDV wins a broad interval, not every cell.
Switch adaptation	Unannounced two-way switches	Cumulative trust is path dependent.

Table 6: Claim-to-control audit. Each positive claim is paired with a result that limits its scope.

correct, and cold start is safest when the graph is known wrong. Target-aware ensembles can also be close once given the same diagnostic target. We therefore view diagnostic falsification as the transferable idea and CDV as a conservative causal-transfer policy around that evidence. The experiments are synthetic and healthcare-style rather than a real clinical deployment, and candidate worlds and veto thresholds are designer-chosen. The fresh defense replication also shows that a no-fallback CDV variant can be unsafe under healthcare wrong-graph shift; robust deployment requires an explicit fallback world rather than diagnostic optimism alone. The public IHDP result does not isolate a diagnostic advantage. Unlike nonstationary causal bandits that posit a temporal causal model (Kwon et al. 2025), our abrupt-switch audit studies trust recovery and shows that it can fail.

Logged-bandit estimators address a complementary uncertainty: how to evaluate or optimize a new policy from actions selected by an earlier logging policy (Dudík, Langford, and Li 2011; Bottou et al. 2013; Swaminathan and Joachims 2015; Joachims, Swaminathan, and Schnabel 2017; Saito, Ren, and Joachims 2023). Propensity correction and doubly robust estimation can reduce statistical bias when overlap and logging assumptions are adequate, but they do not decide whether a proposed causal adjustment graph should control future interventions. CSB instead keeps several interpretations of the same history active and uses online rewards to test their predictive consequences. The two directions are compatible: stronger off-policy estimators could be used inside individual worlds without removing the need to decide which world is reliable.

Related robustness work addresses corrupted feedback (Lykouris, Sridharan, and Tardos 2018), contextual reward-model misspecification via safe-policy reversion (Krishnamurthy, Hadad, and Athey 2021), and model selection (Foster, Krishnamurthy, and Luo 2019). CSB targets a different structured failure: a wrong graph can yield coherent but harmful transfer, whereas a confounded log may remain useful after correct adjustment. It uses selected interventions to test alternative causal readings and an explicit cold fallback. Transportable bandits instead exploit assumed causal invariances across environments and provide regret guarantees (Bellot, Malek, and Chiappa 2023); here those relations are hypotheses that may be false.

The candidate-world view also connects to invariant prediction. Invariant causal methods seek relationships that remain stable across environments (Peters, Bühlmann, and Meinshausen 2016; Arjovsky et al. 2019), whereas CSB receives only an online interaction stream and a finite set of proposed historical interpretations. It does not infer invariance across arbitrary environments. Instead, it asks whether the predictions implied by an interpretation survive the interventions that matter for the current action decision. This narrower goal is why diagnostic arm selection is central: passive observations on arms where worlds agree cannot identify which interpretation should be trusted.

The closest algorithmic precedent robustly combines supervised and bandit feedback under source mismatch, with no-regret guarantees (Zhang et al. 2019). CSB instead treats alternative causal readings of the same log as testable worlds, selects interventions to discriminate them, and couples the evidence to a veto and fallback. Unlike that prior work, our present guarantee is only the simplified probe-identification bound, not a general regret theorem. Classical transfer learning also recognizes that source knowledge can harm a target (Rosenstein et al. 2005; Wang, Zhang, and Chaudhuri 2022); CDV's contribution is the decision-targeted falsification rule, not the observation that transfer can fail.

More broadly, we build on contextual and causal bandits (Li et al. 2010; Chu et al. 2011; Lattimore, Lattimore, and Reid 2016; Lee and Bareinboim 2018; Malek, Aglietti, and Chiappa 2023; Liu, Attias, and Roy 2024; Jamshidi, Etesami, and Kiyavash 2024) and expert aggregation (Herbster and Warmuth 1998; Cesa-Bianchi and Lugosi 2006; Foster, Krishnamurthy, and Luo 2019). CSB differs by actively testing which historical interpretation should control online actions.

Generative-AI use. Generative AI systems assisted with research ideation, manuscript drafting and editing, code generation, experiment orchestration, and result organization. Human authors remain responsible for verifying all submitted content.

Conclusion

Warm-started bandits need to test historical evidence before letting it control decisions. CSB represents alternative interpretations as candidate worlds; CDV uses targeted interventions to support or veto causal transfer. The method reduces risk across unknown but stationary reliability regimes, while public IHDP and switching tests delimit the claim: it is neither an endpoint oracle nor a solved detector of deployment-time change.

In evolving deployments, pair it with monitoring and a validated change-management rule; cumulative trust is not automatically reversible.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. arXiv:1907.02893.

Bellot, A.; Malek, A.; and Chiappa, S. 2023. Transportability for Bandits with Data from Different Environments. In *Advances in Neural Information Processing Systems*, volume 36.

Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 31, 320–326.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual Bandits with Linear Payoff Functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, 1097–1104.

Foster, D. J.; Krishnamurthy, A.; and Luo, H. 2019. Model Selection for Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 32.

Herbster, M.; and Warmuth, M. K. 1998. Tracking the Best Expert. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 286–294.

Hill, J. L. 2011. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Jamshidi, F.; Etesami, J.; and Kiyavash, N. 2024. Confounded Budgeted Causal Bandits. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, 423–461.

