# Four Ways to Manufacture a Coordination Finding

## Technical Supplement and Full Forensic Record

A Control Battery for Mixed-Motive MARL

Murad Farzulla[1,2,*]

[1]Dissensus, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

Local canonical technical supplement — 24 August 2026; August evidence snapshot

**Abstract**

Four ordinary-looking design choices—how target coordinates are sampled, how rewards are scaled, what interaction structure is used, and how target support and the performance reference are constructed—produce three misleading coordination findings and, in the fourth case, a misleading attribution of a real effect in mixed-motive multi-agent reinforcement learning (MARL); each is caught by a cheap, targeted control. Across the factorial and its control variants, including one $n = 2$ diagnostic, the record contains 21,360 stochastic training blocks. Four seductive claims appear in turn: theoretical stakes dominate coordination failure ($\eta^2 = 0.74$); nominal $\alpha$ traces a symmetric U-shape in which the unstructured target arm is uniquely worst; the post-control target-correlation estimate looks context-free; and positive target correlation appears to improve learned coordination. Each is the signature of one design choice. The experiment varies signed correlation among target coordinates (ideal points), not objective- or reward-function alignment. *The sampler:* coupling agents through a shared latent realizes cross-agent correlation $\alpha^2/(\alpha^2 + (1 - |\alpha|)^2)$—never negative—so the U-shape is a magnitude profile and the "adversarial" arm never existed. *The reward scale:* apparent stakes dominance is degree-1 homogeneity, and the $\eta^2$ ranking inverts under reward-scale normalization. *The environment:* the separable IQL expectation is correlation-invariant and the VDN slope is not detected. *The measurement construction:* centring the same signed Gaussian inside the feasible box makes the frozen-policy gradient three to four times steeper and turns a legacy-support mixed result into a consistent $\rho = -.33$ versus 0 penalty in the four-agent design, while the static single-state optimum ignores the reachable lattice, finite-horizon transient and exploration. Against an exact dynamic oracle, feasible-centred runs split the positive-target-correlation gradient between continuous feasibility (about 54–63%) and a smaller frozen-policy residual. The battery that catches all four—realized-correlation verification, stake normalization, a signed re-run under both independent learning and a value-decomposition/team-reward bundle, a separable-resource twin, a partial-sharing overlap response, a feasible-target control, and a dynamic-oracle decomposition—is the contribution of record. The artifacts live in the data-generating process, the environment and the measurement, not in the analysis, which is what lets them survive careful statistics. As the battery's case study we test a pre-stated falsification target, the coordination-friction functional $F = \sigma(1 + \varepsilon)/(1 + \alpha)$ (Farzulla, 2026a): the exploratory controls weigh against it as a maintained single-index predictor in this environment, without supplying a confirmatory H4 verdict. The implemented reward coefficient scales reward by construction. A reduced signed control finds that observation noise raises the mean clean frozen gap for both learners, but its penalty grows rather than falls toward more positive target-coordinate correlation—opposite the ratio's implied target-correlation–noise cross-effect and not universal cell by cell. The post-control target-correlation direction concerns both achievability and, conditionally on in-box support, residual frozen-policy performance. Each control is stated as a procedure over an arbitrary mixed-motive sweep; we do not claim to have measured how often it is needed.

**Keywords:** multi-agent reinforcement learning; construct validity; experiment-design artifact; control battery; coordination failure; signed target correlation; ideal-point correlation; value decomposition; shared-state contention; data-generating process; reproducibility; cooperative AI

**Relationship to the reader paper.** This document is the citable full technical record accompanying the 24 August 2026 reader-facing article. It preserves the complete first-stage tables, legacy and repaired control analyses, implementation chronology, alternative protocols, withdrawn secondary outcomes, non-detections and conditional diagnostics. Material is retained here because it remains part of the provenance of the scientific correction. The checked August raw evidence is public in Zenodo v3.0.0; the 24 August bootstrap and protocol-summary delta instead travels with the arXiv ancillary bundle and awaits repository/Zenodo synchronization. Where wording differs, the reader paper supplies the compressed argument and this supplement supplies the fuller forensic record.

**Terminology boundary.** The theoretical construct *alignment* denotes agreement among objectives or reward functions. The empirical treatment $\rho$ denotes signed cross-agent correlation among target coordinates (ideal points) inside one fixed quadratic reward family. This experiment does not directly manipulate or measure the broader construct. Historical first-stage labels such as "alignment $\alpha$" are retained only when describing the nominal, defective design; repaired empirical claims are stated in terms of target-coordinate correlation.

| Stable ID | Exact destination in this technical record |
|---|---|
| `TS-FIRST` | Section "Results I: The Sign-Blind Factorial" and Tables 10–14 (full ANOVAs and model comparisons) |
| `TS-PROXY` | Section "Dependent variables," paragraph "Three secondary proxies, withdrawn," plus Appendix "First-Stage Descriptive Results" |
| `TS-IMPL` | Section "Robustness: GPU/CPU cross-validation," Table 15, and the implementation-defect discussion in "Limitations" |
| `TS-OPP` | Sections "No detected advantage for strong opposition" and "Control: at latent-target $\rho = -1$" |
| `TS-ENDPOINT` | Section "Partial sharing," paragraph "Validity check," and `results/reviewer3_followup/verify_partial_endpoints.py` |
| `TS-COND` | Conditional opposition tables in "Legacy-support training-window gap," "Partial sharing" and "Observation noise" |
| `TS-ORACLE` | Section "Control: dynamic benchmarking," including Table 7 |
| `TS-RELEASE` | "Declarations," especially Data and Code Availability and the reproducibility note |

## 1 Introduction

An experiment can manufacture its own findings. In reinforcement learning the best-known failure modes live in the *analysis*: too few seeds, cherry-picked baselines, uncorrected multiplicity, reporting conventions that flip conclusions (Henderson et al., 2018; Agarwal et al., 2021). This paper is about a subtler family—artifacts manufactured upstream of any analysis, by the data-generating process, environment, and measurement reference point—and about the controls that catch them. We show, in one controlled mixed-motive setting, that four ordinary-looking design choices each produce a qualitatively wrong coordination finding: a target-coordinate sweep coupled through a shared latent silently fails to vary the *sign* of its claimed correlation and manufactures a symmetric U-shape; a reward coefficient makes theoretical stakes look like the dominant factor; a shared-state environment makes a shared-state-bounded target-correlation effect look general; and a benchmark defined as a per-agent unconstrained ideal makes a rising *achievable* optimum look like agents that have learned to coordinate. The four are not a miscellany: they are the sampler, the reward scale, the environment and the reference point—which between them are most of what an experimental design *is*. None of the four is visible in the analysis, all four passed a first reading, and each is caught by a cheap, targeted control. The control battery—not the initial factorial it corrects—is this paper's contribution of record.

A battery needs a target: a concrete, pre-stated quantitative claim whose apparent confirmation the controls can stress. Ours is a proposed coordination-friction functional. The consent-friction framework (Farzulla, 2026a)—a companion formal paper—proposes a *candidate coordinate triple* $(\alpha, \sigma, \varepsilon)$ of alignment, stake, and entropy as candidate coordinates for coordination cost, and proposes a friction functional whose simplest candidate form is

$$F = \sigma \frac{1+\varepsilon}{1+\alpha}. \tag{1}$$

The public arXiv v3 of that companion predates this audit and repeats the superseded U-shape and symmetric quadratic proposal. The current local AOC reader withdraws both and incorporates the control-battery verdict; synchronizing the public record remains a release-only task. The framework is explicit that (1) is a phenomenological ansatz rather than a theorem: it is the simplest expression satisfying its stated boundary and monotonicity desiderata, and its empirical adequacy is left open. As a falsification target it supplies exactly what a validity exercise wants—three manipulable structural variables, four hypotheses stated in advance of estimation, and an adequacy bar pre-specified before estimation ($R^2 > 0.7$, §3.1)—so every artifact the battery catches is caught against claims fixed before the data arrived. The functional is the case study; the battery is the contribution.

**What we do: a factorial baseline and the control battery.** We first run a $5 \times 5 \times 5$ factorial in a four-agent resource-allocation MARL environment with $m = 3$ shared resources, manipulating nominal $\alpha$ through target coordinates, reward scale $\sigma$ (the weight magnitude on resources), and observation-noise variance parameter $\varepsilon$. These are environment-specific proxies, not direct measurements of general alignment, stakes or entropy. Each of the 125 conditions is replicated 30 times with stochastic replication blocks, training Independent Q-Learning (IQL) agents for 1,000 episodes—3,750 training blocks in total—and we compare (1) against three alternative specifications. That sweep returns a striking and, as it turns out, *misleading* result: a symmetric U-shape in $\alpha$, with neutral preferences the worst case. Diagnosing why exposes a flaw in the design, not the theory. Rather than report the U-shape, we treat it as the first entry in a *control battery*—a sequence of targeted re-runs, each isolating one way the ap-

parent finding could be an artifact of the experimental setup rather than a fact about target dependence. (i) A rebuilt, genuinely *signed* target design—an equicorrelation control whose cross-agent correlation $\rho$ is verified to span $[-.33, +0.8]$, plus a two-team block design reaching full opposition ($\rho \to -1$)—re-run for both IQL and a centralized value-decomposition learner (VDN) on two AMD GPUs (§6). (ii) A *separable-resource* control that removes the jointly controlled state, providing a negative endpoint for whether the surviving effect is about target correlation per se or about agents physically contesting one common pool (§6.5). (iii) A parametric *partial-sharing* control that varies the fraction of resources fed into the common pool, testing whether the estimated gradient changes with the shared-coordinate fraction (§6.6). (iv) An $n = 2$ *latent-target strong-opposition* control that reaches a sign-verified target correlation $\rho = -1$, free of the within-team-cooperation confound and the $-1/3$ positive-semidefinite cap of the $n = 4$ equicorrelated design, but carrying an agent-count confound and realizing only $-0.468$ correlation between feasible optima (§6.7). (v) A matched *target-support and dynamic-benchmark* control that freezes the trained policies, evaluates them on held-out reset seeds, centres the same signed target distribution inside the feasible box, and decomposes the clean-evaluation gap against an exact finite-horizon controller (§6.8). This tests jointly whether the surviving effect is about coordination behaviour, out-of-box ideal points, or the reference against which performance is measured. Reward-scale normalization—dividing through by the degree-1 scale factor—runs alongside as a diagnostic rather than a re-run, since it needs no new data. The battery is the result of record; the first sweep is its cautionary baseline. The exact audited count is 21,360 stochastic training blocks: 3,750 in the GPU factorial, 3,750 in the CPU factorial, 900 in the first-stage learner-bundle slice, 6,840 in the early controls and 6,120 in the follow-up controls. These totals span different environment variants and estimands; they are not replications of one homogeneous four-agent task.

