WHEN THE SAFETY SIGNAL LIES: CORRUPTED COST FEEDBACK IN CONSTRAINED MULTI-AGENT RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe multi-agent reinforcement learning assumes that the cost signal driving constraint enforcement is reported faithfully, yet decentralized deployments receive it over a network as cost reports, an attack surface. We first show that for any connected doubly stochastic mixing matrix, an unclamped fixed point of consensus dual ascent constrains only the *network average* of the agents' cost estimates, not each agent's. We then establish corruption mass conservation: along a fixed primal sequence, the unprojected consensus recursion preserves the aggregate dual bias that corrupted reports inject, so topology can only redistribute it. A biased multiplier scales the constraint gradient, whereas a reward perturbation that shifts every return equally leaves the policy gradient unchanged. We propose Robust Constraint Estimation (RCE), which trims each agent's cost reports and adds a dispersion margin; under an honest-source majority it bounds the estimation error and, unlike gating updates on the untrimmed mean, cannot be stalled by over-reporting. In one instrumented testbed, persistent under-reporting from one of three sources raises true network-average cost by 3.99 against a budget of 25 (paired 95% CI $[3.27, 4.70]$) while the reported cost falls; a matched zero-mean perturbation does not reproduce the effect, an unconstrained control shows that the constraint holds cost near budget, and the effect increases across the tested doses $B/d \in \{0.25, 0.5, 1.0\}$. Within each block RCE removes 80.3% of the effect at $M = 3$, where it reduces to the median, and 82.1% at $M = 5$, where its margin is live; the margin adds clean-run conservatism but no measurable suppression, and residual harm remains. Finally, corrupting a single agent's reports, the case the spreading result addresses, yields the intended concentrated perturbation, but the realized multipliers do not follow the fixed-primal prediction: displacement is not localized on the attacked agent, and a median-deviation monitor pre-declared to catch the attack without communication detects nothing under either topology. Empirical claims are descriptive: one testbed, one operating point, three seeds.

1 INTRODUCTION

Constrained Markov decision processes (CMDPs) formalize safety as a bound on expected cumulative cost (Altman, 1999), constrained policy optimization extends policy search accordingly (Achiam et al., 2017; Tessler et al., 2019b; Chow et al., 2018; Stooke et al., 2020), and multi-agent reinforcement learning (MARL) extensions build on both (Gu et al., 2023; Lu et al., 2021; Ying et al., 2023), typically with PPO-style learners (Schulman et al., 2017; Yu et al., 2022) and safety benchmarks (Ji et al., 2023; Ray et al., 2019). These methods assume that the safety-cost signal used to enforce the constraint is observed correctly, or corrupted only by benign estimation noise. In decentralized methods that assumption cannot be checked locally: each agent's multiplier is driven by cost estimates reported over a network, and these are *learned variables* rather than raw observations (Kahe & Kebriaei, 2026), with no ground truth at the receiver.

The immediate objection is that this is reward poisoning (Zhang et al., 2020c; Rakhsha et al., 2021; Wang et al., 2024) with r relabeled C; in the Lagrangian both are scalars in one objective, so the substitution looks notational. It is not. The cost signal drives a dual variable, that variable multiplies the constraint gradient, and in decentralized methods it is consensus-averaged. Section 4 shows

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.

that in the *unprojected* dual recursion along a fixed primal sequence, doubly stochastic consensus preserves the aggregate dual bias from corrupted reports and only redistributes it (Theorem 2): a constant shift in one agent's cost estimate spreads over all multipliers on a connected graph, evenly up to a bounded term, but stays on that agent without communication (Proposition 3). Whether this structure survives in a live learner, whose multipliers are projected and whose policy adapts, is a separate question; in our testbed it does not (Section 7.5).

Contributions and scope. To ask what corrupted cost reports do to a decentralized primal–dual learner, and whether an estimator can contain them, we formalize the constrained Markov game with corrupted cost feedback and identify which constraint the consensus dual enforces (Proposition 1); establish corruption mass conservation and how topology redistributes the bias (Theorem 2, Proposition 3); and show that the dual channel scales the constraint gradient where a return-level reward offset leaves the policy gradient unchanged (Proposition 4). We present Robust Constraint Estimation (RCE), with a bounded-error guarantee under $M \geq 2f + 1$ (Theorem 5) and liveness: over-reporting cannot stall it, as it can stall updates gated on the untrimmed mean (Proposition 7). In one testbed, validated against pre-declared gates, under-reporting from one of three sources shared by all agents raises true cost while the reported cost falls, RCE removes about four-fifths of that increase at each source count tested, and a concentrated attack's predicted spreading does not appear in the realized multipliers. The results of Sections 3–5 are proved in Appendix A; the experiments report implementation behavior at one operating point and verify no theorem (Section 8).

2 Related Work

Safe RL and safe MARL. The CMDP formalism (Altman, 1999) underlies CPO, RCPO, Lyapunov methods and PID control of the multiplier, the main single-agent approaches (Achiam et al., 2017; Tessler et al., 2019b; Chow et al., 2018; Stooke et al., 2020), surveyed by Gu et al. (2024); Xu et al. (2022), and robust CMDPs harden the formalism against model uncertainty (Russel et al., 2020), a different perturbation channel from a corrupted report. MACPO (Gu et al., 2023) brought constraint-satisfaction guarantees to cooperative MARL, with MAPPO-Lagrangian as its Lagrangian counterpart; a parallel line replaces the central trainer with local updates plus consensus on the duals (Lu et al., 2021; Ying et al., 2023; Kahe & Kebriaei, 2026), building on classical distributed optimization (Nedić & Ozdaglar, 2009). That literature already names the exposure created by exchanging learned variables (Kahe & Kebriaei, 2026); we ask what happens when an adversary exploits it.

Adversarial robustness and poisoning. Byzantine cooperative MARL has been formalized as a Bayesian game (Li et al., 2024), and MARL has been hardened by minimax objectives, uncertainty-aware training, auxiliary attackers, adversarial value decomposition, regularization and certified robustness to observation perturbation (Li et al., 2019; Zhang et al., 2020b; Yuan et al., 2023; Phan et al., 2021; Bukharin et al., 2023; Pinto et al., 2017; Tessler et al., 2019a; Zhang et al., 2020a; Huang et al., 2017); all target perturbed observations, actions, dynamics or teammates, not a cost report feeding a dual variable. Attacks target the reward (Zhang et al., 2020c; Rakhsha et al., 2021), preference data (Wang et al., 2024), backdoor triggers (Kiourti et al., 2020; Wang et al., 2021) or the action channel (Gleave et al., 2020), and regret has been bounded under corruption budgets on reward or transition observations (Lykouris et al., 2025; Ye et al., 2024). Closest to our work, Figura et al. (2021) show that one adversarial agent in consensus-based MARL can steer the network toward its own objective, but through the shared *value* channel and in an unconstrained problem. Optimization also exploits any gap between a measured and an intended signal (Amodei et al., 2016; Skalse et al., 2022); our zero-mean control tests one form of that reading: undirected noise of the same size does not raise true cost (Section 7.2).

Resilient consensus and corrupted measurements. The components we build on are established, and we claim none of them: resilient consensus filters extreme neighbor values under a graph-robustness condition (LeBlanc et al., 2013), fault-tolerant distributed optimization extends that to a sum of local objectives (Su & Vaidya, 2016; Sundaram & Gharesifard, 2019), and Byzantine-robust distributed learning supplies our trimming primitives (Yin et al., 2018; Blanchard et al., 2017; Chen et al., 2017), which Wu et al. (2021) already apply to decentralized TD learning. That a corrupted measurement channel can move a controller while the reported signal stays plausible is the premise of false-data injection and secure state estimation (Liu et al., 2011; Fawzi et al., 2014; Pasqualetti et al., 2013). Our contribution is the composition: a decentralized *constrained* learner in which

a corrupted cost report drives a dual variable that is consensus-averaged and then multiplies the constraint gradient. To our knowledge no threat model exists for that channel; our analysis uses standard consensus and primal–dual algebra and is new only in what it implies for corrupted cost reports (Propositions 1, 3, 4; Theorem 2).

3 Constrained Markov Games with Corrupted Cost Feedback

A *constrained Markov game* is a tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}, P, R, \{C^i\}, \{d^i\}, \gamma, \rho_0 \rangle$ with N agents, states $\mathcal{S}$, actions $\mathcal{A}^i$, kernel P, shared task reward R, per-agent safety cost C^i with budget d^i, discount $\gamma \in (0,1)$ and initial distribution ρ_0. For $\pi_\theta = \prod_i \pi^i_{\theta^i}$ define

$$J_R(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t\geq 0} \gamma^t R(s_t, a_t)\Big], \qquad J^i_C(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t\geq 0} \gamma^t C^i(s_t, a_t)\Big]. \tag{1}$$

The cooperative objective is $\max_\theta J_R(\theta)$ subject to $J^i_C(\theta) \leq d^i$ for all i, the setting of the decentralized primal-dual methods we study (Lu et al., 2021; Ying et al., 2023; Kahe & Kebriaei, 2026). For Proposition 1 we also need the *network-average* cost return $J_{C,\mathrm{avg}}(\theta) \triangleq \frac{1}{N}\sum_i J^i_C(\theta)$ and its constraint $J_{C,\mathrm{avg}}(\theta) \leq \bar{d} \triangleq \frac{1}{N}\sum_i d^i$. The *per-agent* condition implies the network-average one but not conversely, since one agent far under budget can offset another far over it; with heterogeneous C^i, as in our testbed, the two are distinct. All constraints and cost reports are on the return scale.

Each agent i, the *constraint owner* for C^i, holds θ^i, a multiplier $\lambda^i \geq 0$ and an estimate $\hat{J}^i_C$ of its cost return, by default the mean of return-scale reports from M *sources* that estimate J^i_C. Agents mix multipliers through a nonnegative doubly stochastic W ($W\mathbf{1} = \mathbf{1}, \mathbf{1}^\top W = \mathbf{1}^\top$) on a connected graph, or $W = I$ without communication, and run

$$\theta^i_{k+1} = \theta^i_k + \eta_\theta \nabla_{\theta^i}\big[J_R(\theta_k) - \lambda^i_k J^i_C(\theta_k)\big], \qquad \lambda_{k+1} = \Pi_{[0,\lambda_{\max}]}\big[W\lambda_k + \eta_\lambda\, g_k\big], \tag{2}$$

with $g^i_k = \hat{J}^i_C(\theta_k) - d^i$. Mixing λ by W makes the method decentralized, and it also determines which condition the method enforces: not the one the objective states.

Proposition 1 (What the consensus dual enforces)**.** Assume W *is doubly stochastic and* $\lambda^\star$ *is a fixed point of Eq.* (2) *under residual* $g^\star$*, with no coordinate clamped at* $\lambda_{\max}$. Then $\mathbf{1}^\top g^\star \leq 0$*, equivalently* $\frac{1}{N}\sum_i \hat{J}^i_C \leq \bar{d}$*, with equality when* $\lambda^\star$ *is interior, in which case* $(I - W)\lambda^\star = \eta_\lambda g^\star$*. On a connected graph, and for* $\lambda_{\max}$ *large enough, that scalar equation is the* only *restriction the fixed point places on* $g^\star$*, individual residuals being otherwise unconstrained in sign and magnitude. The per-agent condition* $g^{i\star} \leq 0$ *for all* i *is instead the fixed-point condition of the* non-communicating *update* $W = I$.

The consensus dual enforces only the average. As in Theorem 2, the cause is double stochasticity: while the projection is inactive the multipliers' network average performs scalar dual ascent on the average estimate, a proxy for $J_{C,\mathrm{avg}} \leq \bar{d}$ when the estimates are unbiased, and the mixing absorbs any pattern of individual residuals that sums to zero. The fixed points of Eq. (2) therefore satisfy only the average constraint, and $J^i_C \leq d^i$ is design intent rather than something the dual update delivers. This holds for every connected doubly stochastic W, but only for Eq. (2), not for every decentralized safe RL algorithm.

Corrupted cost feedback and the adversary model. An adversary controls at most f of the M sources feeding a constraint owner and perturbs each report it controls by at most B (Lykouris et al., 2025; Ye et al., 2024); the owner's estimate, and hence its residual in Eq. (2), then shifts by δ^i_k, so $\tilde{g}^i_k = g^i_k + \delta^i_k$, where the size of δ_k depends on the aggregator (one corrupted report among M moves a mean by B/M). The adversary observes reports and multipliers on the channels it controls, may choose its perturbation as a function of (θ_k, λ_k), and cannot touch rewards, actions, observations or transitions; this models a deployment whose control plane is authenticated but whose telemetry plane is not. Corrupted reports bias the dual, the biased multiplier distorts the policy update, and the policy violates the constraint while the reported cost, being the corrupted quantity, still looks acceptable to an evaluator that sees only task return and reported cost (Figure 1, Appendix B). Of the attack instances spanned by *direction* (negative / positive / mixed), *support* (persistent / selective), *adaptivity* (static / adaptive) and *consistency* (consistent / equivocating), **the only one we run is (negative, persistent, static, consistent) at** $f{=}1$.