Joachims, T.; Swaminathan, A.; and Schnabel, T. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 781–789.

Krishnamurthy, S. K.; Hadad, V.; and Athey, S. 2021. Adapting to Misspecification in Contextual Bandits with Offline Regression Oracles. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5805–5814.

Kwon, Y.; Choe, Y.; Park, S.; Dhir, N.; and Lee, S. 2025. Non-Stationary Structural Causal Bandits. In *Advances in Neural Information Processing Systems*, volume 38.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal Bandits: Learning Good Interventions via Causal Inference. In *Advances in Neural Information Processing Systems*, volume 29.

Lee, S.; and Bareinboim, E. 2018. Structural Causal Bandits: Where to Intervene? In *Advances in Neural Information Processing Systems*, volume 31.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 661–670.

Liu, Z.; Attias, I.; and Roy, D. M. 2024. Causal Bandits: The Pareto Optimal Frontier of Adaptivity, a Reduction to Linear Bandits, and Limitations around Unknown Marginals. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 30647–30668.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems*, volume 30.

Lykouris, T.; Sridharan, K.; and Tardos, É. 2018. Stochastic Bandits Robust to Adversarial Corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 114–122.

Malek, A.; Aglietti, V.; and Chiappa, S. 2023. Additive Causal Bandits with Unknown Graph. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 23574–23589.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society: Series B*, 78(5): 947–1012.

Rosenstein, M. T.; Marx, Z.; Kaelbling, L. P.; and Dietterich, T. G. 2005. To Transfer or Not to Transfer. In *NIPS Workshop on Transfer Learning*.

Rubin, D. B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5): 688–701.

Saito, Y.; Ren, Q.; and Joachims, T. 2023. Off-Policy Evaluation for Large Action Spaces via Conjunct Effect Modeling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 29734–29759.

Swaminathan, A.; and Joachims, T. 2015. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, 814–823.

Wang, Z.; Zhang, C.; and Chaudhuri, K. 2022. Thompson Sampling for Robust Transfer in Multi-Task Bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 23363–23416.

Zhang, C.; Agarwal, A.; Daumé III, H.; Langford, J.; and Negahban, S. 2019. Warm-Starting Contextual Bandits: Robustly Combining Supervised and Bandit Feedback. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 7335–7344.

Supplementary Material for Testing Before Trusting: Causal Skeptic Bandits under Unreliable Historical Evidence

Anonymous submission

Claim	Supporting test	Boundary test
Diagnostic value	Probe-necessary controls	Two-action IHDP tie with no-diagnostic CSB
Wrong-graph robustness	Veto and fallback ablations	Cold start wins known-wrong endpoints
Unknown-regime hedge	Predeployment mixture curve	Endpoint policies win near mixture extremes
World-set robustness	Shift stress tests	No guarantee outside the world set
Online change	Two-way switches	Path-dependent trust

Table 1: Evidence-to-claim map.

Scope of the Supplement

This supplement gives the algorithmic specification, evaluation protocol, additional tables, public-data construction, misspecification tests, and negative nonstationarity results for the anonymous submission. It deliberately distinguishes three claims:

1. diagnostic actions can be necessary when ordinary reward-seeking actions do not distinguish candidate interpretations;
2. a cold fallback limits negative transfer when an adjusted world is wrong; and
3. neither mechanism makes the current method an endpoint oracle or a general change detector.

Algorithmic Details

Candidate Worlds

All candidate worlds receive the same historical batch and the same online interactions, but interpret the batch differently.

- **Naive warm start**: ridge reward models fit logged action–reward associations on the deployment context directly.
- **Backdoor-adjusted world**: ridge models fit the specified adjustment variables in the log and, at decision time, marginalize them using their online-population mean.
- **Graph-selective world**: reward models use graph-selected feature subsets.
- **Cold fallback**: reward parameters begin from online-only state and therefore do not inherit historical reward bias.
- **Diagnostic bait world**: used only in controlled probe-necessary experiments to create a historically plausible but interventionally falsifiable interpretation.

Candidate worlds are predictive hypotheses, not posterior samples over causal graphs. Their normalized weights are not calibrated graph probabilities.

Trust, Action Score, and Veto

For robust loss

$$\ell_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta(|e| - \frac{1}{2}\delta), & |e| > \delta, \end{cases}$$

the implementation uses

$$L_{m,t+1} = \gamma L_{m,t} + \ell_\delta(r_t - \widehat{r}_{m,t}(a_t)),$$
$$w_{m,t} \propto \rho_m e^{-\eta L_{m,t}} + \epsilon.$$

Writing $b_{m,t}(a)$ for the unscaled per-world ridge uncertainty radius, the action score is

$$S_t(a) = \sum_m w_{m,t}\widehat{r}_{m,t}(a) + \alpha\sqrt{\sum_m w_{m,t} b_{m,t}(a)^2} + \lambda_t D_t(a).$$

For the decision-rescue target, let

$$q_t(a) = \sum_m w_{m,t}\,\mathbf{1}\left[a = \arg\max_b \widehat{r}_{m,t}(b)\right],$$
$$u_t = 1 - \max_a q_t(a), \qquad g_t(a) = \max_b v_t(b) - v_t(a),$$

where $v_t(a)$ is the reward-plus-UCB exploitation value after any rescue or fallback blend. The boundary factor used in the main paper is

$$Q_t(a) = \frac{(0.05 + u_t)(0.05 + q_t(a))}{0.10 + \max\{g_t(a), 0\}}.$$

The decision-rescue diagnostic is based on positive adjusted-versus-naive reward disagreement and is downweighted away from the current decision boundary. Entropy and no-harm gates turn the diagnostic bonus off as trust concentrates or when the diagnostic arm has excessive estimated opportunity cost. CDV-NC uses supported adjusted-world influence for causal rescue. CDV-Veto additionally blends toward the cold world when adjusted-world weight is at most the causal threshold and cold-world weight is at least the fallback threshold.