**The detour: a sign-blind design manufactures a U-shape.** The first sweep realizes its nominal $\alpha$ proxy by interpolating each agent's target between a base vector *shared across all agents* and an idiosyncratic component, with $|\alpha|$ setting the weight on the shared component (§3.2). The cross-agent target correlation it induces is $\alpha^2/(\alpha^2 + (1 - |\alpha|)^2)$—a function of $|\alpha|$ *alone* that is never negative. Empirically, $\alpha = +0.8$ and $\alpha = -0.8$ are *both* correlated at $\approx +0.94$ (§3.2, Figure 1): the "adversarial" arm of the factorial never existed. The only genuinely unstructured cell is $\alpha = 0$ ($\rho = 0$), which is why it looked uniquely bad. What the first sweep measured was the *magnitude* of an always-cooperative correlation, and the apparent "structure of either sign helps" U-shape was an artifact of a magnitude-only manipulation. The signed-linear component of the nominal-$\alpha$ effect is therefore $\approx 0\%$ of its variance *by construction*—not a discovery about objectives, but a tell that the target-correlation sign was not being varied.

**Headline result: signed target correlation matters under shared-state contention, but its estimated size and channel depend on target support.** The verified signed design removes the U-shape. Here and below, "baseline" in a signed contrast means independent target draws at $\rho = 0$, not motivational or moral indifference. Under the legacy target support, the training-window gap declines on the cooperative side; its equicorrelated opposition–baseline contrasts are individually non-significant and not equivalence results (§6.2). The fully separable IQL expectation is $\rho$-invariant by construction because each agent's marginal problem is invariant to $\rho$; the separable VDN slope is not detected despite team-reward coupling, and equivalence is not established. These are negative endpoints, not by themselves a mechanism test. The paired legacy-Gaussian-at-zero-support partial-sharing series preserves interaction while varying the shared-coordinate fraction under the fixed first-$k$ ordering. Its paired frozen-

policy $\rho \times w$ term detects graded slope modulation with problem-block clustering ($p = 3.7 \times 10^{-6}$ IQL; $p = 2.0 \times 10^{-17}$ VDN). Categorical modulation is also detected, while departure from a linear ramp is not (§6.6).

The held-out target-support control changes both magnitude and attribution. Under legacy out-of-box targets, the clean frozen-policy slope is about $-1.16$ for each learner; continuous feasibility contributes more than the point-estimated gradient, while the residual policy term is not detected ($+.136$, $p = .463$, IQL; $+.130$, $p = .051$, VDN). Absence or equivalence is not established. When the same signed Gaussian is centred inside the feasible state box, the slope steepens to $-3.602$ IQL and $-4.210$ VDN. An exact finite-horizon oracle attributes about 54–63% to continuous feasibility and 37–46% to the residual frozen-policy component. In this feasible-centred design, the four-agent $\rho = -.33$ endpoint is worse than $\rho = 0$ at both reward scales for both learners after Holm correction (§6.8). *The honest core is conditional but substantive*: more positive target correlation lowers the gap when agents contest shared state; the degree of sharing modulates that gradient; and in-box support reveals both achievability and support-dependent frozen-policy channels. "Stakes dominate" (reward scale), the symmetric U-shape (sampler), "correlation alone causes the effect" (interaction structure), and "a training-window/static-reference gap directly measures learned coordination" (target support plus benchmark) are the four inferences the controls overturn. The fourth is an attribution error rather than a nonexistent effect.

**The surviving shape is reproduced under a second learner/objective bundle.** The shared-state-bounded target-correlation effect is not a quirk of independent learning. Re-running the signed sweep under VDN (Sunehag et al., 2018)—a centralized additive value decomposition—preserves the curve (IQL-vs-VDN curve correlation 0.976) and uniformly *lowers* the gap by $\approx 0.6$–$0.9$ at every $\rho$, without manufacturing a U or selectively rescuing the opposition region. The shift is uniform rather than shape-changing—and it is not a fact about centralization, as the next paragraph explains, nor even a general advantage for VDN: once the pool is not shared, VDN's gap sits *above* IQL's at every $\rho$ (§6.5). It does not change the shared-state target-correlation ordering. The separable control is structurally $\rho$-invariant for IQL; the VDN slope is not detected despite team-reward coupling, and equivalence is not established. We state the scope of this replication narrowly: IQL and VDN are *both* value-based, and VDN is the strictly additive, most conservative near-independent member of its family, so what we have shown is replication under an additive value-decomposition/team-reward bundle, not architecture-robustness in the broad sense.

There is a further confound in this comparison that the word "architecture" hides, and it deserves stating rather than burying. IQL trains each agent on its *own* reward; VDN trains on the summed team reward. Swapping one for the other therefore changes two things at once—the value-learning structure *and* the objective the learners are given—and the second change converts a mixed-motive game into a fully cooperative one. That a team-reward learner attains a better team-mean outcome is unsurprising—though our own separable control shows it is not guaranteed, since VDN is worse than IQL at every $\rho$ once agents hold independent pools (4.05 versus 3.90 mean normalized gap, Table 4)—so the uniform $\approx 0.6$–$0.9$ level improvement should not be read as evidence about centralization. What the comparison does support is the claim we make from it: the *shape* over $\rho$ survives the swap, and a qualitative pattern that persists when both the architecture and the training objective change is more robust than one that has only been seen under a single learner. The clean design that would separate the two—crossing architecture with reward information, so that individual-reward VDN and team-reward IQL both exist—we did not run, and we do not claim its result. We are also careful to distinguish two senses of "robust across both learners": the *shape and direction* of the target-correlation effect

hold for IQL and VDN alike, whereas the *multiplicity-corrected significance* of the single cooperation-beats-baseline contrast holds only for VDN, with IQL directionally consistent but not surviving Holm correction (§6.2). Monotonic-mixing CTDE (QMIX) and, more importantly, policy-gradient and other non-value-based methods are untested and left to future work (§8).

**Reward scale dominates only the raw scale.** As a co-result we confirm, on the new data, that theoretical stakes' apparent dominance is a measurement convention. Because reward is homogeneous of degree one in $\sigma$, $\sigma$ rescales the reward landscape without changing its geometry: reward scale carries $\eta^2 \approx 0.23$ of the *raw* gap variance but $\eta^2 \approx 0.0003$ once the gap is normalized by $\sigma$, while the structural correlation effect persists (§6.2; the degree-1 mechanism is set out in §5.6). "Stakes dominate" is degree-1 reward scaling, not a structural fact—we report both partitions and lean on neither.

**Two contributions.** The paper makes a methodological and an empirical contribution, and the methodological one is primary. *Methodological (a control battery):* a worked demonstration that four ordinary-looking design choices each produce a misleading coordination finding—three spurious effects and one spurious attribution. Degree-1 reward scaling makes stakes look dominant; coupling agents through a shared anchor manufactures a symmetric U-shape and invites the false conclusion that "baseline is uniquely worst" or "opposition coordinates as well as cooperation"; a shared-pool environment makes a target-correlation effect appear general even though it is detected only in the tested shared-state variants; and an out-of-box target distribution combined with a training-window outcome and static single-state benchmark distorts both the magnitude and attribution of the signed effect—together with the specific control that catches each (stake-normalization, a verified-signed DGP, a separable/partial-sharing series, and a feasible-target held-out evaluation against an exact dynamic oracle). The artifacts are subtle because they live in the data-generating process, the environment and the measurement, not in the analysis. We do not assert these are universal traps (§7.4); we report them as artifacts of *this* setup that a cheap verification would have caught, and that may be worth checking in mixed-motive (Schelling, 1960) sweeps with similar structure. Every post-diagnostic control is exploratory; its estimates are demonstrations to be independently replicated, not confirmatory discoveries. *Empirical (the case study):* a candid characterization of the consent-friction functional under learning. Its empirical proxies have partial, scope-dependent directional support—reward scale expands raw loss, observation noise raises the mean gap, and more positive target correlation lowers it—but their interaction contradicts the proposed ratio, and the exploratory evidence weighs against its specific monotone form $\sigma(1+\varepsilon)/(1+\alpha)$ as a maintained single-index predictor, and the post-control target-correlation direction is *shared-state-bounded*: it is detected in shared-state variants; the separable IQL expectation is $\rho$-invariant and the VDN slope is not detected, without an equivalence claim. Legacy-support opposition is mixed across outcome protocols and reward scales; in the feasible-centred frozen control the four-agent $\rho = -.33$ endpoint is a Holm-significant penalty at both reward scales for both learners. Neither reading is transported across target supports or evaluation protocols.