4 CORRUPTION IN THE DUAL CHANNEL

Let λ_k be the dual trajectory under corruption and λ_k° that of the same algorithm, from the same initialization, under uncorrupted reports; the *dual bias* is $e_k \triangleq \lambda_k - \lambda_k^\circ$. All proofs are in Appendix A.

Theorem 2 (Corruption mass conservation)**.** Assume *Eq.* (2) *with any doubly stochastic W, and that the two trajectories are compared* along the same primal sequence*, hence share residuals g_k. (a) If both remain interior to $[0, \lambda_{\max}]$, then $e_K = \eta_\lambda \sum_{k=0}^{K-1} W^{K-1-k}\delta_k$, and*

$$\mathbf{1}^\top e_K \;=\; \eta_\lambda \sum_{k=0}^{K-1} \mathbf{1}^\top \delta_k, \tag{3}$$

whose right-hand side contains no W. (b) Without interiority, if W is entrywise nonnegative and the corruption under-reports ($\delta_k \le 0$), then $\eta_\lambda \sum_k \mathbf{1}^\top \delta_k \le \mathbf{1}^\top e_K \le 0$: the injected mass bounds the realized bias in magnitude, still with no W.

Double stochasticity preserves the sum, so the property that makes consensus work also leaves it unable to reject corruption mass: in this recursion, no degree of connectivity and no spectral gap removes injected bias. Two caveats are essential. The fixed-primal-sequence condition is a counterfactual device that isolates the dual channel; no experiment here holds the primal sequence fixed. And Eq. (3) says only that the *aggregate* carries no W (the full vector e_K does, as does the primal trajectory), so **Theorem 2 does not establish that the learned system is topology-independent**.

Proposition 3 (Topology redistributes the bias)**.** Assume *the hypotheses of Theorem 2(a), a persistent perturbation confined to owner j, $\delta_k = \delta u_j$ with u_j the jth standard basis vector, and a connected aperiodic graph with second-largest singular value $\sigma_2(W) < 1$.* Then $e_K = \eta_\lambda \delta[\frac{K}{N}\mathbf{1} + r_K]$ *with $\|r_K\|_2 \le (1 - \sigma_2(W))^{-1}$, so every coordinate of the dual bias grows as $\eta_\lambda \delta K/N + O(1)$; under the non-communicating update $W = I$ the bias instead stays on coordinate j and grows as $\eta_\lambda \delta K$.*

Spreading is a property of the recursion. Like Theorem 2, this concerns the *unprojected* recursion under a fixed primal sequence: it says where the perturbation lands *there*, not where displacement appears in the multipliers of a live run, where projection and the policy's response intervene (Section 7.5). With Proposition 1 it also makes topology a trade-off. $W = I$ keeps the bias on the attacked coordinate and enforces the per-agent condition, but for that reason each multiplier tracks its own agent's residual, so with heterogeneous costs the multipliers are dispersed even when nothing is wrong; consensus keeps them close but spreads the bias.

Proposition 4 (Asymmetry of the cost channel)**.** *Let Δ_R^i be the policy-gradient error from a reward corruption that adds $\epsilon(\theta)$ to the return $J_R(\theta)$, and Δ_C^i that from a cost corruption producing dual bias e^i. Then $\|\Delta_R^i\| = \|\nabla_{\theta^i}\epsilon\|$, which vanishes whenever ϵ does not depend on θ^i, whereas $\|\Delta_C^i\| = |e^i| \cdot \|\nabla_{\theta^i} J_C^i(\theta)\|$, which is nonzero whenever $e^i \neq 0$ and the constraint gradient is nonzero.*

Two asymmetries. The first is between channels: a reward perturbation that shifts every return by the same amount is invisible to the policy gradient, which is why reward-poisoning attacks vary the perturbation with state and action (Zhang et al., 2020c), whereas on a cost report even a constant offset is multiplied into $\nabla_\theta J_C$ through the dual bias. The second, which shapes the defense, is between *directions* within the cost channel, because enforcement is a threshold mechanism whose failure modes are not mirror images. Under-reporting relaxes the perceived constraint and accumulates in the dual (Theorem 2); it is the instance we test, and in A1 it holds the reported cost near budget while the true cost moves (Section 7.2). Over-reporting can drive the multiplier only to $\lambda_{\max}$, so its worst case is maximal conservatism: degraded but live, *provided the update is unconditional*. Gating updates on an aggregate that keeps the corrupted report converts over-reporting into a denial of learning (Proposition 7), so a defense must bound both directions inside the estimator without refusing to update.

5 ROBUST CONSTRAINT ESTIMATION

RCE never gates the update; it bounds how far corruption in either direction can move the estimate, by trimming. Each constraint owner i receives return-scale reports $\{\tilde{J}_C^{i,m}\}_{m=1}^M$ from its M sources

and forms

$$\hat{J}_C^i \leftarrow \mathrm{TrimMean}_f\big(\{\tilde{J}_C^{i,m}\}\big), \quad \sigma_k^i \leftarrow \mathrm{MAD}\big(\{\tilde{J}_C^{i,m}\}_{m\in\mathcal{T}}\big), \quad \bar{J}_C^i \leftarrow \hat{J}_C^i + \beta\,\sigma_k^i, \tag{4}$$

discarding the f largest and f smallest reports, measuring dispersion by the median absolute deviation (MAD) over the retained set $\mathcal{T}$ of size $M-2f$, and feeding the pessimistic $\bar{J}_C^i$, on the return scale of d^i, to the unmodified bounded dual ascent of Eq. (2) (Algorithm 1, Appendix C).

Assumption 1. *Fix a constraint owner i and a round k. (i) At most f of that owner's M sources are corrupted, with $M \geq 2f+1$. (ii) Conditional on θ_k, its honest sources report unbiased estimates of $J_C^i(\theta_k)$ with sub-Gaussian parameter ς,* mutually independently across that owner's M sources. *(iii) Each corrupted report is perturbed by at most B; Eq.* (5) *does not depend on B. Nothing here requires independence* across *owners, and the results below do not use it.*

Theorem 5 (Bounded estimation error, per owner and per round)**.** *Under Assumption 1, with probability at least $1-\alpha$,*

$$\big|\hat{J}_C^i - J_C^i(\theta_k)\big| \;\leq\; \underbrace{\varsigma\sqrt{\tfrac{2\log(2/\alpha)}{M-2f}}}_{\textit{statistical error}} \;+\; \underbrace{\tfrac{f}{M-2f}\,\varsigma\sqrt{2\log(2M/\alpha)}}_{\textit{Byzantine error}} \;\triangleq\; \varepsilon(M,f,\alpha). \tag{5}$$

Consequently, if $\beta\sigma_k^i \geq \varepsilon(M,f,\alpha)$ then $\bar{J}_C^i \geq J_C^i(\theta_k)$ with probability at least $1-\alpha$, and any policy satisfying $\bar{J}_C^i \leq d^i$ satisfies $J_C^i \leq d^i$ with the same probability.

The guarantee is conditional. (a) Theorem 5 bounds one owner's estimation error in one round, not cumulative violation, and says nothing about the fixed point of the corrupted dynamics. (b) Its final clause is an implication, not an assertion that Eq. (2) drives $\bar{J}_C^i$ below d^i for each i; by Proposition 1, it does not. What the learner enforces is the average condition, to which the bound transfers by a union bound at confidence $1-N\alpha$ (Proposition 6), and Proposition 7 adds that over-reporting cannot stall the update. (c) The premise $\beta\sigma_k^i \geq \varepsilon$ is a condition, not a consequence, and cannot be checked at run time because ε depends on the unknown ς; against a reference $\hat{\varepsilon}$ calibrated on attack-free data (Appendix C), it holds in a minority of cells (Section 7.3). (d) Eq. (5) is undefined at $2f=M$, the information-theoretic limit rather than a weakness of the analysis, and we do not test that boundary. The bound requires $M \geq 2f+1$ reporting *sources*, not agents, and our source architecture exists to supply them; two deployment assumptions, attack-free calibration and per-owner trimming, bound the threat model (Appendix C).

6 Experimental Setup

A controlled testbed, not a benchmark. All runs use one instrumented environment, built so that the constraint-enforcement loop is observed rather than inferred: `ManySegmentAnt` 6×1 ($N=6$) on the Farama MaMuJoCo backend, with a per-agent cost indicating torso speed above a threshold (0.75) calibrated so that the constraint binds. Three properties of the reference Safe MAMuJoCo suite (Gu et al., 2023) force that choice, so no number here is comparable to a published Safe MAMuJoCo score (Appendix C). We use a common budget $d^i \equiv d = 25$, so the enforced constraint reads $J_{C,\mathrm{avg}} \leq 25$; a ring W except in the $W=I$ arms of Section 7.5; $\eta_\lambda = 0.035$; $\lambda_{\max} = 25$; and independent PPO (Schulman et al., 2017) per agent over 250 rounds of 2000 environment steps. The M sources ($M=5$ in A3, Section 7.3; $M=3$ elsewhere) are parallel batches of current-policy rollouts on separate RNG streams, each reporting every owner's discounted Monte Carlo cost return from its one shared rollout. Corrupting one source therefore shifts every owner's mean aggregate by the same B/M, so Sections 7.2–7.4 inject a *uniform* perturbation $\delta_k = -(B/M)\mathbf{1}$ (only Section 7.5 concentrates it), and the owners' costs are strongly correlated (pairwise 0.836 to 0.998 in a clean seed), so a "k of 6 owners" count is not 6 independent observations. The attack is the instance named in Section 3, applied every round to one source (rotating by seed) at $B = 0.5\,d = 12.5$; the dose sweep varies B alone. RCE uses $f=1$, $\beta=1.5$ and dispersion floor $\sigma_{\min} = 10^{-3}$.

Metrics and statistics. An oracle evaluator reads the uncorrupted simulator signal and computes each agent's true discounted cost return every round; the learner has no access to it. The primary outcome J_C (higher is worse), the run-level counterpart of $J_{C,\mathrm{avg}}$, averages that true cost over agents and all 250 rounds ("whole-run"); the violation rate is the fraction of (round, agent) cells with true cost above d. The *reported cost* J_C^{rep} is the value the dual update consumes (the aggregate, plus the

margin under RCE), averaged the same way, and the *mechanism gap* $\Delta_{\text{mech}} \triangleq J_C - J_C^{\text{rep}}$ is positive when the report understates realized cost; it diagnoses the reporting pipeline, not intent. With three seeds per condition we report per-seed values, paired mean differences with paired 95% t-intervals (2 degrees of freedom), sign consistency and, for scale only, the paired effect size d_z. Three seeds do not support hypothesis testing, so these intervals, and our equivalence checks (a paired 90% interval inside a pre-declared margin), are descriptive rather than inferential.

Provenance. Each campaign's conditions, primary contrasts and decision rules were fixed in a *gates document* before its first run (P1's final adjudication thresholds before its last run completed); claims called pre-declared come from one. Every $M{=}3$ contrast is *paired under common random numbers* (CRN): each condition's sources draw an identical seed stream, so at round 0, before the policies diverge, source values are bit-identical and the treatment is the only difference. The $M{=}5$ block (A3) is a separate campaign paired *within* itself, and no A3 value is comparable to an $M{=}3$ value: A3 ran on a different platform, where seed streams match but policy initializations and source values do not. Across blocks we compare only signs and within-block ratios, never raw values.

7 RESULTS

7.1 INSTRUMENT AND CONTROLLER VALIDATION

Instrument (G10). Three clean constrained runs pass all six pre-declared scientific gates and all nine structural checks (Appendix C). The clean system runs at $J_C = 25.05$ (sd 0.14) against $d = 25$, fixing the operating point for what follows. The average is pinned while seed 0's per-agent true costs are $(21.62, 21.57, 21.69, 23.92, 28.50, 33.99)$, with 46.6% of (round, agent) cells violating and 2–3 of 6 agents above d in every clean seed, which Proposition 1 permits without explaining (Section 8).

Controller (U). Sections 7.2–7.3 compare constrained runs, so none can show the dual doing any work. The unconstrained control lowers $\lambda_{\max}$ from 25 to 0 and changes nothing else; λ is verified to be exactly 0.0 in all $1{,}500$ cells of each run. Removing the constraint raises whole-run J_C from 25.05 to $\mathbf{82.69}$, a paired difference of $+\mathbf{57.64}$ (95% CI $[+52.61, +62.67]$, $d_z = +28.5$, three of three, 6 of 6 agents agreeing); by the final 50 rounds it sits at 97.09 and violates in 100% of cells. Three consequences follow. The constrained system's $J_C \approx d$ is caused by the constraint, the only thing removed, not by a benignly safe environment. The attack below undoes 3.99 of that 57.64, about 7%: corruption shifts the operating point of a controller still doing most of its work, and should not be read as disabling it. And Δ_{mech} stays near zero even at a true cost of 82.69 (a *level* of $+0.32$; the paired $U - A$ *contrast* is $+0.11$, mixed signs): with no adversary, the sources report even this cost honestly. That comparison is within one experiment, not causal identification, but it separates corruption of the signal from absence of a controller.

7.2 EFFECT OF PERSISTENT DIRECTIONAL CORRUPTION (A1)

Condition B applies the attack to one of three sources every round; condition A is the paired clean run under identical seeds. Table 1 collects the conditions of Sections 7.1–7.4.