Parameter	Value	Parameter	Value
Ridge regularization	1.0	UCB scale	0.8
Trust learning rate η	0.08	Trust discount γ	1.0
Huber threshold δ	1.0	Diagnostic warm-up	800
CDV diagnostic scale	8.0	Rescue trust thresh.	0.80
Fallback causal thresh.	0.30	Fallback cold thresh.	0.10

Table 2: Principal healthcare and stress-test settings. Dedicated diagnostic-control sweeps vary the diagnostic scale and gates as stated in their metadata.

Principal Hyperparameters

Probe-Identification Derivation and Scope

The simplified identification statement in the main paper follows from a fixed-sample sub-Gaussian test. Suppose two candidate worlds agree on the ordinary exploitative arm but predict probe-arm means μ_0 and μ_1, with $\Delta = |\mu_1 - \mu_0| > 0$. After n probe pulls, classify using the midpoint $(\mu_0 + \mu_1)/2$. If probe noise is σ^2-sub-Gaussian, then under either world,

$$\Pr\left(|\bar{r}_n - \mu_j| \ge \frac{\Delta}{2}\right) \le 2\exp\left(-\frac{n\Delta^2}{8\sigma^2}\right).$$

Consequently,

$$n \ge \frac{8\sigma^2}{\Delta^2}\log\frac{2}{\delta}$$

suffices for error at most δ. If the one-step opportunity cost of the probe is bounded by c, the corresponding fixed-sample identification cost is at most $c\left\lceil 8\sigma^2\Delta^{-2}\log(2/\delta)\right\rceil$.

This calculation explains why an explicitly diagnostic action can matter: a policy that never selects the probe receives no samples governed by Δ, so ordinary low regret before identification does not imply that the worlds become distinguishable. It is not a regret theorem for the full algorithm. In particular, it assumes a stationary two-world probe model and does not cover context-dependent gaps, fitted-world error, multiple testing, adaptive change points, or the opportunity cost induced by the no-harm gate. Those effects are evaluated empirically below.

Evaluation and Statistical Protocol

Regret is computed against the context-specific action with maximum expected interventional reward. The realized reward observed by the learner is never used as the oracle comparator. All algorithms in a paired cell receive the same environment seed and historical draw. Reported $\pm$ values in the main stationary tables are 95% confidence half-widths over independent seeds. Paired comparisons use per-seed regret differences. For IHDP, split-level results are first averaged within each of the ten public replications and uncertainty is computed across replication means; the 320 splits are not treated as 320 independent datasets.

Hyperparameter sweeps and confirmatory runs use disjoint seed ranges where a setting is selected. The fixed-forgetting study, for example, selects $\gamma \in \{1, .999\}$ on exploratory seeds and tests only those two values on new seeds 140000–140063. Exploratory change-detection attempts are

not used as positive evidence unless they pass a new-seed holdout.

The new CPU experiments used Ubuntu 22.04 on a dual-socket AMD EPYC 7282 host with 64 logical CPUs and 251 GiB RAM. Each process used one BLAS/OpenMP thread; at most 48 worker processes were launched. The simulations did not use GPUs. The anonymous artifact records Python/package versions, seeds, scenario parameters, row counts, and dataset hashes.

Controlled Scenario Definitions

Clean logs share the deployment reward mechanism and contain no historical treatment confounding. They verify that robustness machinery does not create an apparent gain by silently corrupting an easy causal endpoint. **Confounded logs** use a historical action policy that depends on a common cause of treatment and reward; adjustment is correctly specified. **Mixed logs** combine historical subpopulations with different logging bias while retaining a valid adjustment interpretation.

Diagnostic-necessary regimes are constructed so that candidate worlds agree on the action preferred by ordinary exploitation but disagree on a lower-immediate-reward probe. Selecting only reward-maximizing arms can therefore leave the harmful interpretation unfalsified. **Spurious-diagnostic regimes** reverse this logic: candidate predictions disagree on an arm whose acquisition is not worth its opportunity cost. They test whether gating suppresses diagnosis when disagreement is decision irrelevant.

Wrong-graph regimes deliberately provide an invalid adjustment structure while retaining historically plausible fitted predictions. **Candidate-misspecified regimes** additionally change deployment in a way not represented by any fitted world. Finally, **switch regimes** change the reward mechanism after interaction begins and withhold the change time and direction. They are evaluated separately because predeployment regime uncertainty and temporal nonstationarity are different statistical problems.

Baseline Information and Action Permissions

All methods receive the same historical batch, online contexts, and realized rewards within a paired seed. The distinction is how they interpret the log and whether they may target disagreement. Table 3 makes these permissions explicit.