**Why falsify one's own functional?** The target functional comes from our own companion framework, and a fair question is whether the exercise is a general lesson or private theory maintenance. The answer is that the functional is an *instance* of a general class: low-dimensional structural indices claimed to predict coordination loss—the same family as price-of-anarchy-style summaries of strategic inefficiency (Koutsoupias and Papadimitriou, 1999; Roughgarden, 2005) and the scalar alignment couplings swept in mixed-motive MARL (Tampuu et al., 2017; McKee et al., 2020). Such indices can be vulnerable to the

four design choices this paper audits; we do not estimate how often. The reusable object is the battery, not the verdict on this particular index: verify the realized sign of a coupled-preference sweep, normalize for reward homogeneity before declaring a factor dominant, preserve a graded interaction while varying the proposed interaction structure, and evaluate frozen policies against a transition-matched oracle under an audited target support. Testing one's own theory also carries one dry epistemic advantage: there is no motivated reasoning toward rescue—we had every incentive to keep the functional's confirming U-shape, and report instead that it was our own design's artifact.

**Positioning and contribution.** We position the paper primarily within the measurement-and-validity literature on (MA)RL evaluation (Henderson et al., 2018; Patterson et al., 2024; Gorsane et al., 2022; Ellis et al., 2023), and situate friction-as-coordination-loss against the price-of-anarchy tradition in algorithmic game theory (Koutsoupias and Papadimitriou, 1999; Roughgarden, 2005) and the Cooperative AI agenda (Dafoe et al., 2020, 2021). The contribution is twofold. *Methodological:* a control battery (and an environment, a verified signed data-generating process, a separable-resource twin, and a codebase) for separating structural coordination effects from target-design, reward-scaling, environment-structure and benchmark-choice artifacts—offered as a cheap check we recommend for similar mixed-motive sweeps, without claiming the error is common (we have not surveyed the literature for it; §7.4). *Empirical:* a reproducible, honestly scoped test of the consent-friction functional—more positive target correlation lowers the shared-state frozen-policy gap; under feasible-centred support this combines raised achievability with a smaller residual to an exact dynamic oracle; the separable IQL expectation is $\rho$-invariant, the VDN slope is not detected, and on paired legacy-Gaussian-at-zero support point estimates steepen over four tested sharing levels; and the shape is reproduced by a VDN/team-reward bundle whose two changes are not isolated. Apparent stakes dominance, the nominal-$\alpha$ U-shape, the appearance that the surviving effect is a context-free correlation law, and the reference point against which the gap is measured are each shown to be artifacts.

## 2 Related Work

**Measurement and validity in (MA)RL evaluation.** That an RL finding can be an artifact of the experimental setup is by now a lineage, with the established results on the analysis side: Henderson et al. (2018) showed that headline deep-RL comparisons hinge on seeds, code-level choices, and reporting conventions; Ilyas et al. (2020) falsified theory-motivated mechanism claims about deep policy gradients with controlled experiments; and Patterson et al. (2024) codify empirical-design practice for RL experiments—our control battery is a worked instance of that program, moved to the environment side. In multi-agent evaluation specifically, Lanctot et al. (2017) and Balduzzi et al. (2018) treat evaluation itself as a measurement problem whose naive estimators fail in agent-interdependent ways; Gorsane et al. (2022) document evaluation heterogeneity across cooperative MARL and propose a standardized protocol; and Ellis et al. (2023) is the closest single precedent for our central move—a flagship benchmark (SMAC) shown to have a construct-validity failure (open-loop policies suffice), triggering a redesign that restores the property the benchmark was assumed to measure. Cardei et al. (2026) argue that return-style metrics conceal the coordination mechanisms they are meant to capture—the same diagnostic genre as our target-support and dynamic-oracle decomposition—and Freiesleben and Zezulka (2025) supply the construct-validity framing for ML evaluation in general. Within this literature our niche is specific: to our knowledge no prior work takes a closed-form, theory-motivated candidate functional, pre-states an adequacy criterion, runs a controlled factorial against it, discovers that its own data-generating process was sign-blind, and reports the resulting artifact taxonomy—with the control that catches each artifact—

as the contribution. The nearest neighbors diagnose analysis-side artifacts (Henderson et al., 2018; Ilyas et al., 2020) or benchmark-side validity failures (Ellis et al., 2023; Gorsane et al., 2022); ours span the data-generating process, environment structure, and measurement reference point in a mixed-motive sweep, with a falsification target supplying claims fixed in advance.

**The algorithmic-collusion critique literature.** The move we make here—take a headline multi-agent result, then ask with controls whether the design manufactured it—has run furthest in algorithmic pricing, and that literature is our closest genre precedent. Calvano et al. (2020) reported that independent Q-learners converge on supracompetitive prices sustained by punishment strategies; most of the work since has audited that result rather than extended it. Abada and Lambin (2023) trace the apparent collusion to imperfect exploration rather than strategic sophistication; den Boer et al. (2026) show the collusive equilibria are reachable only on timescales irrelevant to a firm's objective and only under identical algorithms, identical settings, and synchronized starts; and Murthy and Pandey (2026) reproduce the effect in deep multi-agent RL and find that asynchrony alone removes roughly half of it. The difference from our battery is where the controls point. Theirs are learning-hyperparameter-side—exploration schedule, update synchronization, the discretization of the action grid—with the economic environment held fixed. Ours are preference-DGP-side and environment-structure-side: what correlation the preference sampler actually realizes, and whether the estimate remains after removing the shared pool. The two sets are complementary, and neither catches the other's artifacts. Nearer in title, Tilbury et al. (2024) locate cooperative coordination failure in the algorithm class under a fixed offline dataset, where we locate it in the preference generator and the environment under a fixed algorithm, online and mixed-motive.

**Coordination and independent learning in MARL.** Independent learners that treat co-players as part of a non-stationary environment are the oldest and simplest MARL baseline (Tan, 1993; Buşoniu et al., 2008). Their pathologies—instability, failure to converge, sensitivity to co-player adaptation—are well documented (Hernandez-Leal et al., 2019, 2017; Matignon et al., 2012). Independent learning nonetheless remains a competitive and widely studied baseline: independent PPO matches or beats centralized-critic methods on standard cooperative benchmarks (de Witt et al., 2020), and systematic comparisons across cooperative tasks confirm that the gap between independent and centralized learners is task-dependent rather than uniform (Papoudakis et al., 2021; Albrecht et al., 2024). We deliberately use Independent Q-Learning (Watkins and Dayan, 1992; Tan, 1993): because coordination failure is the dependent variable, an algorithm that engineers coordination away would confound the manipulation. Our contribution is to vary the strategic structure factorially and ask whether a proposed scalar index predicts the magnitude of failure.

**Centralized training and value decomposition.** The standard response to independent-learner non-stationarity is centralized training with decentralized execution (CTDE). Additive value decomposition (VDN) factorizes the team value into per-agent components trained on a shared team reward (Sunehag et al., 2018); QMIX generalizes this to a monotonic mixing network (Rashid et al., 2018), QTRAN relaxes the structural constraint (Son et al., 2019), and Weighted QMIX corrects the sub-optimality of the monotonic projection (Rashid et al., 2020). Policy-gradient CTDE—centralized-critic actor-critic (MADDPG (Lowe et al., 2017)) and multi-agent PPO (Yu et al., 2022)—spans the actor-critic flank. These methods are usually invoked to *suppress* coordination failure. We use the most conservative, near-independent member of the family (VDN) only as a cross-bundle probe: both the value-learning structure and the objective change, because VDN trains on the summed team reward. The comparison

therefore asks whether the curve shape reappears under that bundle; it does not isolate centralization. We leave crossed objective/architecture controls, monotonic mixing and policy-gradient CTDE to future work.

**Mixed-motive learning and the preference axis.** Whether learning agents cooperate or defect is an emergent function of incentive structure rather than a fixed property (Leibo et al., 2017). A line of work sweeps a scalar coupling between agents' objectives—from a cooperation coefficient (Tampuu et al., 2017) to individual-versus-shared mixing weights (Durugkar et al., 2020) to social-value orientation and incentive correlation (McKee et al., 2020). These establish that the *degree* of alignment matters and report monotone transitions between competitive and cooperative regimes—monotone in the signed coupling, broadly the ordering we recover over the cooperative range once the sign is genuinely manipulated. Radke et al. (2023) show that full common interest is not always optimal in team settings, hinting at non-monotonicity; our controlled signed sweep finds no such interior optimum in a shared-pool game—the gap is ordered by signed correlation, lowest at full cooperation. Our distinct contribution is methodological as much as substantive: we show that a preference-axis sweep which couples agents through a *shared* latent realizes a correlation that depends only on the coupling *magnitude* and never goes negative, so it silently tests $|\alpha|$ rather than signed target correlation and can manufacture a spurious symmetric U-shape unless the realized correlation is verified to span the intended sign. The theoretical intuition that strongly structured games are more tractable than weakly structured ones traces to the special status of zero-sum games, whose unique minimax value makes them more learnable than general-sum objectives (Littman, 1994), with competition-as-curriculum evidence in Bansal et al. (2018); our shared-pool result is the opposite-signed boundary case—no legacy endpoint test detects an advantage for opposition, and feasible-centred opposed targets increase the gap because they contend for one shared state. To our knowledge the specific pairing—a *shared-state-bounded* target-correlation effect (detected in the shared arm but not in the separable arm), together with a control battery that diagnoses the magnitude-only design artifact, the degree-1 reward-scale artifact, the shared-state-contention bound, and the unconstrained-benchmark artifact—is not stated in this corpus; Radke et al. (2023), McKee et al. (2020), and the opponent-modeling work below are precedents we position against rather than results we duplicate.