Primary result. The paired mean difference on whole-run J_C is $\mathbf{B} - \mathbf{A} = +\mathbf{3.99}$ (95% CI $[+3.27, +4.70]$, $d_z = +13.8$, three of three) on a clean baseline of 25.05 against $d = 25$. The violation rate rises by $+17.0$ points ($[+12.6, +21.4]$), agents above budget in the final 50 rounds go from 2–3 to 5–6, and the per-agent increment agrees with the network sign in 6 of 6 agents in every seed. Simultaneously the reported cost *falls* by 0.23 ($[-0.33, -0.14]$, three of three) and Δ_{mech} opens from $+0.21$ to $+4.42$, a paired difference of $+4.22$ ($[+3.56, +4.88]$). With the reported cost held near d, the true cost absorbs most of the injected $B/M \approx 4.17$ shift ($\Delta J_C/(B/M)$ is 0.77–1.05 at every dose and seed), consistent with Proposition 1's estimate-level pinning; the exact B/M shift of the aggregate is an implementation check (Appendix C). We detect no task-return degradation ($+0.53$, CI $[-5.70, +6.76]$, mixed signs): the paired 90% interval $[-3.70, +4.76]$ lies inside the pre-declared equivalence margin of 15.3 return units (clean final return is about -44), but not inside the stricter secondary margin of 1.36, one clean between-seed standard deviation, so equivalence at that margin is not established. **Task performance alone therefore did not reveal the corruption here**, a statement about these observables under these conditions, not a general stealth property.

Table 1: Conditions of Sections 7.1–7.4: whole-run network-average values, three seeds each, $d = 25$; $\Delta_{\text{mech}} > 0$ means the report understates realized cost, and "Agents $> d$" is the range across seeds over the final 50 rounds. The two blocks sit at different source counts and on different platforms and are *not* comparable to one another; contrasts are paired within each block. Dose rows come from a later campaign (Section 7.4).

Condition	Setting	$J_C \downarrow$	Reported	Δ_{mech}	Viol. rate	Agents $> d$
$M = 3$ sources, mean aggregation unless stated (A1, A2, U)						
A clean (G10)	–	25.05	24.84	$+0.21$	0.466	2–3
D zero-mean control	$\sigma = 0.5d$	24.69	24.49	$+0.19$	0.432	1–2
B attack (A1)	$B/d = 0.5$, f=1	29.03	24.61	$+4.42$	0.636	5–6
E clean + RCE (A2)	$\beta = 1.5$	25.17	24.74	$+0.44$	0.467	2
C attack + RCE (A2)	$B/d = 0.5$, f=1	25.95	24.80	$+1.15$	0.502	2–3
U unconstrained	$\lambda_{\max}$=0	82.69	82.37	$+0.32$	1.000	6
dose (undefended)	$B/d = 0.25$	26.77	24.78	$+1.99$	0.538	2–6
dose (undefended)	$B/d = 1.0$	33.02	24.55	$+8.47$	0.785	6
$M = 5$ sources (A3); RCE margin non-degenerate						
A clean	–	25.24	24.88	$+0.36$	0.469	2
B attack	$B/d = 0.5$, f=1	27.45	24.76	$+2.70$	0.569	2–5
E clean + RCE	$\beta = 1.5$	24.47	24.93	-0.46	0.436	2
C attack + RCE	$B/d = 0.5$, f=1	24.87	24.75	$+0.11$	0.448	2

Zero-mean control (D). The pre-declared falsification control replaces the adversary with zero-mean i.i.d. Gaussian noise on the *same* source, with standard deviation equal to the attack's magnitude ($\sigma = B = 12.5$), its mean and scale checked per seed: if undirected noise of equal root-mean-square size reproduced the effect, the phenomenon would be ordinary estimation noise. It does not. $\mathbf{D} - \mathbf{A} = \mathbf{-0.36}$ on J_C, CI $[-1.76, +1.03]$, mixed signs, and -0.02 on Δ_{mech}, again mixed, against $+3.99$ and $+4.22$ with three-of-three sign consistency under the directional attack (Figure 2, Appendix C). At this size, then, the directed perturbation moves true cost and the undirected one does not, consistent with Theorem 2: $\mathbf{1}^\top \delta$ is conserved and zero-mean noise injects no net mass in expectation, though its realized sum is a random walk. Two qualifications. B$-$D ($+4.35$, $[+2.70, +6.00]$) was pre-declared *descriptive only*, on the basis of a power analysis written before any run existed. And the zero-mean noise is not free: it costs 4.7 return units over the final 50 rounds, three of three (Table 3), but it is not a *safety* attack.

7.3 Robust mitigation with RCE (A2, A3)

Both defense campaigns run the same 2×2 attack $\times$ defense factorial over conditions A, B, E and C of Table 1, and the primary quantity in each is the *interaction* $(C - E) - (B - A)$, the change in the attack's effect when RCE replaces the mean. Both use $f = 1$; A2 sits at $M = 3$ and A3 at $M = 5$, the smallest M whose retained set meets the implementation's three-value minimum for a dispersion estimate (Figure 3, Appendix C).

Primary results. At $M = 3$ the interaction on whole-run J_C is $\mathbf{-3.20}$ (95% CI $[-4.73, -1.68]$, $d_z = -5.2$, three of three) against $B - A = +3.99$: the aggregator removes 80.3% of the attack's true-cost effect, and 83.1% on Δ_{mech} (-3.50, $[-5.08, -1.93]$). At $M = 5$ it is $\mathbf{-1.81}$ ($[-2.79, -0.83]$, $d_z = -4.6$, three of three) against a smaller $B - A = +2.21$ (the aggregate shifts by B/M), removing 82.1%, with -1.76 on Δ_{mech} ($[-3.03, -0.50]$, removing 75.6%). Residual harm is positive in every seed of both blocks ($C - E = +0.78$ and $+0.40$), though both 95% intervals include zero: RCE reduces the effect but does not eliminate it.

A2 tests the median; A3 tests the margin. At $M = 3$, $f = 1$ the retained set holds $M - 2f = 1$ value in *every* cell, so its MAD is zero, the margin sits at its floor (0.006% of d), and the point estimate equals the median of the three post-attack reports bit for bit: **A2 tests only the trimming/median special case**, not an aggregator comparison, so no superiority over coordinate-wise median, trimmed mean or Krum is claimed. In A3 the retained set holds exactly 3 values, dispersion exceeds $\sigma_{\min}$ in 99.6–99.8% of cells, and the applied margin is 0.61 under attack against 0.46 when clean, roughly $400\times$ A2's inert value; the corrupted source is trimmed in 100% of cells, as expected for a -12.5 shift. Suppression at $M = 5$ (82.1%) is similar to that at $M = 3$ (80.3%), where the

margin was inert; since each percentage is relative to its own block's $B - A$, the comparison does not require the blocks' raw values to match, and **these experiments do not establish that the live margin increases attack suppression**, which is attributable to trimming in both blocks. What the live margin does buy is clean-time conservatism: $E - A$ is -0.77 on true cost ($[-1.05, -0.50]$) and -0.82 on Δ_{mech} ($[-1.28, -0.36]$), both three of three, so the clean RCE report *overstates* realized cost (-0.46) where the clean unprotected one understates it ($+0.36$), with no detectable change in final-50 task return ($E - A = -1.78$, mixed signs; Table 3).

This is not experimental validation of Theorem 5. The logged flag `guarantee_in_force`, recording whether $\beta\sigma_k^i \geq \hat{\varepsilon}$ held in a cell, is true in only 24.9% of attacked and 15.5% of clean A3 cells: the sufficient condition fails in most cells, and where it holds the flag records the precondition, not the conclusion. (Appendix C checks its independence assumption for consistency.)

7.4 Attack-dose response

To see how the effect scales, a later campaign (DR) with its own gates document varied only the attack budget B, over $B/d \in \{0, 0.25, 0.5, 1.0\}$ at M=3, f=1, paired by seed; its $B/d = 0.5$ arm is A1's condition B. True cost rises at every tested dose: $\Delta J_C = +1.72$ (95% CI $[+1.41, +2.04]$), $+3.99$ ($[+3.27, +4.70]$) and $+7.97$ ($[+6.20, +9.75]$), positive in all three seeds at each, with all nine adjacent-dose increments positive (Figure 4, Appendix C). The reported cost stays near d throughout (24.78, 24.61, 24.55), so the mechanism gap, true minus reported cost, necessarily orders the same way ($+1.78$, $+4.22$, $+8.26$); the informative observation is that the reported cost does not rise with the dose. Even the strongest dose reaches only 13.8% of the unconstrained control's $+57.64$, and ΔJ_C stays between 0.77 and 1.05 times the injected shift B/M at every dose, so the response shows no sign of saturating inside the tested range. Four levels at three seeds establish *ordering over the tested points* and nothing more: the doses are nested contrasts against one shared clean reference, not independent observations, and we fit no functional form. Task return stayed inside the pre-declared equivalence margin at every dose (paired 90% interval within $\Delta_{\text{equiv}} = 15.3$).

7.5 Concentrated corruption and the theory-to-observation gap (P1)

The uniform perturbation of Sections 7.2–7.4 is the limit *toward* which connectivity drives a concentrated one, so those campaigns cannot test Proposition 3. P1 corrupts only `agent_5`'s reports in nine runs at three seeds: attacked ring (R), attacked $W = I$ (I) and its CRN-paired clean $W = I$ control (IC), with G10's CRN-paired clean ring runs as R's reference, at M=3, f=1, B/d=0.5, no defense (Appendix C.5). We write $\text{conc}(x) = \|x\|_\infty / \|x\|_1$ ($1/N = 0.167$ is uniform, 1 a single coordinate). The pre-declared monitor flags round k when the fleet statistic $s_k = \max_i |\lambda_k^i - \text{median}_j \lambda_k^j|$ exceeds a threshold τ, set to the 95th percentile of s_k on the same-topology clean run or swept over a pre-declared grid; its detection excess is the attacked run's firing frequency minus the clean run's.

The premise is instantiated. The measured perturbation is $\delta_k = [0, 0, 0, 0, 0, -B/M]$, the five untargeted entries exactly 0.0 on every round of every run; this is the campaign's one empirical premise. Given it, the unprojected recursion $e_{k+1} = We_k + \eta_\lambda \delta_k$ yields by construction $\text{conc}(e_K) = 0.1725$ under the ring against 1.000 under $W = I$, ring residual norm 0.270 within the bound $(1 - \sigma_2(W))^{-1} = 3$, and $\mathbf{1}^\top e_K = -36.458$ under *both* topologies, matching Eq. (3) to 8.6×10^{-14}; these are computations from the measured δ_k, not observations of the learner.

The realized multipliers depart from the recursion. At the final round, realized displacement (attacked minus CRN-paired clean multipliers) has ℓ_1 norm 2.4, 8.8 and 1.3% of $\|e_K\|_1$ under the ring and 3.8, 117.6 and 46.1% under $W = I$. The comparator is a fixed-primal counterfactual that grows linearly in K, whereas the live policy responds to its biased multiplier, as it does in A1 when the reported cost is held near d; projection also binds (only 26–30 of 250 rounds are interior in both $W = I$ arms), and P1 isolates none of these effects. The pre-declared claim that the monitor catches the attack under $W = I$ is **refuted**, for a structural reason: with no adversary at all, the clean $W = I$ fleet statistic already averages 25–33$\times$ the clean ring's (18.9–22.3 against 0.67–0.75), so the clean-calibrated τ equals $\lambda_{\max}$, which s_k cannot exceed, and at every grid τ both arms fire on at least 96% of rounds; the test has no power there (Appendix C.5). Under the ring the monitor, as predicted, does not fire on the attacked run (excess -0.052), since consensus makes the bias near-uniform.

Displacement is real but not localized. Under $W = I$ the attacked owner's multiplier is displaced by a whole-run mean $|e|$ of 4.40, 4.98 and 5.34, 0.53–0.99 times the attacked fleet's mean λ over the final 50 rounds (a different window, used only as a scale): the attack moves this multiplier materially, yet the monitor does not register it. The displacement is *not* localized on the attacked owner, a pre-declared claim reported as **unsupported**: the owner carries a majority of the whole-run mass in one seed of three (0.61, 0.34, 0.36), and in the other two the largest per-agent displacement sits on an agent that receives exactly zero injected perturbation and is reachable only through the shared environment, since $W = I$ decouples the dual coordinates. The $W = I$ effects on true cost (-1.43, $+9.26$, -2.25) and violation rate (-0.015, $+0.112$, -0.104) flip sign across seeds and are also reported as unsupported, not averaged into a misleading mean.

8 Discussion and Limitations

Three kinds of claim. *Theoretical*: every theorem and proposition above, which holds under its stated assumptions whatever the experiments show. *Descriptive*: the findings of Section 7, each at $n = 3$ in one testbed (one environment, one operating point, two topologies; the main campaigns perturb every owner uniformly, only P1 a single owner), where three-of-three sign consistency, which symmetric noise produces with probability $1/4$, is evidence, not demonstration. *Neither proved nor shown here*: Theorem 5's guarantee as such; Proposition 3's transfer to realized multipliers; any monitor that detects the concentrated attack; the direction of the $W = I$ attack's effect on cost or violation rate; and any safe-MARL baseline comparison, since no MACPO or PID-Lagrangian run exists. Appendix C.10 lists the rest of the untested design space.