Multiplicity and Selection Policy

The tables report effect sizes, paired win counts, and approximate 95% intervals; they do not turn every cell in a large robustness grid into an independent discovery claim. Families such as the 24-cell healthcare surface are reported completely, including negative cells, so their interpretation is the stability pattern rather than 24 separately selected tests. Whenever an operating point was selected from exploratory seeds, only the prespecified choice was rerun on disjoint seeds before being described as holdout evidence. The fixed-forgetting experiment is the clearest example. Bounded-evidence and hybrid detectors failed the exploratory criterion and therefore were not advanced to a confirmatory stage. No uncorrected exploratory p-value is used to claim a new method improvement.

Stationary Mechanism Controls

Diagnostic Necessity and Veto

The no-diagnostic comparison is the key mechanism control because it uses the same candidate-world set and trust update. The difference is the action distribution: ordinary exploitation may not visit the disputed arm, whereas CDV explicitly buys observations that separate the naive and adjusted predictions.

No-Harm and Specificity Controls

The gate is an operating-point choice. Aggressive diagnostic selection retains the largest probe-necessary gain but incurs avoidable regret where worlds are already decision-equivalent. The result does not support a zero-cost diagnosis claim. The low and high gates use exploitation-gap cutoffs 0.2 and 0.6, respectively; the decay gate moves from 0.6 to 0.2 over the first 50 rounds. All three use the median arm-disagreement cutoff.

Implementation and Baseline Audits

Candidate-World Composition

The candidate set is a safety design choice rather than a harmless implementation detail. Table 6 reruns the same three stationary regimes with 500 independent seeds while changing only the available worlds. Removing the cold world is especially damaging under a wrong graph. Conversely, the two-world naive-plus-causal set is strong when its adjusted model is useful, but it lacks an online-only escape route.

Trust-Loss and Graph Perturbations

The Huber-plus-discount update is not uniquely privileged by the evidence. Alternative trust losses can be substantially better in these stationary cells (Table 7); the purpose of Huberization is outlier resistance and stable behavior across the broader campaign. We therefore do not claim that the default update is an optimizer over trust rules.

History, Scale, and Diagnostic-Strength Sweeps

The older robustness campaign used 500 seeds per cell and varied one factor at a time. Table 9 shows that more history benefits all warm starts while cold start is, by construction, unchanged. Scaling the number of actions and covariates makes cold exploration expensive; the skeptical warm start retains substantial value, although a correct causal model remains the endpoint oracle.

Diagnostic pressure also has a non-monotone cost. In the generic skeptical learner, increasing the falsification coefficient past the moderate range causes regret to grow rapidly in every stationary regime (Table 10). This observation motivated the later no-harm gates and the fresh-seed calibration in Table 5; it does not justify choosing a coefficient after seeing the final test cells.

Method	Historical initialization	Online model weighting	Diagnostic target	Cold fallback
Cold	None	No	No	Itself
Naive warm start	Associational log fit	No	No	No
Pure causal	Adjusted log fit	No	No	No
Loss weighted	Multiple warm fits	Prediction loss	No	Optional world only
LWD+Diag	Multiple warm fits	Prediction loss	Same target as CDV	Optional world only
No-diagnostic CSB	Full candidate set	Trust update	No	Candidate world
CDV-NC	Full candidate set	Trust update + rescue	Yes	No veto blend
CDV-Veto	Full candidate set	Trust update + rescue/veto	Yes	Explicit blend

Table 3: Baseline capability audit. LWD+Diag is deliberately granted the same diagnostic action target, so comparisons against it do not withhold the central exploration signal from a strong ensemble baseline.

Method	Diagnostic regime	Wrong graph
Pure causal	26.64 ± 0.41	1585.42 ± 43.00
Loss weighted	106.58 ± 9.20	126.05 ± 5.85
No-diagnostic CSB	152.91 ± 5.59	124.11 ± 5.56
CDV diagnostic	26.53 ± 0.39	—
CDV-Veto	—	127.59 ± 7.50

Table 4: Component controls at horizon 18,000. Diagnostic selection is essential in the probe-necessary condition; the cold fallback prevents the large wrong-graph failure of pure causal transfer.

Method	Diagnostic	Mean ovhd.	Wrong graph
Entropy diag.	26.12	3.05	127.71
No diagnostic	149.25	0.00	125.85
Loss weighted	95.20	-2.69	127.07
Decay gate	33.15	1.52	126.10
Low gate	55.81	1.00	126.92
High gate	26.74	3.51	128.91

Table 5: Fresh 1200-seed gate calibration. Overhead is relative to no-diagnostic CSB averaged across clean, confounded, and mixed regimes.

Strong Target-Aware Ensembles

A strong baseline is allowed to use the same diagnostic target. The independently seeded 1200-seed comparison in Table 11 reports the best target-aware loss-weighted-disagreement (LWD+Diag) variant. Negative differences favor CDV. CDV is lower in ten of thirteen conditions, but LWD+Diag is locally better with five arms, a small gap, or short history. Thus diagnostic targeting is a transferable mechanism; the evidence does not support attributing every gain to CDV-specific weighting.