**Learner-generality and non-stationarity.** Non-stationarity is the canonical pathology of independent learning (Hernandez-Leal et al., 2017), and agents that explicitly exploit the structure of others' learning can escape it—learning with opponent-learning awareness engineers such exploitation directly (Foerster et al., 2018). Our signed-correlation shape reappears under VDN (curve correlation 0.976), but that arm jointly changes architecture and the reward objective. It therefore documents recurrence across a second learner/objective bundle, not that centralization or reduced non-stationarity caused the pattern. The value-decomposition arm shifts the level uniformly while leaving the shape intact: the gap falls as target correlation becomes positive, while the legacy opposed endpoint is a non-detection.

**Price of anarchy.** The price-of-anarchy (PoA) literature in algorithmic game theory quantifies coordination loss as the ratio between the social cost of selfish (Nash) equilibrium play (Nash, 1951) and the social optimum (Koutsoupias and Papadimitriou, 1999; Papadimitriou, 2001; Roughgarden, 2005; Nisan et al., 2007). Our reward gap $\Delta R$ is only a loose cousin of that object: it is an absolute, $\sigma$-scaled shortfall against an infeasible per-agent ideal, not a welfare fraction, and much of it is the feasibility floor rather than welfare lost to selfish play (§6.8). We therefore use PoA as motivation, not as a formal mapping.

The PoA tradition establishes *worst-case* bounds for fixed game classes; we instead measure the *learned* coordination loss that emerges when adaptive agents play a parameterized family of games, and ask how that loss varies with structural parameters. The two views are complementary: PoA bounds the equilibrium cost analytically, while our factorial maps how the realized cost moves across the implemented proxy grid. We find that realized coordination loss is ordered by signed target-coordinate correlation, decreasing as that correlation becomes positive. This is not a direct manipulation of objective alignment. It would be natural to read that as evidence that the *learnability* of the strategic structure, and not only its equilibrium efficiency, governs realized coordination loss—and the first-stage version of this paper did. The exact dynamic decomposition (§6.8) qualifies that reading. Under legacy target support, achievability dominates the point-estimated decomposition and the residual is not detected, without establishing a zero residual; under feasible-centred support both achievability and the residual frozen-policy component improve as target correlation becomes positive. The target-correlation axis therefore orders the attainability of a good shared state in both designs and, conditionally on in-box support, also orders performance relative to a matched finite-horizon controller.

**Cooperative AI, evaluation, and external validity.** The Cooperative AI agenda frames the construction of agents that can find common ground—under mixed motives, partial information, and heterogeneous objectives—as a first-class research problem (Dafoe et al., 2020, 2021). Our experiment is a controlled probe of one of its central questions: when does learning suffice for coordination, and when does it fail? The finding speaks to this agenda only within its measured scope: when agents contend for one shared state, more positive target correlation lowers the frozen-policy gap; in the feasible-centred design this combines improved achievability with a smaller residual to the matched dynamic oracle. At the fully separable endpoint the IQL expectation is $\rho$-invariant and the VDN slope is not detected; on paired legacy-Gaussian-at-zero support, point estimates steepen over the four tested shared-coordinate fractions. Methodologically, we follow few-seed RL reliability guidance (Agarwal et al., 2021), reporting a 30-replication factorial with GPU/CPU cross-validation, and position our bespoke controlled environment against recognized mixed-motive evaluation suites (Leibo et al., 2021) as the external-validity target.

**The case-study functional.** The falsification target (Farzulla, 2026a) proposes the coordinate triple $(\alpha, \sigma, \varepsilon)$ as candidate inputs for a resource-allocation configuration and proposes (1) as the simplest friction functional consistent with its desiderata. Its proposed dynamical extension—that lower-friction configurations persist longer under a specified selection process—is a conditional modeling hypothesis developed in the Replicator-Optimization Mechanism (Farzulla, 2026b), not a premise tested here. The present paper does not re-derive that model; it restates the testable predictions so the case study is self-contained (§3.1) and subjects them to the battery.

## 3 Environment and Experimental Design

The design has two components: a falsification target (the case-study functional and its pre-stated criteria, §3.1) and the environment and factorial that probe it (§3.2 onward). Stating the target first fixes, in advance, exactly which apparent findings the control battery of §6 will stress.

### 3.1 The target functional and pre-stated criteria

We recap only what is needed to make the case study self-contained; the formal construction is in the framework paper (Farzulla, 2026a) and the conditional dynamical model in (Farzulla, 2026b).

**The candidate coordinate triple.** The general proposal asks whether coordination difficulty can be organized by three coordinates: *alignment* $\alpha$ (agreement among agents' objectives or reward functions, $\alpha \in [-1, 1]$), *stake* $\sigma$ (how much agents care, $\sigma \geq 0$), and *entropy* $\varepsilon$. The experiment implements $\varepsilon$ only as a Gaussian observation-noise variance parameter; it is neither an amplitude nor a direct entropy measure. A model-derived friction index $F$ is then proposed to follow the composite form (1): increasing in stakes, increasing in entropy, decreasing in alignment, and diverging as $\alpha \to -1$. Its adequacy is tested against the independently measured coordination gap. The experiment instantiates these coordinates more narrowly: $\rho$ is correlation among target coordinates within one fixed quadratic reward family, and $\sigma$ is its reward-scale coefficient. Neither manipulation by itself identifies the broader construct named by the theory.

**Desiderata and functional-form candidates (self-contained statement).** So that the test does not depend on the framework paper, we restate the boundary and monotonicity desiderata that $F$ is required to satisfy, and the competitors they leave open. The framework asks for a nonnegative friction scalar $F(\alpha, \sigma, \varepsilon)$ obeying: (D1) *stake homogeneity*—$F$ scales with $\sigma$, so doubling every agent's stake doubles friction ($F \propto \sigma$, degree-1 in $\sigma$); (D2) *entropy monotonicity*—$F$ is nondecreasing in $\varepsilon$, with $\varepsilon = 0$ (perfect information) the friction-minimizing floor; (D3) *alignment monotonicity*—$F$ is nonincreasing in $\alpha$ ($\partial F / \partial \alpha \leq 0$); (D4) *cooperative vanishing*—friction $\to$ its floor as $\alpha \to +1$ (perfect common interest), which as stated follows from (D3) and so carries no content independent of it: any $F$ nonincreasing in $\alpha$ attains its $\alpha$-infimum there; and (D5) *adversarial divergence*—friction $\to \infty$ as $\alpha \to -1$ (perfect opposition). The canonical ansatz (1) is the *simplest* closed form meeting (D1)–(D5): the $\sigma$ prefactor gives (D1), the $(1+\varepsilon)$ numerator gives (D2), and the $1/(1+\alpha)$ denominator gives (D3)–(D5) simultaneously (it is $1/2$ at $\alpha = +1$, blows up at $\alpha = -1$). It is explicitly a phenomenological choice, not a theorem; several alternatives meet the same desiderata. *Theoretical competitors.* An *additive* combiner $F = \sigma + \varepsilon - \alpha$ (or $\sigma(1 + \varepsilon - \alpha)$) satisfies the monotonicities but not the divergence (D5); a *multiplicative* combiner $F = \sigma\varepsilon(1-\alpha)$ satisfies (D3)/(D5)-style growth but collapses friction to zero whenever $\varepsilon = 0$, violating the idea that misaligned agents conflict even under perfect information; a *symmetric* refinement $F = \sigma(1+\varepsilon)/(1+\alpha^2)$—which an even-in-$\alpha$ empirical shape would seem to motivate—keeps (D1)/(D2) but *breaks* (D3)/(D5), treating cooperation and opposition as equally frictionless. It does *not* break (D4) as we have stated it: its $\alpha$-minimum is $\sigma(1+\varepsilon)/2$, attained at $\alpha = \pm 1$, so it does reach the floor at $\alpha = +1$—it simply reaches the same floor at $\alpha = -1$ as well, which is exactly what (D3) and (D5) rule out. This is the sense in which (D4) is redundant: what separates the symmetric refinement from the canonical form is a *strict* minimum at $\alpha = +1$, not the floor condition we wrote down. A fully *separable* linear model $F = \gamma_0 + \gamma_\alpha \alpha + \gamma_\sigma \sigma + \gamma_\varepsilon \varepsilon$ keeps the directions but abandons the single-index, divergence, and homogeneity structure. We estimate the additive and multiplicative composites and the separable model against the canonical form (M1–M4, §4). The symmetric refinement is not fit: it is the post-hoc form later withdrawn because the even-in-$\alpha$ shape it was built to capture was the sampler artifact (§7.4).

**Hypotheses.** The framework's comparative statics yield three directional predictions and one prediction about functional form. We stated all four in advance of estimation.