Limits of the evidence. Per-agent safety is neither achieved nor claimed: 2–3 of 6 agents exceed d even in clean ring runs. Proposition 1 permits this under consensus, but it constrains cost *estimates* at an unclamped fixed point, not the true costs of a still-learning policy, and clean $W = I$ runs, whose fixed point would enforce the per-agent condition, also show per-agent violations (all six agents exceed d in one of three seeds; Appendix C.5), so these runs do not isolate the consensus dual as the cause. The dose result establishes *ordering*, not a law, and covers the *undefended* line only; P1 adds one topology, not a sweep. Theorem 2 assumes a fixed primal sequence and, for its equality, interior dual trajectories; P1's realized multipliers meet neither condition (attacked and clean runs follow different policies, and 32–52% of clean $W = I$ multiplier cells sit at the floor and 6.7–20.3% at the ceiling), so neither clause applies to them. Theorem 5 bounds one owner's error in one round, not cumulative violation, and a cumulative bound under adaptive corruption (Lykouris et al., 2025) remains open. And we claim no detection: RCE bounds damage without identifying a source, Δ_{mech} needs an oracle a deployment lacks, and the monitor of Section 7.5 detects nothing. Attributing corruption to a source is the open problem this threat model raises.

9 Conclusion

Two consequences of $\mathbf{1}^\top W = \mathbf{1}^\top$ shape the problem. At an unclamped fixed point the consensus dual constrains the network average of the cost estimates, not the N per-agent budgets its objective states, so heterogeneous agents can sit above budget while the reported costs satisfy the average constraint. And in the unprojected recursion along a fixed primal sequence, double stochasticity conserves injected corruption mass, so mixing can change where the total bias lands but not its size. To reduce the total, corruption must be bounded where it enters, as RCE does without ever refusing an update, which is why over-reporting cannot stall it.

Empirically, under-reporting raised true network-average cost by 3.99 while the reported cost fell, matched zero-mean noise did not, and the effect increased at every dose tested; RCE removed 80.3% ($M = 3$) and 82.1% ($M = 5$) of each block's attack effect, left residual harm, and gained no measurable suppression from its live margin. With one agent's reports corrupted, the realized multipliers did not show the structure the fixed-primal recursion predicts: displacement was not localized on the attacked agent, the $W = I$ outcomes had no consistent sign, and the pre-declared claim that a median-deviation monitor would catch the attack without communication was refuted. In this testbed, dual-side analysis described where corruption lands in the recursion but did not predict what an operator observes. Per-agent guarantees, attribution and adaptive corruption remain open.

AI USE STATEMENT

Large language models and AI tools were utilized exclusively to aid in editing, grammatical refinement, and polishing the prose of this manuscript. All theoretical definitions, proofs, algorithmic designs, software implementations, and empirical findings were independently conceived, conducted, and verified by the authors.

ETHICS STATEMENT

This paper describes an attack on safety mechanisms in systems that may be physically deployed. The attack requires write access to a cost-reporting channel; an adversary holding that access in a fielded system already possesses a serious capability, and the vulnerability we describe does not create it. The enabling property is a published algorithmic structure shared by decentralized primal–dual methods (doubly stochastic consensus on dual variables) rather than an implementation flaw in any specific system, so there is no vendor to notify privately and no patch to coordinate. We present the defense alongside the attack; Theorem 5 gives operators a necessary design condition, $M \geq 2f + 1$ reporting sources per constraint, although its guarantee also requires the margin condition $\beta\sigma_k^i \geq \varepsilon$, which held in a minority of our cells. The released artifact contains the attack code needed to reproduce our results; it perturbs a cost report inside a simulator and includes no tooling for reaching a deployed system. No human subjects, personal data or deployed systems were involved.

REPRODUCIBILITY STATEMENT

All theoretical claims are stated with explicit assumptions in Sections 3–5 and proved in Appendix A. Every campaign reported here is governed by a gates document written before its first run, and every quantitative claim in Sections 7.1–7.5 is traceable to an analysis artifact in the released repository; Appendix C gives per-seed values, exact metric definitions and the artifact map. The dose-response campaign of Section 7.4 and the topology campaign of Section 7.5 are later extensions governed the same way, not part of the original A1–A3 pre-declarations, which excluded them by name. P1's final adjudication thresholds were fixed before its last run completed, and its verdicts are reported in Section 7.5 as they were computed: the monitor's detection claim under $W = I$ is refuted, and the localization claim and the signs of the $W = I$ cost and violation effects are unsupported. The environment deviations of Section 6, the disabled reliability weighting, the degenerate margin at $M = 3$ and the minority `guarantee_in_force` frequency at $M = 5$ are also documented in Appendix C. Three seeds per condition do not support inference, so every campaign is analyzed descriptively; the repository enforces this in code with a five-seed floor for inferential analysis. An anonymized artifact accompanies the submission and contains the attack and defense implementations, the environment and oracle evaluator that computes the withheld ground-truth cost, every run configuration, the per-run seed and round logs and consolidated reports, all gates documents, the analysis scripts that produce each reported contrast, and a single script that re-derives every quantitative claim in this paper from those artifacts. Trained policy checkpoints are *not* included, for size reasons; every number reported here is recomputed from the logged records rather than from a checkpoint.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Eitan Altman. *Constrained Markov Decision Processes*. CRC Press, 1999. URL https://www.routledge.com/Constrained-Markov-Decision-Processes/Altman/p/book/9780849303821.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565. URL https://arxiv.org/abs/1606.06565.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 118–128, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html.

Alexander Bukharin, Yan Li, Yue Yu, Qingru Zhang, Zhehui Chen, Simiao Zuo, Chao Zhang, Songan Zhang, and Tuo Zhao. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68121–68133, 2023. doi: 10.52202/075280-2979. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d6f8517fceeca1e2cd61721dff786c14-Abstract-Conference.html.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017. doi: 10.1145/3154503. URL https://doi.org/10.1145/3154503.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8103–8112, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/4fe5149039b52765bde64beb9f674940-Abstract.html.

Hamza Fawzi, Paulo Tabuada, and Suhas Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014. doi: 10.1109/TAC.2014.2303233.

Martin Figura, Krishna Chaitanya Kosaraju, and Vijay Gupta. Adversarial attacks in consensus-based multi-agent reinforcement learning. In *American Control Conference (ACC)*, pp. 3050–3055, 2021. doi: 10.23919/ACC50511.2021.9483080.

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://iclr.cc/virtual_2020/poster_HJgEMpVFwB.html.

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

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, 2024. doi: 10.1109/TPAMI.2024.3457538. URL https://doi.org/10.1109/TPAMI.2024.3457538.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. URL https://openreview.net/forum?id=ryvlRyBKl.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18964–18993, 2023. doi: 10.52202/075280-0831. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3c557a3d6a48cc99444f85e924c66753-Abstract-Datasets_and_Benchmarks.html.

Ali Kahe and Hamed Kebriaei. A distributed primal-dual method for constrained multi-agent reinforcement learning with general parameterization. *IEEE Transactions on Automatic Control*, 2026. doi: 10.1109/TAC.2026.3691607. URL https://doi.org/10.1109/TAC.2026.3691607.

Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. TrojDRL: Evaluation of backdoor attacks on deep reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2020. doi: 10.1109/DAC18072.2020.9218663. URL https://doi.org/10.1109/DAC18072.2020.9218663.

Heath J. LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4): 766–781, 2013. doi: 10.1109/JSAC.2013.130413.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4213–4220, 2019. doi: 10.1609/aaai.v33i01.33014213. URL https://ojs.aaai.org/index.php/AAAI/article/view/4327.

Simin Li, Jun Guo, Jingqiao Xiu, Ruixiao Xu, Xin Yu, Jiakai Wang, Aishan Liu, Yaodong Yang, and Xianglong Liu. Byzantine robust cooperative multi-agent reinforcement learning as a bayesian game. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/86c408308bec72ecce4d86edf243634c-Abstract-Conference.html.

Yao Liu, Peng Ning, and Michael K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):1–33, 2011. doi: 10.1145/1952982.1952995. URL https://doi.org/10.1145/1952982.1952995.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8767–8775, 2021. doi: 10.1609/aaai.v35i10.17062.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. *Mathematics of Operations Research*, 50(2):1277–1304, 2025. doi: 10.1287/moor.2021.0202.

Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. doi: 10.1109/TAC.2008.2009515.

Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013. doi: 10.1109/TAC.2013.2266831.

Thomy Phan, Lenz Belzner, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, and Claudia Linnhoff-Popien. Resilient multi-agent reinforcement learning with adversarial value decomposition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11308–11316, 2021. doi: 10.1609/aaai.v35i13.17348. URL https://ojs.aaai.org/index.php/AAAI/article/view/17348.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2817–2826. PMLR, 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching in reinforcement learning via environment poisoning attacks. *Journal of Machine Learning Research*, 22(210):1–45, 2021. URL https://www.jmlr.org/papers/v22/20-1329.html.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning, 2019. URL https://cdn.openai.com/safexp-short.pdf.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-MDPs: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020. doi: 10.48550/arXiv.2010.04870. URL https://arxiv.org/abs/2010.04870.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143, 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

Lili Su and Nitin H. Vaidya. Fault-tolerant multi-agent optimization: Optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 425–434, 2016. doi: 10.1145/2933057.2933105.

Shreyas Sundaram and Bahman Gharesifard. Distributed optimization under adversarial nodes. *IEEE Transactions on Automatic Control*, 64(3):1063–1076, 2019. doi: 10.1109/TAC.2018.2836919.

Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6215–6224. PMLR, 2019a. URL https://proceedings.mlr.press/v97/tessler19a.html.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/pdf?id=SkfrvsA9FX.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. RLHFPoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2551–2570. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.140. URL https://aclanthology.org/2024.acl-long.140/.

Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. BACKDOORL: Backdoor attack against competitive reinforcement learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3699–3705, 2021. doi: 10.24963/ijcai.2021/509. URL https://www.ijcai.org/proceedings/2021/509.

Zhaoxian Wu, Han Shen, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized policy evaluation with linear function approximation. *IEEE Transactions on Signal Processing*, 69:3839–3853, 2021. doi: 10.1109/TSP.2021.3090952. URL https://doi.org/10.1109/TSP.2021.3090952.

Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *arXiv preprint arXiv:2209.08025*, 2022. doi: 10.48550/arXiv.2209.08025. URL https://arxiv.org/abs/2209.08025.

Chenlu Ye, Jiafan He, Quanquan Gu, and Tong Zhang. Towards robust model-based reinforcement learning against adversarial corruption. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56982–57017. PMLR, 2024. URL https://proceedings.mlr.press/v235/ye24a.html.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659, 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

Donghao Ying, Yunkai Zhang, Yuhao Ding, Alec Koppel, and Javad Lavaei. Scalable primal-dual actor-critic method for safe multi-agent RL with general utilities. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36524–36539, 2023. doi: 10.52202/075280-1586. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/72a1ec14aed36985ffba175e0bba3fec-Abstract-Conference.html.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624, 2022. doi: 10.52202/068431-1787. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

Lei Yuan, Ziqian Zhang, Ke Xue, Hao Yin, Feng Chen, Cong Guan, Lihe Li, Chao Qian, and Yang Yu. Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11753–11762, 2023. doi: 10.1609/aaai.v37i10.26388. URL https://ojs.aaai.org/index.php/AAAI/article/view/26388.

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, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/f0eb6568ea114ba6e293f903c34d7488-Abstract.html.

Kaiqing Zhang, TAO SUN, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10571–10583. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/774412967f19ea61d448977ad9749078-Paper.pdf.

Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11225–11234. PMLR, 2020c. URL https://proceedings.mlr.press/v119/zhang20u.html.

A Proofs

A.1 Proof of Proposition 1

Write $u = W\lambda^\star + \eta_\lambda g^\star$, so the fixed-point condition is $\lambda^\star = \Pi_{[0,\lambda_{\max}]}(u)$ coordinate-wise. For each i exactly one of three cases holds: $\lambda^{i\star} = u_i$ (interior), or $\lambda^{i\star} = 0$ with $u_i \le 0$, or $\lambda^{i\star} = \lambda_{\max}$ with $u_i \ge \lambda_{\max}$. The third is excluded by hypothesis, so $\lambda^{i\star} \ge u_i$ for every i. Summing and using $\mathbf{1}^\top W = \mathbf{1}^\top$,

$$\mathbf{1}^\top\lambda^\star \;\ge\; \mathbf{1}^\top u \;=\; \mathbf{1}^\top W\lambda^\star + \eta_\lambda \mathbf{1}^\top g^\star \;=\; \mathbf{1}^\top\lambda^\star + \eta_\lambda \mathbf{1}^\top g^\star,$$

hence $\eta_\lambda \mathbf{1}^\top g^\star \le 0$ and so $\mathbf{1}^\top g^\star \le 0$. Since $g^{i\star} = \hat{J}_C^i - d^i$, dividing by N gives $\frac{1}{N}\sum_i \hat{J}_C^i \le \bar{d}$. If $\lambda^\star$ is interior then $\lambda^\star = u$, the displayed inequality is an equality, and $\mathbf{1}^\top g^\star = 0$; rearranging $\lambda^\star = W\lambda^\star + \eta_\lambda g^\star$ gives $(I - W)\lambda^\star = \eta_\lambda g^\star$.