Predeployment Regime-Uncertainty Risk

The mixture analysis is not an online switch experiment. Before deployment, let p denote the probability that the supplied graph is wrong. For a fixed policy π, its expected regret is the affine combination

$$R_\pi(p) = (1-p)\bar{R}_{\pi,\text{valid}} + pR_{\pi,\text{wrong}},$$

where the valid endpoint averages the clean, diagnostic, and high-triage diagnostic settings. No method observes p or the regime label online. The curve asks which fixed deployment policy minimizes risk under a specified prior over reliability.

On the evaluated grid with spacing .025, CDV-Veto has the lowest mean for $p_{\text{wrong}} \in [.05, .775]$. The endpoint policies correctly retake the lead near certainty: pure causal transfer when validity is nearly assured and cold start when severe misspecification is nearly certain. Linear interpolation places the CDV–causal and CDV–cold crossings near .031 and .791, respectively. These values summarize this simulator and are not universal deployment thresholds.

Healthcare-Style Stress Test

The simulator instantiates four treatments, observed patient features, a latent risk factor, confounded historical triage, and either a valid or deliberately incorrect adjustment structure. It is a controlled healthcare-style mechanism benchmark, not clinical evidence.

The 24-cell sensitivity grid varies triage strength and hidden-risk level with 1500 paired seeds per cell. CDV-NC beats LWD+Diag in 24/24 diagnostic cells (mean paired gain 3.41) and no-diagnostic CSB in 24/24 cells (gain 25.61). CDV-Veto beats pure causal transfer in 24/24 wrong-graph cells (gain 274.55) and the strongest ordinary warm-start comparator in 18/24. Cold start remains better in all 24 known-wrong cells by 118.01 on average. The six warm-start losses occur at the smallest hidden-risk level, where incorrect adjustment is less damaging.

Complete Triage–Risk Sensitivity Surface

Tables 14 and 15 expose all 48 environment cells rather than only the across-cell win counts. Each cell aggregates 1500 paired seeds. For the diagnostic regime, "gain" is LWD+Diag regret minus CDV-NC regret. For the wrong-graph regime, it is the best ordinary warm-start regret minus CDV-Veto regret. Positive values therefore favor CDV in both tables.

Across the same 24 wrong-graph cells, the paired gain over pure causal transfer ranges from 87.61 to 560.32 and is positive in every cell. The paired difference relative to cold start has the opposite sign in every cell, ranging from -259.24 to -40.10. This joint pattern is the intended non-oracle result: the veto limits harmful transfer when validity is unknown,

Candidate set	Confounded	Mixed	Wrong graph
Full (naive, adjusted, graph, cold)	195.42 ± 4.61	193.36 ± 4.50	383.41 ± 10.06
Naive + adjusted	37.44 ± 1.51	36.26 ± 1.58	206.59 ± 11.21
Naive + cold	137.27 ± 6.08	132.62 ± 5.28	137.27 ± 6.08
No cold fallback	237.86 ± 5.43	236.34 ± 5.37	543.10 ± 16.22
No graph-selective world	54.51 ± 2.18	52.98 ± 2.17	175.32 ± 8.39

Table 6: Candidate-set audit at 500 seeds per cell. Entries are mean regret and 95% confidence half-width. This audit studies the generic skeptical learner before the final rescue/veto policy and is used to expose implementation sensitivity, not to select the reported CDV variant.

Trust update	Conf.	Mixed	Wrong
Huber + discount	195.42	193.36	383.41
No discount	23.00	21.80	127.05
No Huber	10.59	9.71	123.55
Squared loss	87.87	86.64	203.56

Table 7: Mean regret in the 500-seed trust-update audit. The apparent stationary advantage of undiscounted or non-Huber alternatives is not used to retrofit the method after observing the later shift tests.

Specified graph	Pure causal	Skeptical set
Correct	6.39 ± 0.38	195.42 ± 4.61
Missing edge	63.79 ± 3.59	227.09 ± 5.67
Extra edge	1466.97 ± 42.84	361.87 ± 9.00
Random graph	1802.95 ± 83.27	367.16 ± 11.17
Wrong graph	1585.42 ± 43.00	383.41 ± 10.06

Table 8: Graph-perturbation sweep with 500 seeds per cell. A conservative candidate set pays a large price at the known-correct endpoint but limits the catastrophic transfer caused by severe graph errors.

while discarding the log remains better when a severe graph error is known in advance.

Public IHDP Semi-Synthetic Benchmark

Data and Construction

We use the public NPCI/IHDP CSV replications distributed with the CEVAE repository. Each row contains treatment, factual outcome, two potential outcomes, and 25 covariates. Algorithms receive only factual historical treatment/outcome tuples and online bandit feedback. Potential outcomes are held by the environment for reward generation and regret evaluation.

The original ten replications are nearly a one-action policy problem: treatment 1 is optimal for 93.5% of held-out units. These original conditions are retained as negative controls. The balanced policy stress test subtracts the historical-split median treatment effect from treatment-1 potential and factual outcomes. This uses no held-out outcomes and preserves covariates, individual treatment-effect heterogeneity, and factual residual noise. The resulting held-out optimal-treatment rate is 49.4%.