**Hypothesis 1** (Theoretical alignment)**.** *Coordination failure decreases with alignment:* $\partial \Delta R / \partial \alpha < 0$.

**Hypothesis 2** (Stakes)**.** *Coordination failure increases with stakes:* $\partial \Delta R / \partial \sigma > 0$.

**Hypothesis 3** (Observation-noise proxy)**.** *Coordination failure increases with the implemented observation-noise proxy:* $\partial \Delta R / \partial \varepsilon > 0$.

**Hypothesis 4** (Functional form)**.** *Measured coordination failure is well predicted by the canonical composite:* $\Delta R \approx \beta_0 + \beta_1 \cdot \sigma(1+\varepsilon)/(1+\alpha)$*, with the composite explaining the bulk of cross-condition variance.*

For H4 we adopted an *adequacy criterion* in advance of the first-stage factorial: the full canonical composite was to be treated as adequate only if it attained $R^2 > 0.7$ on condition-level reward gaps. The first stage crosses all three factors but has an invalid signed axis; the repaired signed grid is post-diagnostic and fixes $\varepsilon = 0$. Consequently neither design supplies a valid confirmatory pass/fail verdict for the full H4 claim. We retain the threshold only as a descriptive calibration for the reduced $\sigma/(1+\rho)$ restriction, not as a hypothesis about a population parameter. Hypotheses H1–H3 concern *direction* and are, as the framework itself notes, invariant to the exact functional form—any specification satisfying the monotonicity conditions generates them. H4 concerns *shape* and is what distinguishes (1) from its competitors.

**Two identification boundaries.** The framework reads $\alpha$ directionally: positive $\alpha$ is cooperative, negative $\alpha$ adversarial, and friction is supposed to diverge as $\alpha \to -1$. The first-stage environment substitutes correlation among agents' target coordinates for that theoretical construct, and its sampler makes even this narrower correlation a function only of $|\alpha|$. The first stage is therefore unidentified twice over: it neither delivers signed target dependence nor establishes that target dependence is equivalent to objective- or reward-function alignment. The repaired design in §6 solves the first problem by manipulating $\rho$ directly and holding marginal target variance fixed. It does not solve the second. The repaired experiment therefore tests the direction only for this target-coordinate instantiation.

### 3.2 Resource-allocation environment

We implement a resource-allocation environment in which $n$ agents coordinate to allocate $m$ shared resources.

**Definition 3.1** (Resource-allocation MDP)**.** The environment is specified by a state space $\mathscr{S} = [0, C]^m$ (resource levels, capacity $C = 10$), $n = 4$ agents, $m = 3$ shared resources, a per-agent action space $\mathscr{A}_i = \{-1, 0, +1\}^m$ (request a decrease, maintain, or request an increase for each resource, so $|\mathscr{A}_i| = 3^m = 27$), transition dynamics that increment each resource by the summed agent requests and clip to $[0, C]$, an independent episode reset $s_0 \sim \text{Uniform}([0, C]^m)$, and an episode length of $T = 100$ timesteps. All agents observe the same underlying resource vector, each corrupted by its own independent observation noise (Definition 3.3).

**Definition 3.2** (Parameterized rewards)**.** Agent $i$'s reward function is

$$R_i(s) = \sum_{j=1}^{m} w_{ij}\, u(s_j - \tau_{ij}), \tag{2}$$

with weight $w_{ij} \in [0, 1]$ (stake on resource $j$), target $\tau_{ij} \in \mathbb{R}$ (preferred level), and quadratic loss $u(x) = -x^2$.

This specification gives direct control over an environment-specific proxy triple, and the exact sampling is load-bearing for the analysis. The implemented reward scale $\sigma$ sets the weight magnitudes:

each $w_{ij} \sim \mathcal{N}(\sigma, 0.05^2)$, clipped to $[0,1]$, so every weight is approximately $\sigma$ and the reward (2) is, to leading order, homogeneous of degree one in $\sigma$ ($R_i \approx \sigma \sum_j u(s_j - \tau_{ij})$); we use this in the effect-size analysis (§5.6). The implemented information-loss proxy $\varepsilon$ is the observation-noise variance parameter (Definition 3.3). Nominal $\alpha$ is implemented through the targets: a base vector $b \sim \mathcal{N}(0, I_m)$, shared by all agents within a run, is drawn afresh for each replication as a sequential draw within its condition stream, and agent $i$'s target is

$$\tau_i = \alpha\, b + (1 - |\alpha|)\, \xi_i, \qquad \xi_i \sim \mathcal{N}(0, I_m) \text{ i.i.d.}, \tag{3}$$

so the cross-agent target correlation is a function of $|\alpha|$ alone,

$$\operatorname{corr}(\tau_i, \tau_j) = \frac{\alpha^2}{\alpha^2 + (1 - |\alpha|)^2} \qquad (i \neq j), \tag{4}$$

ranging from 0 at $\alpha = 0$ to $\approx 0.94$ at $|\alpha| = 0.8$. Two consequences matter for interpretation. First, (4) depends only on $|\alpha|$: the design varies the *strength* of shared target structure, and—because the anchor $b$ is sign-symmetric—the conditions $\alpha = +c$ and $\alpha = -c$ are statistically exchangeable. The factorial therefore does *not* implement a cooperative-versus-adversarial contrast; its negative-$\alpha$ cells are not anti-aligned but carry the same shared structure as their positive counterparts. Second, the degree-1 homogeneity in $\sigma$ means the reward gap scales mechanically with stakes—a fact we exploit, and partly correct for, in §5.6.

**Definition 3.3** (Noisy observations)**.** Agent $i$ observes $\tilde{s}_i = s + v_i$ with $v_i \sim \mathcal{N}(0, \varepsilon I_m)$, so larger $\varepsilon$ corresponds to less accurate information about the true state. The code therefore uses standard deviation $\sqrt{\varepsilon}$: $\varepsilon$ indexes covariance, not amplitude. For $\varepsilon > 0$ the Gaussian differential entropy is monotone in $\log \varepsilon$; $\varepsilon = 0$ is the degenerate no-noise boundary. This is a variance-level observation-noise proxy, not a direct entropy measure.

### 3.3 Factorial design

We manipulate the three factors in a full $5 \times 5 \times 5$ factorial, summarized in Table 9, over the levels

$$\alpha \in \{-0.8, -0.4, 0, 0.4, 0.8\}, \quad \sigma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}, \quad \varepsilon \in \{0, 0.25, 0.5, 0.75, 1.0\}.$$

This yields 125 conditions, each with $k = 30$ replication blocks (3,750 blocks in total). Each replication trains for 1,000 episodes of 100 timesteps, with the final 100 episodes used to compute the reported reward.

**That window is training, not evaluation.** We call it the *final training window* rather than an evaluation phase, because that is what it is: learning updates continue throughout it, exploration remains at the 1% floor, and episodes are not drawn from held-out resets. Any quantity computed on it therefore mixes converged policy quality with residual exploration cost and late-stage learning, and "evaluation" would over-claim. A separate frozen protocol—parameters fixed, greedy actions, independent reset seeds, with noise-free and matched-noise scores—is available in the post-fix factorial and is used for the bounded observation-noise result (§7). The archived signed, separable and partial-sharing CSVs, however, contain only the late-training window; they have no frozen-evaluation columns. Their results must therefore be read as late-training performance, not as direct estimates of learned-policy quality. The completed follow-up runners implement frozen greedy and frozen 1%-exploration evaluations on a common held-out reset bank. Those outputs are public in the checked Zenodo v3.0.0 record 10.5281/zenodo.22004215.

### 3.4 Dependent variables

We operationalize coordination failure through a single dependent variable, the reward gap. Three secondary proxies appeared in earlier versions and are withdrawn below; the analysis never rested on them. **Reward gap** $\Delta R = R^* - \bar{R}$: the shortfall of realized mean payoff $\bar{R} = \frac{1}{T}\sum_t \bar{R}_t$ against the cooperative benchmark $R^*$ (defined below). It is the only dependent variable the analysis uses.

**Three secondary proxies, withdrawn.** Earlier versions of this work reported three further dependent variables—Pareto inefficiency, policy variance, and convergence time—alongside the reward gap. All three are withdrawn here, because an audit of the implementation found that each measured something other than what its definition stated, and a paper whose subject is the gap between a stated construct and its implementation cannot leave that standing in its own instrument.

*Pareto inefficiency* was defined as distance from the Pareto frontier, but the implementation took the frontier to be the non-dominated subset of the 30 observed reward vectors *in that condition*. That set is sample-relative: it moves with the runs drawn, at least one point has distance zero by construction, and the quantity shrinks as replication variance shrinks. It measured dispersion, not inefficiency, and it is not the environment's attainable frontier. *Policy variance* was defined as cross-replication dispersion of evaluation action frequencies, but the implementation evaluated greedy policies on 128 random probe states with the exploration mixture folded in, then aggregated over agents before taking the cross-run variance—a different estimand. *Convergence time* sampled the policy at six points $\{0, 200, 400, 600, 800, 999\}$ and, on failure to converge, returned the number of samples and multiplied by the sampling interval, yielding 1,200 episodes inside a 1,000-episode budget. Since nearly every condition failed to converge under the 1% exploration floor, the measure was both censored and capable of impossible values.

None of this touches the results of record: the analysis rested on the reward gap throughout, and these three were labeled corroborating or weak wherever they appeared. Removing them costs a robustness gesture, not an argument. We note the episode chiefly because it is the same failure mode, one level down—a measurement whose name promised more than its code delivered.