For the second claim, that this scalar equation is the only restriction, let the graph be connected. Then W and $W^\top$ are both doubly stochastic and irreducible, so $\ker(I - W) = \ker(I - W^\top) = \mathrm{span}(\mathbf{1})$ and therefore $\mathrm{range}(I - W) = \ker(I - W^\top)^\perp = \mathrm{span}(\mathbf{1})^\perp$. Consequently $(I - W)\lambda = \eta_\lambda g^\star$ is solvable precisely when $\mathbf{1}^\top g^\star = 0$, and its solution set is $\lambda_p + \mathrm{span}(\mathbf{1})$. Choosing the free constant c in $\lambda_p + c\mathbf{1}$ places the solution strictly inside $[0, \lambda_{\max}]^N$ whenever $\lambda_{\max}$ exceeds the spread of λ_p, so the single scalar equation $\mathbf{1}^\top g^\star = 0$ is the whole of what an interior dual fixed point requires; the $N - 1$ remaining degrees of freedom in $g^\star$ are unconstrained.

For $W = I$ the mixing term vanishes and $u = \lambda^\star + \eta_\lambda g^\star$, so the same case analysis applies coordinate-wise: $g^{i\star} = 0$ at interior coordinates and $g^{i\star} \le 0$ at coordinates clamped to zero. Hence $\hat{J}_C^i \le d^i$ for every i separately, which is the per-agent condition. □

Numerical check. Iterating the implemented dual update to convergence on heterogeneous instances $J_C^i(\lambda) = a_i - b_i\lambda^i$ reproduces this exactly: for `ring`, `complete`, `star` and $W = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ at $N = 6$, the fixed point satisfies $\mathbf{1}^\top g^\star \le 0$ with equality at interior fixed points ($|\mathbf{1}^\top g^\star| < 10^{-11}$) while individual agents sit well above and below d; for $W = I$ every agent satisfies $\hat{J}_C^i \le d$. The mean multiplier trajectory agrees with scalar dual ascent on $J_{C,\mathrm{avg}} - \bar{d}$ to 10^{-15} across all of these W. This is a numerical check of the proposition's algebra, not an experiment.

A.2 Proposition 6 and its proof

Proposition 6 (Transfer to the average constraint)**.** Assume *Assumption 1 holds separately for each of the N constraint owners, owner i having its own parameters (M_i, f_i, ς_i) and hence its own $\varepsilon_i \triangleq \varepsilon(M_i, f_i, \alpha; \varsigma_i)$ from Eq.* (5)*; write $\bar{\varepsilon} \triangleq \frac{1}{N}\sum_i \varepsilon_i$.* Then *with probability at least $1 - N\alpha$, $|\frac{1}{N}\sum_i \hat{J}_C^i - J_{C,\mathrm{avg}}(\theta_k)| \le \bar{\varepsilon}$; and if $\beta\sigma_k^i \ge \varepsilon_i$ for every i, then $\frac{1}{N}\sum_i \bar{J}_C^i \le \bar{d}$ implies $J_{C,\mathrm{avg}}(\theta_k) \le \bar{d}$ with the same probability. With common parameters this reduces to $\bar{\varepsilon} = \varepsilon(M, f, \alpha)$.*

Apply Theorem 5 to each owner and take a union bound: with probability at least $1 - N\alpha$, $|\hat{J}_C^i - J_C^i(\theta_k)| \le \varepsilon_i$ simultaneously for all i. The union bound is Boole's inequality and needs no relationship between owners, so cross-owner dependence, which is present in our runs (Section 7.3), does not affect it. On that event the triangle inequality gives $\left|\frac{1}{N}\sum_i \hat{J}_C^i - J_{C,\mathrm{avg}}(\theta_k)\right| \le \frac{1}{N}\sum_i |\hat{J}_C^i - J_C^i(\theta_k)| \le \bar{\varepsilon}$. If additionally $\beta\sigma_k^i \ge \varepsilon_i$ for every i, then $\bar{J}_C^i = \hat{J}_C^i + \beta\sigma_k^i \ge J_C^i(\theta_k)$ for every i by the second statement of Theorem 5, and averaging preserves the inequality. Hence $\frac{1}{N}\sum_i \bar{J}_C^i \le \bar{d}$ implies $J_{C,\mathrm{avg}}(\theta_k) \le \bar{d}$. □

Heterogeneous owners. Allowing (M_i, f_i, ς_i) to differ matters in deployment, where owners may be instrumented differently, but it does not weaken the requirement: $M_i \ge 2f_i + 1$ must still hold *at every owner*, since one owner below its own threshold leaves ε_i, and therefore $\bar{\varepsilon}$, undefined. All runs in this paper use common parameters.

A.3 Proof of Theorem 2

Part (a). Under Eq. (2), the corrupted and uncorrupted dual iterates evolve, while both remain interior to $[0, \lambda_{\max}]$, as $\lambda_{k+1} = W\lambda_k + \eta_\lambda(g_k + \delta_k)$ and $\lambda_{k+1}^\circ = W\lambda_k^\circ + \eta_\lambda g_k$. Subtracting gives $e_{k+1} = We_k + \eta_\lambda\delta_k$ with $e_0 = 0$, since both trajectories share an initialization and, by construction, the same primal sequence $\{\theta_k\}$ and hence the same residuals g_k. Unrolling, $e_K = \eta_\lambda\sum_{k=0}^{K-1} W^{K-1-k}\delta_k$. Left-multiplying by $\mathbf{1}^\top$ and using $\mathbf{1}^\top W = \mathbf{1}^\top$, hence $\mathbf{1}^\top W^m = \mathbf{1}^\top$ for every $m \ge 0$ by induction, gives $\mathbf{1}^\top e_K = \eta_\lambda\sum_{k=0}^{K-1}\mathbf{1}^\top\delta_k$, which contains no dependence on W. □

Part (b), the projected case. Let $\tilde{e}_k$ solve the unprojected recursion $\tilde{e}_{k+1} = W\tilde{e}_k + \eta_\lambda\delta_k$, $\tilde{e}_0 = 0$, so $\mathbf{1}^\top\tilde{e}_K = \eta_\lambda\sum_k \mathbf{1}^\top\delta_k$ by the argument above. Suppose W is entrywise nonnegative, $\delta_k \le 0$ coordinate-wise (under-reporting), and $e_k \ge \tilde{e}_k$ coordinate-wise, which holds at $k = 0$. Writing $u = W\lambda_k + \eta_\lambda(g_k + \delta_k)$ and $u^\circ = W\lambda_k^\circ + \eta_\lambda g_k$, entrywise nonnegativity of W gives $u \ge u^\circ + \tilde{e}_{k+1}$. The coordinate-wise projection $\Pi_{[0,\lambda_{\max}]}$ is monotone and 1-Lipschitz, so $\Pi(u) \ge \Pi(u^\circ + \tilde{e}_{k+1}) \ge \Pi(u^\circ) + \tilde{e}_{k+1}$, the last step because $\tilde{e}_{k+1} \le 0$ and $\Pi(a) - \Pi(a + b) \le |b|$. Hence $e_{k+1} \ge \tilde{e}_{k+1}$, and by induction $\mathbf{1}^\top e_K \ge \eta_\lambda\sum_k \mathbf{1}^\top\delta_k$. Monotonicity also gives the upper bound: if $e_k \le 0$ then $u \le u^\circ$, because $W \ge 0$ and $\delta_k \le 0$, so $e_{k+1} = \Pi(u) - \Pi(u^\circ) \le 0$; since $e_0 = 0$, $\mathbf{1}^\top e_K \le 0$. The mirror argument with $\delta_k \ge 0$ reverses both inequalities. The injected mass therefore bounds the realized bias in magnitude, with no W in the bound. The realized amount itself can depend on W, because projection withholds bias on any coordinate held at a rail: at the floor $\lambda = 0$, where the multiplier is fully relaxed and further under-reporting buys the adversary nothing, or at the ceiling in both arms, where the clean multiplier is pinned at $\lambda_{\max}$ and the under-report does not move the attacked one below it. Which coordinates reach a rail depends on how W distributes the bias; Section 7.5 reports how often the rails bind in P1.

A.4 Proof of Proposition 3

With $\delta_k = \delta u_j$ for all k, Theorem 2 gives $e_K = \eta_\lambda \delta \sum_{m=0}^{K-1} W^m u_j$. Since W is doubly stochastic and the graph is connected and aperiodic, $W^m u_j \to \frac{1}{N}\mathbf{1}$, and $\|W^m u_j - \frac{1}{N}\mathbf{1}\|_2 \le \sigma_2(W)^m$. Writing $r_K = \sum_{m=0}^{K-1}(W^m u_j - \frac{1}{N}\mathbf{1})$ yields $e_K = \eta_\lambda \delta[\frac{K}{N}\mathbf{1} + r_K]$ with $\|r_K\|_2 \le \sum_{m=0}^{K-1}\sigma_2(W)^m \le (1-\sigma_2(W))^{-1}$, bounded uniformly in K. For $W = I$ we have $\sigma_2 = 1$, the series does not converge, and the sum is Ku_j, recovering the concentrated case. □

A.5 Proof of Proposition 4

The agent-i Lagrangian gradient is $\nabla_{\theta^i}\mathcal{L}^i = \nabla_{\theta^i} J_R - \lambda^i \nabla_{\theta^i} J_C^i$. Under a reward corruption replacing $J_R(\theta)$ by $J_R(\theta) + \epsilon(\theta)$, the gradient error is $\nabla_{\theta^i}\epsilon$, which vanishes whenever ϵ does not depend on θ^i, in particular for a constant offset to the return. A constant *per-step* reward offset qualifies only when the expected discounted horizon does not depend on θ^i; with policy-dependent termination, which our testbed has, it need not. Under a cost corruption producing dual bias $e^i = \lambda^i - \lambda^{i\circ}$, the gradient error is $-(\lambda^{i\circ} + e^i)\nabla_{\theta^i} J_C^i + \lambda^{i\circ}\nabla_{\theta^i} J_C^i = -e^i \nabla_{\theta^i} J_C^i$, whose norm is $|e^i| \cdot \|\nabla_{\theta^i} J_C^i\|$, nonzero for any $e^i \neq 0$ at any θ where the constraint gradient is nonzero. □

A.6 Proof of Theorem 5

The argument adapts the coordinate-wise trimmed-mean analysis of Yin et al. (2018) to a scalar constraint estimate; we claim no novelty for the technique and reproduce it for completeness. Let $\mathcal{H}$ index the honest sources, $|\mathcal{H}| \ge M - f$, each reporting $\tilde{J}_C^{i,m} = J_C^i + \xi_m$ with ξ_m independent, zero-mean and sub-Gaussian with parameter ς. Trimming the f largest and f smallest values leaves $M - 2f$ retained sources. Two error sources contribute. First, the retained honest values form an average of at least $M - 2f$ sub-Gaussian variables, giving, with probability at least $1 - \alpha/2$, a deviation of at most $\varsigma\sqrt{2\log(2/\alpha)/(M-2f)}$. Second, each corrupted value that survives trimming is bounded between the f-th smallest and f-th largest honest order statistics, since at most f values are corrupted; a union bound over the honest sources bounds every honest deviation $|\xi_m|$ by $\varsigma\sqrt{2\log(2M/\alpha)}$ with probability at least $1 - \alpha/2$, and at most f corrupted values each displace the retained mean by at most this deviation divided by $M - 2f$. Summing and applying a union bound gives Eq. (5). The second statement is immediate: on that event, $\bar{J}_C^i = \hat{J}_C^i + \beta\sigma_k^i \ge J_C^i - \varepsilon + \varepsilon = J_C^i$ whenever $\beta\sigma_k^i \ge \varepsilon$, so $\bar{J}_C^i \le d^i$ implies $J_C^i \le d^i$. □

A.7 Proposition 7 and its proof

Proposition 7 (Liveness under over-reporting)**.** *Under Assumption 1, arbitrary positive-direction corruption can inflate the RCE estimate by at most $\varepsilon(M, f, \alpha)$, with probability at least $1 - \alpha$, and drive the multiplier to at most $\lambda_{\max}$, so RCE performs a policy update every round with excess conservatism bounded by $\lambda_{\max}\|\nabla_\theta J_C^i\|$. By contrast, if costs are nonnegative, a defense that updates only when $\hat{J}_{\text{mean}}^i + \beta\sigma_k^i \le d^i$, where $\hat{J}_{\text{mean}}^i$ is the untrimmed mean of the M reports, can be made to reject every update for all k by one corrupted source over-reporting by $B > Md^i$.*

The inflation bound is the upper half of Eq. (5) applied to positive-direction corruption: trimming removes the f largest reports, so at most f corrupted values survive and each is bounded by the honest range, giving inflation at most $\varepsilon(M, f, \alpha)$ regardless of the magnitude of δ. An unbounded over-report is trimmed away, so the adversary's per-round budget B does not appear in the bound. The multiplier update is executed unconditionally and projected onto $[0, \lambda_{\max}]$, so $\lambda_k^i \le \lambda_{\max}$ for all k and a policy update occurs every round, with conservatism bounded by $\lambda_{\max}\|\nabla_\theta J_C^i\|$. For the gating comparison, consider a defense updating only if $\hat{J}_{\text{mean}}^i + \beta\sigma_k^i \le d^i$. With nonnegative costs every report is nonnegative, so one corrupted source over-reporting by B gives $\hat{J}_{\text{mean}}^i \ge B/M$; if $B > Md^i$ the gate fails at every round and no update is ever performed. The same argument applies to any aggregate that gives the corrupted report a weight $w > 0$, once $wB > d^i$; RCE trims the f largest reports, so a large over-report receives weight zero. □

This is an analytical result about two update rules, and a design implication rather than an empirical finding: no run in this paper uses a gated update or a positive-direction adversary, so the denial-of-learning failure mode is derived, not observed.