Results

CDV improves over cold start by 180.43 under full adjustment and 180.52 under partial adjustment. Replication-clustered approximate 95% intervals are [55.61, 305.25] and [56.48, 304.56]. Paired gains for CDV over no-diagnostic CSB are -0.47 ([-1.01, 0.06]) and -0.29 ([-0.99, 0.42]); hence the two methods are tied. Intervals for CDV against the other warm-start methods also include zero after clustering. The defensible interpretation is external-validity support for robust warm starting, not a unique diagnostic-action win.

Candidate-Set Misspecification

Historical batches are generated by the original linear structural model. Deployment then adds an unseen linear component, an unseen nonlinear component, or an abrupt reward drift; none is represented by a fitted candidate world. Each cell uses 128 paired seeds and horizon 9000.

CDV's paired gains over causal transfer are 215.04, 505.65, and 345.87, respectively, with approximate paired 95% intervals excluding zero. Cold start is nevertheless better in the in-set and nonlinear endpoints. Under valid-graph unseen linear and nonlinear components, CDV is competitive but not uniformly better than target-aware or ordinary loss-weighted ensembles. In the separate abrupt-drift condition, CDV has mean regret 647.99 versus 632.02 for pure causal and 635.17 for LWD+Diag, so the drift test supplies no CDV gain. Candidate coverage is therefore a limitation rather than an assumption that the experiment silently enforces.

Priced Diagnostic Actions

Probe cost is subtracted from both realized and expected reward and therefore affects action selection and regret online; it is not a post-hoc penalty. The grid contains 5760 runs: three scenarios, five costs, six algorithms, 64 paired seeds, and horizon 9000.

Across all ten wrong-graph/cost cells, CDV is significantly better than pure causal transfer. It does not uniformly beat cold start or strong noncausal warm starts. In the correctly specified diagnostic world, the robustness mechanism has a significant opportunity cost relative to the causal endpoint oracle at every price.

History-size sweep				Action/context scale sweep			
n_{hist}	Causal	Naive	Skeptical	$K = d$	Causal	Skeptical	Cold
50	17.07	199.07	226.37	5	6.39	195.42	266.32
100	11.26	175.19	214.01	10	16.03	440.96	1528.32
400	6.39	122.78	195.42	20	81.05	1137.02	7247.63
1000	4.99	92.82	182.66				
3000	3.04	57.93	166.50				

Table 9: Mean regret in selected 500-seed robustness sweeps. These generic candidate-set results establish scaling and data-volume behavior; they are not substituted for the final CDV rescue/veto evaluation.

Diagnostic coefficient	Confounded	Wrong graph
0	189.13	376.07
.25	174.83	363.42
.50	195.42	383.41
1	334.91	515.53
2	977.27	1077.71
4	3328.38	3079.11

Table 10: Mean regret in a 500-seed diagnostic-strength sweep.

Unannounced Regime Switches

Protocol

The historical batch and initial online segment share a regime. Without warning, the reward mechanism changes from valid to wrong graph or from wrong graph to valid after 25%, 50%, or 75% of horizon 9000. Six algorithms and 128 paired seeds yield 4608 completed runs. Evaluation records total, pre-change, post-change, early-500, and late-500 regret, as well as final candidate-world weights.

Cumulative-Trust Result

CDV does not significantly improve on causal or no-diagnostic CSB for valid-to-wrong switching. It is best in mean total regret for wrong-to-valid switches at 25% and 50%, but cold start is best at 75%. Most importantly, the final weights do not implement the claimed veto/re-trust transition. Static mixture curves must therefore not be interpreted as change-point experiments.

Fixed Forgetting Holdout

An exploratory sweep considered $\gamma \in \{1, .999, .995, .99, .98\}$ at the 50% switch. Only $\gamma = 1$ and .999 were carried to the independently seeded 2304-run holdout covering all six direction/timing cells.

Fixed forgetting helps only when the initially wrong segment is short enough to leave substantial recovery time. It significantly harms every valid-to-wrong timing and the late wrong-to-valid timing. With matched $\gamma = .999$, the diagnostic method significantly beats its no-diagnostic counterpart only in the 50% wrong-to-valid cell. A single global discount cannot provide symmetric adaptation.

Additional Tracking Attempts

We also evaluated two oracle-free attempts to make trust reversible:

1. an exponentially smoothed prediction-loss increase detector; and
2. a cap on cumulative-loss gaps relative to the currently best world.

The loss detector could observe the large valid-to-wrong residual increase but not the wrong-to-valid change, whose aggregate prediction-loss distribution barely moves. In a 896-run cap sweep, cap 120 was numerically identical to the base method and cap 80 changed regret by less than one point. Cap 40 produced an unstable early wrong-to-valid gain whose 95% interval crossed zero and significantly harmed the 75% valid-to-wrong switch by 39.74 (baseline-minus-cap interval [-69.50, -9.99]).

A final 1024-run hybrid combined bounded evidence with a one-sided windowed loss-increase detector. The mid detector fired after 100% of valid-to-wrong switches, but its regret gains relative to base CDV were -157.46, -402.30, and -124.67 at the 25%, 50%, and 75% switch times; the first and third intervals excluded zero. The largest wrong-to-valid mean gain was unresolved, and the most sensitive detector significantly harmed the stationary valid control. Detection is therefore not the missing ingredient: a global reset discards useful state while the online predictors are already adapting. Because no exploratory configuration met the selection criterion, no confirmatory holdout was run and none is promoted as an improved method.