**The cooperative benchmark $R^*$ is an infeasible per-agent ideal, by design.** The per-agent reward (2) is a negated quadratic loss $u(x) = -x^2$, so each agent's reward is bounded above by zero and attains zero exactly when every resource sits at *that* agent's target ($s_j = \tau_{ij}$ for all $j$). Our benchmark is the average of these *per-agent unconstrained optima*,

$$R^* \;=\; \frac{1}{n}\sum_{i=1}^{n} \max_{s\in\mathbb{R}^m} R_i(s) \;=\; 0, \tag{5}$$

in which the maximization is performed *separately for each agent* and—this matters—over $\mathbb{R}^m$ rather than over the state space $\mathcal{S} = [0, C]^m$. The distinction is not pedantic here. Targets are drawn Gaussian about zero (Equation (3)), so roughly half of all target coordinates are negative and therefore lie outside $\mathcal{S}$; for those, the box-constrained per-agent optimum sits at the boundary $s_j = 0$ and is *strictly negative*, not 0. Taking the maximization over $\mathcal{S}$ would thus make (5) false as written and would make $R^*$ depend on the realized targets, defeating the purpose of a constant benchmark. Over $\mathbb{R}^m$ each agent's ideal is 0 by construction, so their average is 0 for *every* condition, independent of the weights $w_{ij}$ and targets $\tau_{ij}$. $R^*$ is accordingly an *ideal*, not an attainable reward—which is exactly the reading the rest of this section relies on, and the reason the finite-horizon benchmark decomposition of §6.8 is needed at all. It

is essential to be precise about what $R^*$ is *not*. It is *not* the achievable joint maximum

$$R^{\text{cont}} \equiv R^{\text{joint}} = \max_{s \in \mathscr{S}} \frac{1}{n} \sum_{i=1}^{n} R_i(s), \qquad R^{\text{cont}} \leq R^* = 0, \tag{6}$$

which optimizes the team average over a *single shared* state $s$. The two coincide ($R^{\text{joint}} = 0$) only when all agents' targets coincide; whenever targets differ—i.e. for any $|\alpha| < 1$, and most severely at $\alpha = 0$—no single shared state can sit at every agent's target at once, so $R^{\text{joint}}$ is *strictly negative*. Our reported gap $\Delta R = R^* - \bar{R} = -\bar{R}$ is therefore measured against an *infeasible* per-agent ideal. Algebraically it can be split into $R^* - R^{\text{joint}}$ and $R^{\text{joint}} - \bar{R}$, but the second term is *not* by itself a learning shortfall: $R^{\text{joint}}$ is a continuous, static single-state optimum, whereas a policy starts from a random reset, moves on an integer-reachable lattice, and accrues reward during the transient. Section 6.8 replaces that coarse residual with matched frozen-policy and exact finite-horizon-oracle components. We use $R^*$ rather than $R^{\text{joint}}$ deliberately: it is a single global constant, so $\Delta R = -\bar{R}$ is an *exact* affine transform of mean reward (identical factorial $F$ and $\eta^2$ for mean reward and the gap, a sign flip and additive constant leaving the variance decomposition unchanged), whereas $R^{\text{joint}}$ would itself vary with $\alpha$ and require a separate per-condition optimization. The cost of this choice is that the gap's *level* is partly the infeasibility floor $R^* - R^{\text{joint}}$, which is largest exactly where targets diverge most (at $\alpha = 0$), so part of the elevated gap at neutral alignment reflects reduced feasibility of the reference, not policy quality alone. This is precisely why we never interpret raw gap *levels* as structural facts: the condition-level ($n = 125$) and per-replication ($n = 3{,}750$) analyses agree on the *ordering* of factor importance but differ in magnitude (the condition-level partition averages out within-condition replication variance and assigns larger $\eta^2$ to the systematic factors; §5.6), and we lead level summaries with standardized differences. The signed and contention controls of §6 compare gaps *across* $\rho$, where the infeasibility floor $R^* - R^{\text{joint}}$ varies smoothly and monotonically with target correlation (largest at opposition, smallest at positive correlation). This confounds the opposition-versus-baseline comparison *in the direction of our own conclusion*: opposition carries the larger floor, so a non-detection on the total gap is close to what the floor alone would predict, and we should not read it as an independent finding. §6.8 separates incompatibility, reachable-grid, transient, exploration/noise and residual frozen-policy components. Meanwhile the positive-target-correlation dividend includes a feasibility component, consistent with the contention mechanism, but its learning interpretation requires the dynamic rather than the static benchmark (§6.8).

### 3.5 Learning algorithm

Agents use Independent Q-Learning, implemented in PyTorch (Paszke et al., 2019): a two-hidden-layer MLP (64 units per hidden layer, ReLU; three linear layers including the output), learning rate 0.001 with Adam (otherwise default settings), discount $\gamma = 0.99$, and plain MSE temporal-difference loss (no gradient clipping, weight decay, or learning-rate schedule). Each agent keeps a replay buffer of capacity 100,000 transitions, sampled in minibatches of 64 (uniformly *with* replacement in the batched GPU implementation whose results are of record; *without* replacement in the CPU cross-validation implementation, a difference we return to in §5.8), and a separate target network updated by a Polyak soft update with coefficient $\kappa = 0.01$ (target $\leftarrow \kappa\,\theta + (1-\kappa)$ target) applied at *every* environment step rather than by a periodic hard copy. The $\varepsilon_{\text{exp}}$-greedy exploration rate (distinct from the entropy/observation-noise candidate coordinate $\varepsilon$) is annealed *linearly* from 0.1 to a 0.01 floor over the first 5,000 agent steps and then held at that floor for the remainder of training; the 0.01 floor never decays to zero. An earlier version of this paper leaned on that fact when interpreting policy-versus-reward stability; §A.1 withdraws the argument, since the stored policy vector is a deterministic greedy blend in which the

exploration term cancels between measurements. IQL is chosen deliberately. Because coordination failure is the quantity we measure, a method engineered to coordinate—MADDPG (Lowe et al., 2017), QMIX (Rashid et al., 2018)—would suppress the very signal of interest and confound the manipulation. The VDN comparison below inherits that objection because it also replaces individual rewards with a team reward; we therefore read it for curve shape, not for its level. The cost is that absolute friction levels are specific to independent learning; we return to this in §8. To test whether the qualitative shape recurs in a second learner/objective bundle—not to isolate the independent-learning choice—we additionally re-run the experiment with a centralized value-decomposition learner (VDN (Sunehag et al., 2018)) on a reduced grid (§5.7).

**VDN specification.** The centralized robustness learner is a Value-Decomposition Network (Sunehag et al., 2018). Each agent $i$ keeps its own per-agent utility network $Q_i(\tilde{s}_i, a_i; \theta_i)$ with the *same* architecture as the IQL agents (two 64-unit hidden layers with ReLU; three linear layers including the output), and the joint action-value is the additive factorization $Q_{\text{tot}}(s, \mathbf{a}) = \sum_{i=1}^{n} Q_i(\tilde{s}_i, a_i; \theta_i)$. Training is centralized on the team return: a single joint temporal-difference target $y = r_{\text{team}} + \gamma \sum_i \max_{a_i'} Q_i(\tilde{s}_i', a_i'; \theta_i^-)$ is regressed against $Q_{\text{tot}}$ under the same MSE loss, so gradients flow into each $Q_i$ through the sum. All remaining hyperparameters—Adam at learning rate 0.001 with otherwise default settings, discount $\gamma = 0.99$, replay capacity 100,000, minibatch 64, the Polyak target-network coefficient $\kappa = 0.01$ applied at every step, and the $\varepsilon_{\text{exp}}$-greedy schedule—are identical to the IQL configuration above. Two things change together: per-agent TD targets become an additive team factorization, and each learner's individual reward is replaced by the summed team reward. The latter turns the training objective from mixed-motive to fully cooperative, so the comparison cannot isolate architecture. Execution remains decentralized (each agent acts greedily on its own $Q_i$), so VDN is the strictly additive, most conservative near-independent member of the CTDE family.

## 4 Analysis Plan and Model Comparison

### 4.1 Competing specifications

We operationalize H4 as a direct horse race between the canonical single index and alternatives that either combine the factors differently or retain them separately. We compare four functional forms for the reward gap, selected by AIC/BIC (Akaike, 1974; Burnham and Anderson, 2002):

1. **M1 (Friction):** $Y = f\big(\sigma(1+\varepsilon)/(1+\alpha)\big)$ — the canonical composite.
2. **M2 (Additive composite):** $Y = g(\sigma + \varepsilon - \alpha)$.
3. **M3 (Multiplicative composite):** $Y = h(\sigma \cdot \varepsilon \cdot (1-\alpha))$.
4. **M4 (Independent effects):** $Y = \gamma_0 + \gamma_\alpha \alpha + \gamma_\sigma \sigma + \gamma_\varepsilon \varepsilon$.

An auxiliary M0 uses reward scale $\sigma$ alone. M1–M3 are single-index models (intercept plus one composite slope, 2 parameters each); M4 is a linear main-effects model in the three raw factors (4 parameters). All four are estimated by ordinary least squares on the 125 first-stage condition means. To avoid inheriting the sign-blind target-correlation artifact, the identifiable candidates are re-estimated on the signed target-equicorrelation data in §5.3 (Table 12). M4 is deliberately a *linear* main-effects model, not a saturated dummy-coded ANOVA: its advantage over M1 is therefore not flexibility in functional form but the refusal to collapse three separately-acting factors into one ratio. M3 is recorded as unidentifiable on that grid because $\varepsilon = 0$ makes its composite constant. If the consent-friction theory's headline composite is correct, M1 should dominate.