B THE CORRUPTED SAFETY-FEEDBACK LOOP

Figure 1 draws the pipeline described in Section 3: the path the learner closes its loop through, the point at which the adversary writes, and the two quantities a standard evaluator observes.

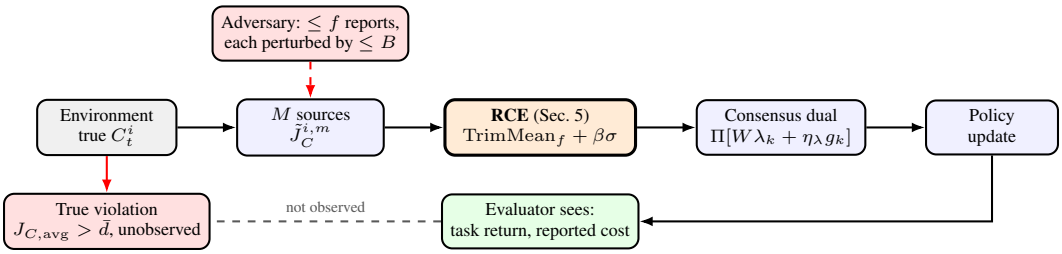


Figure 1: The corrupted safety-feedback loop. Arrows show data flow. Honest sources estimate the true cost from rollouts in the environment, but the dual sees it only through the reports, of which the adversary perturbs up to f; the evaluator observes only task return and reported cost (green), so the true violation that the resulting policy produces in the environment (red) goes unobserved (dashed). RCE (orange) is the only component that limits the influence of extreme reports; without it, corruption reaches the dual unfiltered, where, in the unprojected fixed-primal recursion, Proposition 3 describes how the resulting bias is distributed across the dual coordinates (Section 7.5 measures how far that carries into the realized multipliers).

C EXPERIMENTAL DETAILS AND PER-SEED VALUES

Algorithm 1 Robust Constraint Estimation (RCE) for constraint owner i

Require: budget d^i, source count M, trim parameter f, margin β, dispersion floor $\sigma_{\min}$, dual step η_λ, cap $\lambda_{\max}$, mixing matrix W

1: Initialize $\lambda_0^i \leftarrow 0$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Collect return-scale reports $\{\tilde{J}_C^{i,m}(\theta_k)\}_{m=1}^M$
4: $\mathcal{T} \leftarrow$ indices retained after discarding the f largest and f smallest reports $\{|\mathcal{T}| = M - 2f\}$
5: $\hat{J}_C^i \leftarrow \mathrm{mean}\big(\{\tilde{J}_C^{i,m}\}_{m\in\mathcal{T}}\big)$
6: $\sigma_k^i \leftarrow \max\big(\mathrm{MAD}(\{\tilde{J}_C^{i,m}\}_{m\in\mathcal{T}}),\ \sigma_{\min}\big)$ {degenerate at $|\mathcal{T}| = 1$}
7: $\bar{J}_C^i \leftarrow \hat{J}_C^i + \beta\,\sigma_k^i$ {pessimistic, return-scale}
8: $\lambda_{k+1} \leftarrow \Pi_{[0,\lambda_{\max}]}\big[W\lambda_k + \eta_\lambda(\bar{J}_C - d)\big]$ {network step; no gating}
9: $\theta_{k+1}^i \leftarrow \theta_k^i + \eta_\theta \nabla_{\theta^i}\big[J_R - \lambda_{k+1}^i J_C^i\big]$
10: **end for**

Lines 3–7 run at each owner i; line 8 is the network step, which mixes all owners' multipliers using the vector $\bar{J}_C = (\bar{J}_C^1, \ldots, \bar{J}_C^N)$; and line 9 writes each agent's policy step with exact gradients, which the implementation estimates with PPO. The implementation also carries per-source reliability weights shaping the trimming order; they are **disabled in every run reported in this paper**, no result depends on them, and they are described here only because the released code contains them. Line 9 uses λ_{k+1}^i because the implementation updates the multiplier before the policy step within each round; Eq. (2) writes the simultaneous form, and the one-step offset changes no fixed-point statement. Line 6 makes the $M = 3$, $f = 1$ degeneracy of Section 7.3 explicit: at $|\mathcal{T}| = 1$ the MAD is zero and σ_k^i collapses to $\sigma_{\min}$. The implementation also flags any retained set of fewer than three values as degenerate: at $|\mathcal{T}| = 2$ the MAD, $|a-b|/2$, is positive but rests on two values. This design rule, not a property of the MAD, is why A3 uses $M = 5$, the smallest M with $|\mathcal{T}| \geq 3$ at $f = 1$.

C.1 INSTRUMENT VALIDATION (G10) GATE VALUES

The six scientific gates are aggregate calibration against a withheld reference (G10-A), source precision (B), learning health (C), dual dynamics (D), the joint false-safe rate (E) and constraint control (F); the nine structural checks (S1–S9) verify source counts, that every source ran the current policy, seed-stream disjointness within and across seeds, that no source trajectory or oracle value reaches training, and the dual arithmetic. Source calibration: the owner-mean bias of the aggregate against the withheld reference, pooled over the 15 checkpoints of the three seeds, is $+0.108$, 95% CI $[-0.318, +0.534]$, against a pre-declared bar of 0.6; per-checkpoint $|z|$ is at most 1.361 against a bar of 3.0. Dual wiring: the logged mixed multipliers equal W times the previous ones for the ring W to 1.8×10^{-15}, and each unclipped update equals η_λ times the residual to 8.7×10^{-16} over 1,388 unclipped cells, with λ positive on 92.5% of cells. Learning health: task return improves $-197.1 \to -44.0$ (gain 153.2 against a bar of 75.0), PPO KL median 0.0023 and p95 0.0054 inside bars of 0.0069 and 0.0155, and the fraction of cells at $\lambda_{\max}$ is exactly 0.0. Seed independence: each run issues 23,100 distinct (env, torch) seed pairs, 22,500 to the M=3 reporting sources (250 rounds $\times$ 3 sources $\times$ R_m=30 rollouts) and 600 to the withheld validation reference (5 rounds $\times$ R_{ref}=120), with 0 duplicates within a run; comparing the 22,500 reporting-source seeds across training seeds gives 0 shared values in all three pairings, against a chance expectation of 0.236. Source independence: the between-*source* variance ratio $\text{mean}(s^2_{\text{between}})/\text{mean}(\hat{\varsigma}^2/R_m)$ is 1.085, 0.862 and 0.971 across the three seeds, inside the pre-declared band $[0.80, 1.25]$, which is consistent with Assumption 1(ii); that assumption concerns an owner's sources, not the relation across owners. `scientific_misses` is empty on all three seeds.

Cross-owner dependence and effective sample size. The six constraint owners share one rollout, so their per-round J_C values are strongly dependent: pairwise correlations over the 250 rounds of clean seed 0 run from 0.836 (minimum) through 0.931 (median) to 0.998 (maximum). No result in this paper requires them to be independent: Assumption 1(ii) is within-owner, and Proposition 6 uses a union bound, which holds for arbitrarily dependent events. The consequence is on precision, not validity: the $\approx$1,500 owner-cells of a run carry at most roughly 250 effective observations, fewer once autocorrelation across rounds is counted, any bootstrap treating cells as independent is correspondingly too narrow, and a "k of 6 owners" statement is one observation rather than six. The analysis scripts record this note alongside the affected gate values.

C.2 METRIC DEFINITIONS

J_C is the whole-run network-average *true* discounted cost return: an oracle evaluator, with read-only access to the uncorrupted simulator signal and no path to the learner, computes each agent's true cost return every round, and J_C averages it over agents and over all 250 rounds; higher is worse; the budget is $d = 25$. "Reported" is the reported cost return J_C^{rep}, the value consumed by the dual update (the aggregate under the mean aggregator, the pessimistic estimate $\hat{J}^i_C + \beta\sigma^i_k$ under RCE), averaged the same way. The mechanism gap is $\Delta_{\text{mech}} = J_C - J_C^{\text{rep}}$; positive means the report understates the realized cost, negative that it overstates it. The violation rate is the fraction of (round, agent) cells with true cost above d. Contrasts are paired by seed under common random numbers; CIs are paired 95% t-intervals on $n = 3$ (2 degrees of freedom), and an equivalence check asks whether the paired 90% t-interval lies inside the margin; d_z is the paired mean over the paired standard deviation, reported for scale only.

Why not the own-critic gap. An alternative diagnostic, the *own-critic gap* $J_C - \hat{J}_C^{\text{own critic}}$ between true cost and each agent's own cost-critic estimate, measures critic estimation error rather than corruption of the reports. Under generalized advantage estimation (GAE) it carries a condition-independent offset of ≈ 7.5 in every condition, clean ones included, and the attack moves it by only $+0.97$ ($B - A$, a 13% increment on that offset), against $+4.22$ on Δ_{mech}, whose clean baseline is $+0.21$. Both quantities are logged on every oracle row of every run; the recomputation of Δ_{mech} from the logs matches both the identity $\Delta_{\text{mech}} = J_C - J_C^{\text{rep}}$ and the value fed to the dual in the round log at a maximum absolute error of 0.0.

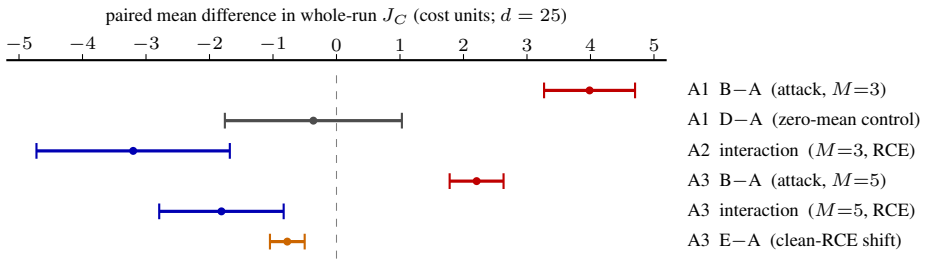


Figure 2: Paired contrasts on whole-run network-average true cost; three seeds, common random numbers, bars paired 95% CIs and dots observed means. The directional attack moves true cost with all three seeds agreeing in sign (red); the zero-mean perturbation of equal root-mean-square size on the same source does not (gray). The RCE interaction is negative in both blocks (blue), and at M=5 the margin also shifts the *clean* RCE condition conservatively (orange). The M=3 and M=5 contrasts share an axis here but are not comparable with one another. The unconstrained control is off scale at +57.64. Per-seed values: Tables 2–3.

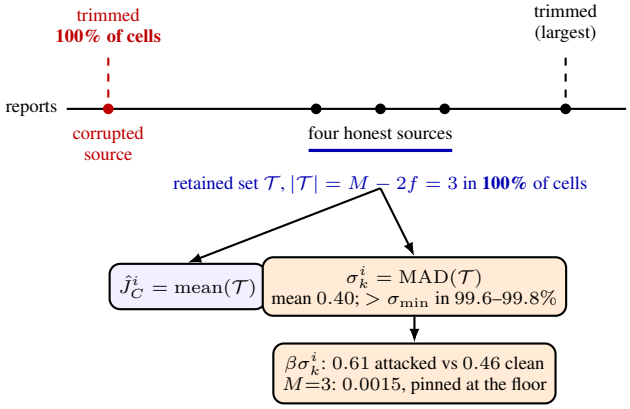


Figure 3: Why A3 uses $M = 5$. Trimming the f largest and f smallest of M reports leaves $|\mathcal{T}| = M - 2f$. At M=3, f=1 this is a single value, its MAD is zero, and the margin collapses to the floor $\sigma_{\min}$, so A2 tests only the trimming/median special case (Section 7.3). The implementation treats fewer than three retained values as degenerate, so M=5 is the smallest M it accepts as live at f=1. At M=5, f=1 three values are retained, the MAD is a live measurement in 99.6% or more of cells, and the applied margin is roughly 400× larger and responds to corruption. All annotated values are measured over 1,500 cells per run in three seeds.