Experiment Accounting and Selection

The experimental campaign was organized by reviewer objection rather than by one favorable benchmark. Endpoint sanity tests ask whether causal and cold oracles win when validity is known. Diagnostic controls ask whether ordinary reward-seeking actions can identify the world. Target-aware baselines isolate the action-target contribution. Candidate, graph, and trust audits expose implementation dependence. Public-data and healthcare tests examine transport beyond the minimal simulator, and the change experiments define a negative boundary. Settings chosen during exploration are never relabeled as holdout evidence.

Condition	CDV	LWD+Diag	CDV–TA	Condition	CDV	LWD+Diag	CDV–TA
Base	26.12	26.20	-0.08	Gap .35	26.12	26.20	-0.08
5 arms	5.62	5.07	0.55	Gap .60	20.11	20.37	-0.26
12 arms	62.00	63.25	-1.25	History 80	58.08	55.24	2.85
20 arms	138.93	143.36	-4.43	History 240	26.12	26.20	-0.08
Gap .12	49.36	47.09	2.27	History 960	0.10	0.33	-0.23
Noise .02	25.49	29.46	-3.97	Noise .10	28.27	28.32	-0.05
Noise .20	33.80	33.82	-0.02				

Table 11: Independent 1200-seed target-aware comparison at horizon 18,000. TA denotes the best LWD+Diag operating point. Negative CDV–TA means lower regret for CDV.

p_{wrong}	Causal	LWD+Diag	CDV-Veto	Cold	Lowest mean
0.000	15.07	20.71	24.88	343.30	Causal
0.025	36.05	36.96	37.96	346.31	Causal
0.050	57.03	53.22	51.03	349.32	CDV-Veto
0.250	224.88	183.26	155.61	373.36	CDV-Veto
0.500	434.68	345.80	286.35	403.42	CDV-Veto
0.775	665.47	524.60	430.15	436.48	CDV-Veto
0.800	686.45	540.86	443.22	439.49	Cold
1.000	854.29	670.89	547.81	463.54	Cold

Table 12: Selected points on the 2700-seed healthcare regime-mixture curve. The valid endpoint is fixed before mixing, and no policy is tuned separately at each value of p_{wrong}.

Regime	Method	Regret	95% half-width
Diagnostic	Causal	14.32	0.92
Diagnostic	CDV-NC	17.48	1.18
Diagnostic	LWD+Diag	20.46	1.04
Diagnostic	No diagnostic	44.27	1.49
Wrong graph	Cold	463.54	14.51
Wrong graph	CDV-Veto	547.81	17.17
Wrong graph	No diagnostic	633.59	14.67
Wrong graph	Loss weighted	674.04	26.99
Wrong graph	Causal	854.29	12.00

Table 13: Healthcare-style benchmark with 2700 seeds per method and setting.

	Risk 1.0		Risk 1.4		Risk 1.8		Risk 2.2	
Triage	CDV	Gain	CDV	Gain	CDV	Gain	CDV	Gain
1.5	11.58	1.38	13.91	3.08	16.68	4.87	20.03	5.59
2.0	11.76	1.43	13.91	2.96	17.67	4.76	21.02	6.46
3.0	12.75	1.34	15.46	2.06	19.10	4.56	22.49	5.94
4.0	13.33	1.29	16.69	1.68	20.57	3.35	23.72	6.08
5.0	13.86	1.31	16.99	2.43	21.28	3.18	24.11	6.27
6.0	14.65	0.99	17.48	2.21	21.50	3.52	24.60	5.10

Table 14: Complete diagnostic healthcare sensitivity surface. Entries give CDV-NC mean regret and paired gain over LWD+Diag. All 24 gains are positive; the range is 0.99–6.46 and the mean is 3.41.

	Risk 1.0		Risk 1.4		Risk 1.8		Risk 2.2	
Triage	CDV	Gain	CDV	Gain	CDV	Gain	CDV	Gain
1.5	249.01	-19.68	430.83	2.45	670.64	45.53	1003.38	111.67
2.0	246.56	-18.55	447.52	11.93	671.73	75.09	999.18	243.10
3.0	251.83	-21.53	461.36	4.80	679.58	120.72	1027.07	285.62
4.0	262.59	-30.91	440.27	36.35	711.85	122.70	974.47	326.03
5.0	257.38	-27.54	442.06	26.96	699.95	133.39	992.87	395.52
6.0	262.65	-27.09	443.28	40.82	670.37	161.74	1028.42	350.56

Table 15: Complete wrong-graph healthcare sensitivity surface. Entries give CDV-Veto mean regret and paired gain over the best ordinary warm-start comparator. The six negative cells are exactly the lowest-risk column.

Method	Full adjustment	Partial adjustment
No-diagnostic CSB	121.95	122.04
CDV-Veto	122.42	122.33
LWD+Diag	127.30	128.28
Loss weighted	127.85	127.99
Pure causal	135.10	136.71
Naive	140.38	140.38
Cold	302.85	302.85

Table 16: IHDP balanced-policy mean regret over ten public replications and 32 paired splits per replication, at horizon 3000.