### 4.2 Primary analysis family and multiplicity

We designate one *primary analysis family*: the four signed contrasts of §6.2 (opposition-versus-baseline and cooperation-versus-baseline, for IQL and VDN), to which a Holm–Bonferroni correction is applied at family-wise $\alpha_{\text{fw}} = 0.05$ (Table 2). This controls multiplicity within that quartet; it does not turn the battery into confirmatory evidence. The signed redesign and subsequent controls were developed after diagnosing the first-stage artifact, and no immutable pre-run protocol exists for them. Accordingly, every inferential test in the paper—including the primary quartet, factorial ANOVAs, functional-form comparison, team-block and $n = 2$ contrasts, separable and partial-sharing slopes, and benchmark decomposition—is *exploratory*. We report effect estimates and uncertainty, preserve the design timestamps in the repository, and treat replication on an independent run as the gate for stronger claims. After correction the only positive contrast in the designated quartet is VDN cooperation-versus-baseline ($p_{\text{Holm}} = 0.0071$); the IQL direction corroborates it but is not a second corrected result.

## 5 Results I: The Sign-Blind Factorial

**Reading this section.** The results below come from the first-stage design of §3, in which the cross-agent correlation depends only on $|\alpha|$ (4). They are faithful descriptions of *that* design, and we report them in full because the conclusion they invite—a symmetric U-shape with neutral preferences the worst case—is precisely the artifact we dissect and overturn in §6. Read this section as the cautionary baseline: every "U-shape" and every "$|\alpha|$" claim here is a property of a sign-blind manipulation, not of signed alignment. The stakes and observation-noise-proxy directions and the degree-1 stake-scaling result (§5.6) do not depend on the *sign* of $\alpha$, so they are unaffected by the DGP repair, which bears only on the alignment shape; the policy-versus-reward stability finding is likewise a sign-blind descriptive property of the first-stage design. The inequality/Gini pattern (Appendix A, Table A1) is itself $|\alpha|$-governed and we do *not* re-estimate it under the signed DGP—we report it as a first-stage descriptive, not as a claim about signed alignment. What the correction overturns is the alignment *shape*.

### 5.1 Summary

Across all 3,750 replications, mean reward ranged from $-4.568$ (worst condition: $\alpha = 0.0$, $\sigma = 1.0$, $\varepsilon = 0.75$) to $-0.418$ (best condition: $\alpha = 0.4$, $\sigma = 0.2$, $\varepsilon = 0$). Since $R^* = 0$, the reward gap equals the negated mean reward, so this is a ratio of approximately $10.94\times$ in reward-gap magnitude between the worst and best conditions. The grand mean was $-1.745$ (SD 1.285). We report such multiplicative ratios for orientation only: they are sensitive to the $R^* = 0$ benchmark (§3.4), and we lead level effects with standardized differences. The transformation-dependent Gini is confined to the first-stage descriptive appendix (Appendix A). Table 9 records the design parameters; Table 10 reports the factorial ANOVA.

### 5.2 Directional predictions: H1–H3

The $F$-statistics and $\eta^2$ quoted in this subsection are from the per-replication mean-reward ANOVA ($n = 3{,}750$, Table 10), computed on the post-fix factorial and regenerated together with that table by `regenerate_tables.py` (§8). The condition-level classical $\eta^2$ ($n = 125$, with within-replication noise removed) puts stakes at 0.74, alignment at 0.18, and the observation-noise proxy at 0.015. We carry these forward only as one of *two* co-equal descriptions: §5.6 shows the stakes-versus-alignment ranking inverts on the reward-scale-normalized gap, so neither ordering is a structural fact. The nominal-$\alpha$ U-shape and observation-noise direction are invariant to normalization; the stakes direction is a raw-gap statement, and dividing by $\sigma$ removes it by construction.

**Stakes raise friction (H2).** Stakes exhibit the strongest *raw* main effect by a wide margin ($F = 749.66$, $\eta^2 = 0.3894$, $p < 0.001$, per-replication), confirming the direction of H2. The regression coefficient $\hat{\gamma}_\sigma = 2.834$ makes stakes the largest raw predictor, accounting for roughly 40% of per-replication variance. Mean reward falls monotonically from $-0.599$ at $\sigma = 0.2$ to $-2.865$ at $\sigma = 1.0$—a near-5$\times$ increase in coordination loss—and approximately linearly in $\sigma$. This linearity is exactly the degree-1 homogeneity built into the reward (§3.2); we show in §5.6 that it is also why the *raw* dominance of stakes is a measurement convention rather than a structural fact.

**The observation-noise proxy contributes weakly (H3).** The implemented noise-variance parameter shows a significant but modest main effect ($F = 15.76$, $\eta^2 = 0.0082$, $p < 0.001$), with $\hat{\gamma}_\varepsilon = 0.292$: observation noise increases friction, supporting H3, but its effect is roughly an order of magnitude weaker than stakes. Mean reward falls from $-1.549$ at $\varepsilon = 0$ to $-1.861$ at $\varepsilon = 1.0$ (a 20% increase in loss). The $\alpha \times \varepsilon$ interaction is not detected ($F = 1.15$, $p = 0.302$): we find no evidence for the multiplicative amplification implied by the canonical $(1+\varepsilon)/(1+\alpha)$ form, but this is a non-detection rather than a demonstration of additivity.

**Nominal alignment is U-shaped—and symmetric in sign (H1).** Nominal alignment has a significant main effect ($F = 185.44$, $\eta^2 = 0.0963$, $p < 0.001$), but it is *not* monotonically inverse as H1 predicts. The relationship is U-shaped: neutral alignment ($\alpha = 0$) produces the worst outcomes (mean reward $-2.526$), while every nonzero $\alpha$ is markedly better—$-1.659$, $-1.457$, $-1.468$, $-1.615$ at $\alpha = -0.8, -0.4, +0.4, +0.8$ respectively—with the minimum friction at moderate magnitude $|\alpha| = 0.4$. This nominal relationship is non-monotone, but it cannot falsify H1: the sampler neither realizes signed target correlation nor directly measures the theoretical alignment construct. It diagnoses the experimental proxy instead. Figures 8 and 9 display the effect.

The effect is essentially *even* in $\alpha$. On the reward-scale-normalized condition means, a signed-linear term in $\alpha$ explains $R^2 \approx 0.000$ of total between-condition variance while a magnitude term $|\alpha|$ explains $R^2 \approx 0.087$: to three decimals, *none* of the nominal-$\alpha$ variance reflects the sign of $\alpha$. This is not a surprising empirical discovery so much as a consequence of the design: the cross-agent target correlation (4) depends only on $|\alpha|$, and the $\pm\alpha$ conditions are statistically exchangeable (§3.2). What the U-shape registers is therefore the *availability of shared target structure*, not a cooperative-versus-adversarial direction. At $\alpha = 0$ each agent draws an independent target, so there is no cross-agent correlation for any learner to exploit and no jointly satisfiable configuration; as $|\alpha|$ grows, agents share an anchor and the joint problem becomes more nearly feasible. Because this axis is sign-blind, neither its apparent learnability nor its feasibility pattern is evidence about opposition. The trough reappears under the VDN/team-reward bundle (§5.7), but that swap changes learner and objective together, so it does not isolate non-stationarity. It shows only that the known-artifactual shape is not confined to the original IQL/individual-reward bundle.

### 5.3 H4: no valid confirmatory verdict; adverse reduced-grid comparison

Table 11 reports the model comparison on the first-stage factorial. The canonical friction specification M1 attains $R^2 = 0.110$ on condition-level reward gaps—far below the $R^2 > 0.7$ adequacy criterion stated in advance. The poor fit has two sources: the U-shaped nominal-$\alpha$ effect violates the monotone $1/(1+\alpha)$ assumption, and the dominance of stakes means that collapsing all three factors into a single composite discards substantial information.

On this first-stage factorial, the independent-effects model M4 attains $R^2 = 0.748$ with the lowest

AIC (173.7), beating M1 by $\Delta\text{AIC} = 153.6$. That gap is large on information-theoretic grounds (Burnham and Anderson, 2002), but the comparison is circular as evidence about signed target correlation because M1 was guaranteed to misfit the sign-blind axis. The signed re-estimation below repairs the validity of the comparison. The additive (M2, $R^2 = 0.158$) and multiplicative (M3, $R^2 = 0.186$) composites fare only marginally better than M1 and remain far below M4. All four models agree on the directional content for stakes and the observation-noise proxy; the disagreement is about the alignment shape and the multiplicative interaction structure. The canonical form captures the right variables but not the right shape for alignment. M4's fit is carried almost entirely by its strong linear stakes term ($\hat{\gamma}_\sigma = 2.834$), not by alignment: its near-zero linear alignment coefficient ($\hat{\gamma}_\alpha = -0.019$) is not evidence that alignment is irrelevant—the U-shape is invisible to any linear term, which is precisely why even a linear main-effects model that simply keeps stakes separate from alignment outperforms the ratio composite—and a quadratic $\alpha^2$ predictor would improve M4 further.