C.3 Per-seed values and paired contrasts

Dose-response detail. The two additional doses appear in Table 1 on the same whole-run basis as every other condition; their per-seed values and adjacent-dose increments are in Tables 2–3. One consistency check is worth recording separately: the ratio of the observed ΔJ_C to the injected mean-aggregate shift B/M lies between 0.77 and 1.05 at every dose and in every seed. That checks the injection arithmetic against the realized effect; it is not a test of Theorem 2, whose statement is about $\mathbf{1}^\top e_K$ in the unprojected recursion rather than about a realized cost difference.

C.4 Attack-injection identity (A1)

The attack subtracts B from one of M sources before mean aggregation, so a mean aggregate must shift by exactly B/M. Measured on the run logs, mean(pre-hook sources) − post-hook aggregate = $B/M = 4.1\overline{6}$ on 100% of cells of every attacked A1 run, at a maximum absolute error of 9.8×10^{-15} against a tolerance of 10^{-9}; the corresponding clean identity ($= 0$) holds on 100% of cells of every clean run at a maximum absolute error of exactly 0.0. For condition D, which draws a fresh zero-mean perturbation per cell rather than applying a constant, the distributional check gives per-

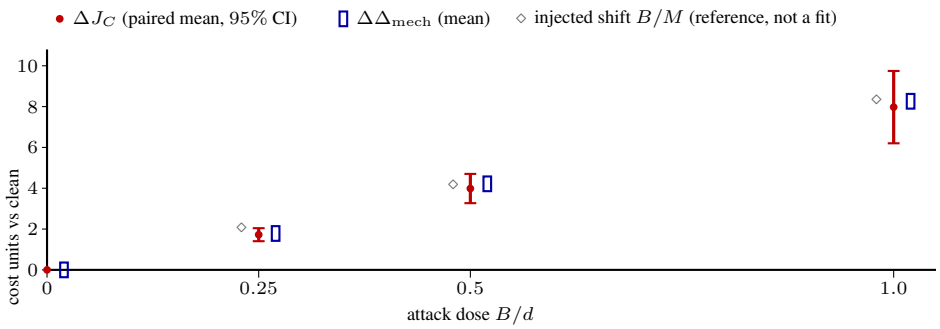


Figure 4: Attack-dose response, M=3, f=1, three seeds paired under common random numbers; clean ($B/d = 0$) is the matched reference, zero by construction. Red: paired mean change in whole-run network-average true cost, 95% CIs; blue: mean change in Δ_{mech}; gray diamonds: the injected shift B/M (a reference value, *not* a fitted curve). Series are offset horizontally for legibility; per-seed values in Tables 2–3.

Table 2: Per-seed whole-run J_C (network-average true cost) for the conditions of Table 1. Campaigns: G10, clean validation; A1, attack and zero-mean control; A2 and A3, RCE at M=3 and M=5; U, unconstrained control; DR, dose response, whose B/d=0.5 arm is A1's condition B.

Condition	Campaign	seed 0	seed 1	seed 2	mean (sd)
A clean	G10 (M=3)	25.2143	24.9550	24.9782	25.0492 (0.1435)
B attack	A1 (M=3)	29.0559	29.2728	28.7761	29.0349 (0.2490)
D zero-mean	A1 (M=3)	25.4654	24.4589	24.1307	24.6850 (0.6955)
dose B/d=0.25	DR (M=3)	26.9116	26.8194	26.5901	26.7737 (0.1656)
dose B/d=1.0	DR (M=3)	33.9986	32.6530	32.4185	33.0234 (0.8527)
E clean + RCE	A2 (M=3)	25.6805	25.3636	24.4701	25.1714 (0.6277)
C attack + RCE	A2 (M=3)	26.4550	25.8098	25.5997	25.9548 (0.4457)
U unconstrained	U (M=3)	81.6027	81.5089	84.9516	82.6878 (1.9611)
A clean	A3 (M=5)	25.5352	25.2017	24.9966	25.2445 (0.2715)
B attack	A3 (M=5)	27.5446	27.5085	27.3001	27.4511 (0.1319)
E clean + RCE	A3 (M=5)	24.6657	24.5479	24.1964	24.4700 (0.2440)
C attack + RCE	A3 (M=5)	25.2442	24.6360	24.7147	24.8650 (0.3308)

seed means of the implied perturbation of $+0.127/+0.160/-0.374$ against a four-standard-error tolerance and standard deviations $12.849/12.250/12.159$ against the declared $\sigma = 12.5$ (ratios $1.028/0.980/0.973$, band $[0.85, 1.15]$), consistent with a live control at the declared scale.

C.5 P1: CONCENTRATED PERTURBATION AND TOPOLOGY

The nine P1 runs are `R_seed{0,1,2}` (attacked, ring), `I_seed{0,1,2}` (attacked, $W = I$) and `IC_seed{0,1,2}` (clean, $W = I$); the ring arm's clean reference is the three CRN-paired clean ring runs of G10. Every run uses the G10/A1 architecture with only three changes: `attack.corrupted_owner_ids` is restricted to `agent_5`; `topology.name` is changed for the I and IC arms; and the attack is removed for IC. The attacked owner was selected by a rule fixed in advance (the owner with the largest clean-run mean multiplier, hence furthest from the projection floor), computed on those clean runs, which give per-agent mean λ of $(1.40, 1.08, 1.11, 1.45, 1.92, 2.09)$, $(1.46, 1.10, 1.14, 1.56, 2.08, 2.25)$ and $(2.04, 1.71, 1.75, 2.14, 2.63, 2.79)$ and select `agent_5` in all three seeds.

Three measurement conventions matter when reading Table 4. *Unprojected* quantities are obtained by driving $e_{k+1} = We_k + \eta_\lambda \delta_k$ with the δ_k measured on the run's own logs; they are properties of that linear recursion, not of the projected run, and $\text{conc}(e_K) = 1$ under $W = I$ is an algebraic consequence of the coordinates decoupling rather than a measurement of the multipliers. *Realized* quantities are differences of logged multipliers between an attacked run and its CRN-paired clean run at the same topology. And P1's cost and violation figures are computed over the final 50 rounds,

Table 3: Paired contrasts, three seeds, paired under common random numbers. "Signs" records how many of the three seeds agree with the mean's sign ("mixed": one or two). Interaction is $(C-E)-(B-A)$. A dash in the d_z column marks the dose range, a pre-declared gate quantity for which d_z is not computed, and $U-A$ violation rate, whose paired standard deviation is near zero.

Campaign	Contrast	Mean	95% CI	d_z	Signs
On whole-run J_C					
U (M=3)	U − A	+57.639	[+52.611, +62.666]	+28.48	3/3
A1 (M=3)	B − A	+3.986	[+3.269, +4.702]	+13.82	3/3
A1 (M=3)	D − A	−0.364	[−1.758, +1.030]	−0.78	mixed
A1 (M=3)	B − D *(descriptive)*	+4.350	[+2.703, +5.997]	+6.20	3/3
A2 (M=3)	C − E	+0.783	[−0.066, +1.632]	+2.29	3/3
A2 (M=3)	E − A	+0.122	[−1.236, +1.480]	+0.22	mixed
A2 (M=3)	**interaction**	**−3.202**	**[−4.725, −1.680]**	−5.22	3/3
A3 (M=5)	B − A	+2.207	[+1.782, +2.631]	+12.92	3/3
A3 (M=5)	C − E	+0.395	[−0.269, +1.059]	+2.56	3/3
A3 (M=5)	E − A	−0.775	[−1.048, −0.501]	−7.03	3/3
A3 (M=5)	**interaction**	**−1.812**	**[−2.792, −0.832]**	−4.59	3/3
Attack dose, on whole-run J_C; each dose paired against the same clean runs					
DR (M=3)	B/d=0.25 − clean	+1.725	[+1.405, +2.044]	+13.42	3/3
DR (M=3)	B/d=0.5 − clean	+3.986	[+3.269, +4.702]	+13.82	3/3
DR (M=3)	B/d=1.0 − clean	+7.974	[+6.202, +9.746]	+11.18	3/3
adjacent-dose increments (the ordering evidence; 9/9 cells positive)					
DR (M=3)	0 → 0.25	+1.725	[+1.405, +2.044]	+13.42	3/3
DR (M=3)	0.25 → 0.5	+2.261	[+1.845, +2.678]	+13.48	3/3
DR (M=3)	0.5 → 1.0	+3.988	[+1.910, +6.067]	+4.77	3/3
DR (M=3)	range 1.0 vs 0.25	+6.250	[+4.448, +8.051]	—	3/3
On the mechanism gap Δ_{mech}					
A1 (M=3)	B − A	+4.217	[+3.558, +4.875]	+15.91	3/3
A1 (M=3)	D − A	−0.016	[−1.170, +1.139]	−0.03	mixed
DR (M=3)	B/d=0.25 − clean	+1.783	[+1.276, +2.290]	+8.74	3/3
DR (M=3)	B/d=1.0 − clean	+8.261	[+6.503, +10.018]	+11.68	3/3
A2 (M=3)	**interaction**	**−3.502**	**[−5.077, −1.928]**	−5.53	3/3
A3 (M=5)	B − A	+2.333	[+2.102, +2.565]	+25.05	3/3
A3 (M=5)	E − A	−0.820	[−1.285, −0.356]	−4.39	3/3
A3 (M=5)	**interaction**	**−1.764**	**[−3.027, −0.500]**	−3.47	3/3
U (M=3)	U − A	+0.112	[−1.002, +1.225]	+0.20	mixed
Other primary outcomes					
A1 (M=3)	B − A, violation rate	+0.170	[+0.126, +0.214]	+9.61	3/3
A1 (M=3)	B − A, reported cost	−0.231	[−0.327, −0.136]	−4.64	3/3
A1 (M=3)	B − A, task return (last 50)	+0.531	[−5.696, +6.757]	+0.21	mixed
A1 (M=3)	D − A, task return (last 50)	−4.706	[−9.153, −0.258]	−2.63	3/3
A3 (M=5)	E − A, task return (last 50)	−1.781	[−38.809, +35.247]	−0.12	mixed
U (M=3)	U − A, violation rate	+0.534	[+0.518, +0.549]	—	3/3

not on the whole-run basis used in Table 1; the two are not interchangeable. Concentration ratios are scale-free and non-monotone under saturation, because a coordinate pinned at a rail in both runs has zero displacement and drops out of the ℓ_1 norm, which raises the ratio mechanically; Table 4 therefore reports realized behavior as an ℓ_1 ratio and a displacement share rather than as a concentration, and carries the saturation fractions beside them. Holding the per-coordinate magnitude equal to A1's necessarily changes the total injected mass ($\mathbf{1}^\top \delta_k = -4.167$ here against -25.0 in A1), so no P1 effect *size* can be compared against A1's.

Why the monitor claim is refuted. Under $W = I$ the clean-calibrated τ is $25.0 = \lambda_{\max}$ (the full range of λ), because the clean $W = I$ fleet is already dispersed enough that no smaller threshold achieves a 5% false-positive rate; since s_k cannot exceed $\lambda_{\max}$, neither arm can fire there. At every τ on the pre-declared grid $\{0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$ both arms fire on 96–100% of rounds and the attacked run fires no more often than clean; in seeds 0 and 2 its mean fleet statistic is lower than clean (17.62 against 22.30; 17.46 against 22.12). The $W = I$ test therefore had no power, because of the monitor's design and the topology rather than anything the attack did, and the pre-declared

Table 4: P1 per-seed values. M=3, f=1, $B/d = 0.5$, mean aggregation, no defense; concentrated perturbation on `agent_5`; $\mathrm{conc}(x) = \|x\|_\infty / \|x\|_1$ ($1/N = 0.167$ is uniform, 1 a single coordinate). Unprojected quantities are computed from the measured δ_k, which is identical in every seed, so they are not repeated measurements. Realized ℓ_1 ratios compare final-round vectors. The fleet statistic is $s_k = \max_i |\lambda_k^i - \mathrm{median}_j \lambda_k^j|$, averaged over rounds; detection excess is attacked-minus-clean firing frequency against the CRN-paired clean run at the same topology. Ceiling-arrival delay is the round at which the attacked owner's multiplier first reaches $\lambda_{\max}$ in the attacked run minus the same round in its clean pair ("never": the attacked run never reaches it, against round 26 in its clean pair). Cost and violation rows are attacked − clean over the final 50 rounds. All values are descriptive at $n = 3$; no contrast in this table is inferential.