Wrong-graph deployment	CDV	Causal	Cold
In candidate set	411.22	626.26	303.50
Unseen linear component	239.23	744.88	215.94
Unseen nonlinear component	596.30	942.17	365.66

Table 17: Mean regret under wrong-graph and candidate-misspecified deployment.

Scenario	CDV probe rate	
	Cost 0	Cost 1
Correct diagnostic graph	.760	.317
In-set wrong graph	.063	.003
Wrong graph + unseen linear	.081	.016

Table 18: Diagnostic selection responds to explicit probe price.

Switch	25%	50%	75%
V→W: CDV regret	1633.12	2876.87	2745.17
V→W: causal regret	1622.62	2866.30	2731.64
V→W: final adjusted weight	.953	.930	.940
W→V: CDV regret	3457.68	3035.36	1847.49
W→V: cold regret	3864.21	3160.72	1783.72
W→V: final adjusted weight	.172	.216	.221

Table 19: Unannounced switch results (V: valid; W: wrong graph). Low late-window regret can coexist with path-dependent trust because online reward predictors continue updating.

Switch cell	Gain: .999 vs. 1	Paired 95% CI
V→W, 25%	-174.16	[-265.66, -82.66]
V→W, 50%	-268.58	[-384.44, -152.72]
V→W, 75%	-106.80	[-147.42, -66.17]
W→V, 25%	840.56	[469.61, 1211.51]
W→V, 50%	24.84	[-158.32, 208.00]
W→V, 75%	-104.22	[-149.94, -58.49]

Table 20: Independent holdout for fixed trust forgetting (V: valid; W: wrong graph). Positive gain means lower regret for $\gamma = .999$.

Attempt	Cell	Gain over base	Paired 95% interval	Decision
Fixed forgetting .999	V→W, 50%	-268.58	[-384.44, -152.72]	Holdout harm
Fixed forgetting .999	W→V, 25%	840.56	[469.61, 1211.51]	Holdout gain
Bounded evidence, cap 40	W→V, 25%	197.91	[-228.81, 624.64]	Unresolved
Bounded evidence, cap 40	V→W, 75%	-39.74	[-69.50, -9.99]	Exploratory harm
Hybrid, mid detector	V→W, 25%	-157.46	[-266.18, -48.74]	Exploratory harm
Hybrid, mid detector	V→W, 50%	-402.30	[-815.27, 10.67]	Unresolved
Hybrid, mid detector	V→W, 75%	-124.67	[-203.01, -46.33]	Exploratory harm

Table 21: Selection-aware summary of adaptation attempts. Positive gain means lower regret than cumulative-trust CDV. Only the fixed-forgetting rows are from an independent holdout; bounded and hybrid rows remain exploratory and are included to document failed routes rather than enlarge the positive claim.

Evidence block	Scale	Role in the submission
Stationary mechanism and implementation audits	500 seeds/cell	Mechanism, graph, candidate-set, and trust-update controls
No-harm and target-aware defenses	1200 seeds/cell	Fresh-seed gate calibration and strongest diagnostic baseline
Healthcare primary benchmark	2700 seeds/method-setting	High-confidence controlled application-style stress test
Healthcare triage–risk surface	504,000 per-seed rows	336 complete algorithm cells; 24 diagnostic and 24 wrong-graph environments
Public IHDP replications	8960 runs	Original-policy negative control plus balanced-policy stress test
Candidate misspecification	5376 runs	Unseen linear/nonlinear components and abrupt drift
Priced diagnostics	5760 runs	Online price response over three scenarios and five costs
Unannounced switches	4608 runs	Two directions, three switch times, pre/post and terminal diagnostics
Fixed-forgetting holdout	2304 runs	New seeds, two preselected discounts, all six switch cells
Bounded/hybrid tracking	896 + 1024 runs	Exploratory failures retained for falsification and boundary reporting

Table 22: Accounting of the main defense experiments. Run counts refer to completed simulation jobs unless explicitly identified as per-seed output rows. The table separates confirmatory evidence from exploratory method search.

Reproducibility Artifact

The anonymous code-and-data archive contains:

- the candidate-world models, trust update, diagnostic score, rescue, and fallback implementation;
- standalone runners for stationary controls, healthcare sensitivity, IHDP, candidate misspecification, diagnostic cost, regime switches, fixed forgetting, and tracking attempts;
- unit tests for environment pairing, trust updates, cost accounting, and tracking wrappers;
- raw per-seed CSV files for the new stress tests, compact aggregate tables, metadata, dataset hashes, and human-readable assessments; and
- commands and seed ranges sufficient to reproduce every new table.

The archive excludes machine addresses, credentials, private paths, scheduler traces unrelated to scientific results, and unpublished author identity information. Included public IHDP files are identified by their literature citation and SHA-256 hashes. They are not covered by the authors' future code license and remain subject to the original dataset terms.

Limitations

The finite world set is designed rather than learned. Trust weights measure relative online predictive utility, not graph truth. Diagnostic interventions may be costly or unavailable. The healthcare study is simulated, and IHDP has only two actions and no dedicated diagnostic arm. The current cumulative trust rule is path dependent under deployment-time mechanism changes, while fixed forgetting and bounded evidence introduce direction-dependent trade-offs. No experiment in this submission establishes clinical safety, universal dominance, or adaptively optimal regret over arbitrary graphs and shifts.