Quantity	**seed 0**	**seed 1**	**seed 2**
Injected perturbation δ_k, measured, identical in both topologies			
$\mathrm{conc}(\delta_k)$, every round	1.000	1.000	1.000
untargeted entries, every round	0.0	0.0	0.0
Unprojected dual recursion e_K at $K = 250$, computed from the measured δ_k			
$\mathrm{conc}(e_K)$, ring	0.1725 (every seed)		
$\mathrm{conc}(e_K)$, $W = I$	1.000 (every seed)		
$\mathbf{1}^\top e_K$, ring	−36.458 (every seed)		
$\mathbf{1}^\top e_K$, $W = I$	−36.458 (every seed)		
$\|r_K\|_2$, ring (bound 3)	0.270 (every seed)		
Realized (projected) multipliers, attacked − CRN-paired clean			
$\|e\|_1/\|e_K\|_1$, ring	0.024	0.088	0.013
$\|e\|_1/\|e_K\|_1$, $W = I$	0.038	1.176	0.461
attacked owner whole-run mean $\|e\|$, $W = I$	4.396	4.980	5.343
its share of the whole-run mass, $W = I$	0.607	0.340	0.358
largest per-agent mean $\|e\|$, $W = I$	4.396	8.350	8.038
ceiling-arrival delay (rounds), $W = I$	119	56	never
Topology baseline: clean runs only, no adversary in either arm			
mean fleet statistic s_k, clean $W = I$	22.30	18.86	22.12
mean fleet statistic s_k, clean ring	0.669	0.745	0.703
clean $W = I$ cells at floor / ceiling	.396/.125	.522/.067	.322/.203
rounds interior in both arms (of 250), $W = I$	30	29	26
Median-deviation monitor, detection excess over the clean run			
$W = I$, maximum over the pre-declared τ grid	+0.000	+0.000	+0.000
ring, at the calibrated τ	−0.052	−0.052	−0.052
$W = I$ outcomes, final 50 rounds (signs inconsistent; reported unsupported)			
network-average true cost	−1.434	+9.256	−2.255
violation rate	−0.015	+0.112	−0.104

claim that the monitor catches the concentrated attack under $W = I$ was adjudicated *refuted*. Under the ring the clean run fires on 5.2% of rounds at its calibrated τ and the attacked run on 0.0%, as predicted for a near-uniform bias. This is the campaign's central negative result, and it concerns transfer rather than the theory: the monitor was the pre-declared operational test of Proposition 3, but it reads a statement about e_K across to the multipliers an operator can see, and here that reading did not survive projection and a topology-dependent baseline.

Why this does not contradict Proposition 1. That proposition characterizes a fixed point at which *no* coordinate is clamped at $\lambda_{\max}$, and it constrains the cost estimate $\hat{J}_C^i$. Neither condition holds in these runs: 6.7–20.3% of clean $W = I$ multiplier cells sit at the ceiling and 32–52% at the floor, and the per-agent quantity P1 reports is the oracle's *true* cost rather than the cost estimate $\hat{J}_C^i$ the proposition bounds; on that measure, all six agents exceed d in one of the three clean $W = I$ seeds. What P1 measures is a projected, still-moving $W = I$ trajectory, not the per-agent guarantee its unclamped fixed point would carry.

Attribution of the off-target displacement. Under $W = I$ the dual coordinates decouple exactly, so the dual channel reaches only `agent_5`; any displacement on another coordinate is primal feedback through the shared environment, not dual propagation. Whole-run per-agent mean

$|e|$ under $W = I$ is $(0.24, 0.22, 0.11, 0.38, 1.90, 4.40)$, $(0.11, 0.10, 0.07, 1.01, 8.35, 4.98)$ and $(0.15, 0.12, 0.10, 1.17, 8.04, 5.34)$, the attacked owner being the last entry. The clearest single case is `agent_4` in seed 1, which receives exactly zero injected perturbation and carries the largest displacement in the fleet: its final-50 mean true cost is 17.16 in the clean run and 27.09 in the attacked one, above $d = 25$. The two runs differ only in the corruption of one owner's reports, yet a different agent ends over budget. This is one seed, and we have not measured how far two $W = I$ runs diverge under a perturbation unrelated to the attack, so we do not attribute the size or direction of this displacement to the attack's content. Claims about the injected perturbation ($\mathrm{conc}(\delta_k) = 1$, $\mathrm{conc}(e_K) = 1$) and claims about realized displacement are therefore different claims and can disagree, as they do here.

C.6 RCE DIAGNOSTICS

Figure 3 summarizes the diagnostics below. At $M = 3$, $f = 1$ (A2): retained set size 1 in every cell; dispersion pinned at $\sigma_{\min} = 10^{-3}$ in every cell; applied margin $\beta\sigma = 1.5 \times 10^{-3}$, i.e. 0.006% of d; the point estimate equals the median of the three post-attack reports at a maximum absolute error of 0.0 over 250 rounds $\times$ 6 owners $\times$ 3 seeds.

At $M = 5$, $f = 1$ (A3), per seed: retained set size 3 in 100% of cells; fraction of cells with dispersion strictly above $\sigma_{\min} = 0.996/0.998/0.996$; fraction at the floor $= 0.0$; mean applied margin under attack $= 0.631/0.575/0.613$ and when clean $= 0.450/0.459/0.478$, an attacked-to-clean ratio of $1.40/1.25/1.28$; offline ε calibration $= 0.863/0.967/0.783$, strictly above $\sigma_{\min}$ in every seed; corrupted source trimmed in 100% of cells and the retained set equal to the three lowest honest sources in 100% of cells, so RCE also discards the largest honest report in every cell; `guarantee_in_force` true in $0.269/0.179/0.300$ of attacked cells and $0.145/0.117/0.203$ of clean cells. The aggregation map reproduces every logged aggregate quantity (point estimate, spread, retained count, applied margin, pessimistic estimate, reported cost, constraint residual) at a maximum absolute error of 0.0 over $1{,}500$ cells per run, with the recorded attack shift -12.5. A within-run counterfactual replaying each attacked A3 cell through the mean aggregator gives a mean-aggregator shift of -2.50 by construction (B/M) against RCE's measured shift of $-0.284/-0.267/-0.292$, a suppression factor of $8.8/9.4/8.6$ at the aggregation step. That factor compares shifts in the estimate itself, whereas the 82.1% of Section 7.3 compares closed-loop effects on true cost, so the two need not agree.

C.7 HYPERPARAMETERS AND SOFTWARE

PPO clip 0.2, $\gamma = 0.99$, $\lambda_{\mathrm{GAE}} = 0.95$, policy and critic learning rate 3×10^{-4}, 4 epochs, 4 minibatches, entropy coefficient 10^{-3}, value coefficient 0.5, gradient clip 0.5, hidden width 64. Environment: `ManySegmentAnt` 6×1 on the Farama MaMuJoCo backend (MuJoCo 3.12.0, `gymnasium-robotics` 1.4.2), per-agent cost 1 when the agent's torso speed exceeds the calibrated threshold 0.75 (simulator units) and 0 otherwise, budget $d = 25$. Dual: $\eta_\lambda = 0.035$, $\lambda_{\max} = 25$ (and $\lambda_{\max} = 0$ in the unconstrained control), Lagrangian controller, ring topology with Metropolis weights ($1/3$ on each agent and on each of its two neighbors) except in P1's $W = I$ arms. Constraint estimator `mc_window`: each source reports an explicit discounted Monte Carlo sum of sampled costs over its rollout window, with no critic. Attack: the corruption is injected after a source produces its legitimate value and before aggregation, with the corrupted source rotating by seed. RCE: $\beta = 1.5$ is the only margin setting run. Because ε depends on the unknown ς, `guarantee_in_force` compares $\beta\sigma_k^i$ with a reference $\hat{\varepsilon}$, the 95th percentile of the retained-set MAD over a dedicated 20-round attack-free run of the same configuration and seed ($\alpha = 0.05$).

RCE deployment assumptions. Two assumptions bound the threat model rather than the mathematics. $\hat{\varepsilon}$ is calibrated on an attack-free run, so an adversary already writing on the channel during calibration is outside what we evaluate. And RCE is per-owner by design, trimming corruption before an owner's estimate reaches the average: robustifying only the averaged residual would operate on a scalar in which a corruption confined to one owner is already blended with $N-1$ honest signals.

Why not the reference Safe MAMuJoCo configuration. Three properties of the suite of Gu et al. (2023) rule it out for this study. The first two force the change: it does not implement `ManySegmentAnt`, the scenario whose 6×1 factorization gives the $N = 6$ per-agent structure we need, and its safety cost is a single *shared* torso-velocity indicator, whereas Section 3 requires

a per-agent C^i for the per-agent and network-average conditions to differ at all. The third leaves a choice: its thresholds leave the constraint non-binding at $d = 25$, so the dual would sit at zero and nothing about constraint enforcement could be measured. Lowering d would also have made the constraint bind; we kept the pre-specified budget $d = 25$ and instead calibrated the per-agent threshold by measurement to 0.75. As a result, no number in this paper is comparable to a published Safe MAMuJoCo score.

Rollout window versus episode boundary. Each round collects a fixed window of 2000 consecutive environment steps, for the learner and for every source rollout alike; 250 rounds therefore give $500{,}000$ PPO environment steps per run. Episodes inside that window are *shorter*: the backend builds `ManySegmentAnt` through `gymnasium`'s `Ant-v5` registration, which applies a `TimeLimit` of 1000 steps, and MuJoCo additionally terminates on an unhealthy state. The wrapper resets internally on either signal and flags the step, so a 2000-step window spans at least two episodes and GAE bootstraps through a time-limit truncation rather than zeroing it.

C.8 Source architecture and compute

Each constraint owner's M reports come from parallel trajectory-batch sources, each collecting $R_m = 30$ rollouts of the current policy on a disjoint RNG stream in a separate process, with periodic validation rounds against a larger withheld reference batch ($R_{\text{ref}} = 120$ at rounds $\{25, 75, 125, 175, 225\}$). Sequential source collection was tested during instrument development and failed: when the batches are collected one after another the policy changes between them, so the sources estimate the costs of different policies; in the configuration we tested, collecting them in parallel under one fixed policy removed that drift. The architecture dominates compute. Source environment steps per round are $M \times R_m \times 2000$, i.e. $180{,}000$ at $M = 3$ and $300{,}000$ at $M = 5$, against $2{,}000$ learner steps; over 250 rounds that is 45M ($M = 3$) or 75M ($M = 5$) source steps against 500k PPO steps per run, as recorded in each run's `env_steps` metadata. This ratio, roughly 90–$150 : 1$, is what bounds the number of conditions and seeds the study contains. Wall-clock time is roughly 3.5–10 hours per run depending on M and host.

Platform and software state. The M=3 campaigns (G10, A1, A2, the unconstrained control, the dose sweep and all three P1 arms) ran on Windows 11 under CPython 3.11.9; A3 (M=5) ran on Apple Silicon macOS under CPython 3.11.15. A cross-platform probe shows which quantities the platform changes: source seed streams are pure integer arithmetic and are *identical* across the two platforms, whereas `torch` network initialization and hence policy checksums and source values are not (round-0 maximum absolute source difference 0.886). A3 is therefore not bitwise comparable to any M=3 run, which is one reason Table 1 keeps the two blocks apart. Each run records its own commit in `run_metadata.json`; campaigns are fixed by pre-run configuration verification against their gates document rather than by a single commit, and those verification artifacts are stored alongside the runs.

C.9 ARTIFACT MAP

Claim	Artifact
Instrument validation (G10)	`results/runs_constraint_batch_g10/g10_report.json`
Constraint-binding control (U)	`results/runs_unconstrained/unconstrained_report.json`
Corruption effect + control (A1)	`results/runs_a1/a1_report.json`
Robust mitigation (A2)	`results/runs_a2/a2_report.json`
Full RCE, live margin (A3)	`results/runs_a3/a3_report.json`
Attack dose response	`results/runs_dose/dose_report.json`
Concentrated perturbation (P1)	`results/runs_p1/p1_consolidated.json`; clean $W = I$ arm and final adjudication in `results/runs_p1_clean_i/p1_final_consolidated.json`
Mechanism gap Δ_{mech}	`results/mechanism_gap_reanalysis.json`
Pre-declarations	`docs/{g10,a1_attack,a2_rce,a3}_gates.md`, `docs/unconstrained_control.md`, `docs/dose_response_gates.md`, `docs/p1_concentrated_attack_gates.md`, `docs/p1_clean_identity_control.md`

C.10 THE UNTESTED DESIGN SPACE

Everything listed here is outside the evidence reported in this paper. None of it was run, and no number, figure or table anywhere in the paper should be read as bearing on it. We list it explicitly so the boundary of the study is unambiguous.

Threat model. $f \geq 2$; the $2f \geq M$ breakdown regime; adaptive corruption chosen as a function of (θ_k, λ_k); Byzantine equivocation, where a source reports differently to different consumers; selective rather than persistent support; positive-direction (over-reporting) attacks and hence the empirical side of Proposition 7; more than one attacked owner; and any attacked owner other than `agent_5` in P1.

Defense. A β sweep ($\beta = 1.5$ is the only margin setting run, so the margin's sensitivity is unmeasured); RCE at any dose other than $B/d = 0.5$; per-source reliability weighting, which is implemented but disabled in every run reported here; calibration of ε under corruption rather than on a clean window; and any aggregator comparison against coordinate-wise median, Krum or a trimmed mean as separate arms.

System and environment. Source counts other than $M \in \{3, 5\}$ and the scaling of the guarantee in M; topologies other than the ring and $W = I$, including complete, star and random graphs, and therefore any spectral-gap sweep; other environments, budgets d, agent counts N or cost definitions; and controller families other than the Lagrangian dual of Eq. (2) (no MACPO, PID-Lagrangian or other published safe-MARL algorithm was run, so this paper contains no baseline comparison).

Statistics. More than three seeds per condition. Three seeds do not support inference, so every campaign here is analyzed descriptively (the repository enforces a five-seed floor for inferential analysis in code), and B−D in Section 7.2, pre-declared descriptive, cannot be upgraded by adding seeds after the fact.