**Exploratory comparison on the signed design.** Because the first-stage target proxy is sign-blind (§6), and the monotone $1/(1+\alpha)$ denominator was *guaranteed* to misfit an even-in-$|\alpha|$ U-shape, rejecting M1 on that grid alone would be circular. We therefore re-estimate the identifiable forms on the signed target-equicorrelation data of §6, mapping the empirical proxy to realized target-coordinate correlation $\rho$ (so $\rho < 0$ is negative target dependence, not the $|\alpha|$ mirror). The signed re-run fixes $\varepsilon = 0$ and varies only $\rho \in \{-.33, \ldots, +0.8\}$ and $\sigma \in \{0.6, 1.0\}$, so this is a re-test of the form's *signed target-correlation and reward-scale* dependence—precisely the axis the artifact corrupted—and *not* of the entropy term or of general objective/reward-function alignment. At $\varepsilon = 0$ the multiplicative composite $\text{M3} = \sigma\varepsilon(1-\rho)$ is identically zero and not identifiable, and M4's entropy slope is absorbed into the intercept. We estimate M0–M4 where identifiable, then add $\text{M5} = \sigma(\beta_0 + \beta_\rho \rho)$, which enforces the implemented degree-one reward-scale homogeneity, M6 with a free intercept, and quadratic/cubic canonical-index responses M7/M8. All use the 16 signed $(\rho, \sigma)$ conditions per learner and the same strict AICc convention.

Table 12 reports the result. Because $n = 16$ with $K = 3$–4, we use the strict small-sample correction $\text{AICc} = \text{AIC} + 2K(K+1)/(n-K-1)$ (Burnham and Anderson, 2002). M4 beats M1 by $\Delta\text{AICc} = 22.5$ (IQL) and 17.7 (VDN), but the stronger result is M5: it reaches $R^2 = .959/.958$ and beats M1 by $33.45/30.51$ and M4 by $10.99/12.81$ AICc units. M6's free intercept does not improve AICc, and the flexible M7/M8 response curves do not rescue the canonical denominator. This is exploratory model comparison, not a confirmatory H4 pass/fail test: the signed grid is post-diagnostic and omits variation in $\varepsilon$.

M1's absolute fit nevertheless improves sharply, from $R^2 = 0.110$ in the first stage to 0.666 (IQL) / 0.717 (VDN). Numerically, only the latter crosses the historical $R^2 > 0.7$ threshold, but that crossing is a descriptive calibration of the reduced $\sigma/(1+\rho)$ restriction, not an H4 pass. At the same parameter count, the reward-scale-only baseline M0 attains $R^2 = 0.842$ / 0.777 and beats M1 by $\Delta\text{AICc} = 11.9$ / 3.8. The negative $\rho$ slopes describe the narrower target-correlation treatment, not general alignment (§6.2).

The signed and first-stage information-criterion differences arise from different data and are not placed on a common per-condition scale. The signed re-estimation repairs the comparison's *validity*. Among the estimable theory-motivated candidate composites, M1 also beats the additive M2 ($R^2 = 0.491$ IQL / 0.569 VDN); M3 is unidentifiable at $\varepsilon = 0$ and is not ranked.

The verdict is therefore specific: retain the negative target-correlation direction and degree-one reward scaling in this environment while rejecting the implemented reciprocal ratio as a maintained exploratory description, not as a confirmatory H4 verdict, and treat the split between feasibility and the

frozen-policy residual as target-support dependent. The best tested response is scale-homogeneous and linear in target correlation; that is not evidence for a universal law.

### 5.4 Interaction effects

Table 14 reports the interaction structure. The $\alpha \times \sigma$ interaction is significant ($F = 10.34$, $\eta^2 = 0.0215$, $p < 0.001$): the worst conditions cluster at neutral alignment *and* high stakes ($\alpha = 0$, $\sigma \geq 0.8$), and the U-shape amplifies with stakes (Figure 8). The $\sigma \times \varepsilon$ interaction is marginal ($F = 1.83$, $\eta^2 = 0.0038$, $p = 0.023$), and the three-way interaction is non-significant ($F = 0.94$, $p = 0.622$)—limited support for the fully multiplicative structure the canonical form assumes.

### 5.5 First-stage descriptive results (deferred to appendix)

Two further first-stage descriptions are known-artifactual properties of the sign-blind design and bear on alignment *magnitude*, not sign; we therefore defer them to Appendix A and do not re-estimate them under the signed DGP. The first is a *policy-versus-outcome convergence gap* (a near-zero policy-stability fraction beside near-universal reward stability), which we read as a descriptive fact and explicitly *not* as evidence of strategic cycling—not because exploration precludes a fixed point (it does not; §A.1), but because an arg max sampled every 200 episodes on a fixed probe set cannot separate strategic cycling from drift. The second is *agent-level reward inequality* (Gini and variance), $|\alpha|$-governed and lowest at the alignment extremes (§A.2).

### 5.6 Effect sizes: two descriptions of one dataset

**Why there is no single "dominant" factor.** A factorial ANOVA on the 125 condition means (classical $\eta^2 = \mathrm{SS}_{\mathrm{effect}}/\mathrm{SS}_{\mathrm{total}}$, the same aggregation as the $R^2$/AIC comparison) assigns stakes $\eta^2 = 0.74$, alignment 0.18, and observation-noise proxy 0.015; the main effects sum to 0.933 and the interactions to the remaining 0.067 ($\alpha \times \sigma = 0.039$, $\sigma \times \varepsilon = 0.007$, $\alpha \times \varepsilon = 0.004$, three-way 0.014). This "stakes dominate" ranking is real but *contingent on the $\sigma$ grid*. Stakes enter the reward with degree-1 homogeneity (§3.2), so $\sigma$ rescales the reward landscape by a common factor without changing its geometry, and the reward gap inherits a mechanical $\sigma$-scaling that any chosen $\sigma$ spacing will amplify. We emphasize that this linear gap-versus-$\sigma$ relationship is a *deterministic algebraic consequence* of the reward being homogeneous of degree one in $\sigma$—$R_i \approx \sigma \sum_j u(s_j - \tau_{ij})$, so $\Delta R \propto \sigma$ up to the weight-clipping that affects only the $\sigma = 1.0$ cell (empirically negligible: the reward-scale-normalized $\eta^2$ collapses to 0.004)—and *not* an emergent property of what the agents learned. The agents do not "discover" that higher stakes raise friction; the reward algebra fixes that scaling before any learning occurs. The result is therefore valuable as a *measurement-convention* caution—raw $\eta^2$ rankings inherit whatever $\sigma$ grid one picks—rather than as a behavioral finding about theoretical stakes. Dividing the gap by $\sigma$ removes exactly this algebraically-defined scaling. On the resulting reward-scale-normalized gap the cross-factor ranking *inverts*: alignment $\eta^2 = 0.82$, observation-noise proxy 0.064, stakes 0.004. The collapse of the stakes term from 0.74 to 0.004 confirms the scaling is essentially exact (degree-1), and the $\sigma^1$ exponent is dictated by the reward algebra, not chosen to produce the inversion. "Essentially" is doing real work in that sentence, and one specific thing explains it. Weights are drawn $w_{ij} \sim \mathcal{N}(\sigma, 0.05^2)$ and clipped to $[0,1]$. The dispersion is a *fixed absolute* 0.05 rather than a fixed fraction of $\sigma$, so relative weight dispersion falls as stakes rise; more sharply, at the top of our grid the clip binds. At $\sigma = 1.0$ exactly half the sampled weights exceed 1 and are pinned there, giving that cell a weight distribution with an atom at the boundary, a mean near 0.98 rather than 1.0, and roughly half the intended variance—while at $\sigma \leq 0.8$ the clip is effectively never reached ($\Pr = 3.2 \times 10^{-5}$ at $\sigma = 0.8$, and below $10^{-15}$ for every lower level—so of the ~45,000 weights drawn at $\sigma = 0.8$ roughly one is clipped, against half of them at

$\sigma = 1.0$). The top stakes level is therefore not a clean rescaling of the others, which is precisely why the normalized stakes term lands at 0.004 rather than at zero. The residual is small and it does not touch the signed arms (run at $\sigma \in \{0.6, 1.0\}$ and analyzed within $\sigma$), but the honest generator is $w_{ij} = \sigma v_{ij}$ with a scale-free $v$, and a re-run on that draw would remove the artifact rather than bound it. We therefore report *both* partitions and treat neither $\eta^2$ ranking as a structural fact: "stakes scale the magnitude of realized welfare loss" and "preference structure is the dominant *structural* axis once that scaling is removed" are two true descriptions of the same data.

**Inference level and effect size of record.** The condition-level partition is a *saturated* descriptive decomposition—the 125 cell means are fit exactly, leaving zero residual—so it carries no $p$-values and serves only to apportion variance. Inference comes from the per-replication ANOVA ($n = 3{,}750$, Table 10), with 30 replication blocks per condition. On the batched GPU path, those runs are separate blocks drawn from one deterministic per-condition pseudorandom stream, not 30 separately seeded processes; each block nevertheless receives its own base vector, preferences, environment resets, and noise draws. The CPU path instead assigns an explicit seed to each replication. The design is therefore not agent-level pseudoreplication. At this $N$ and these $F$-values the bias-corrected $\omega^2$ is numerically indistinguishable from $\eta^2$ (within 0.001: stakes $\omega^2 \approx 0.39$, alignment $\approx 0.095$, observation-noise proxy $\approx 0.007$), so we quote $\eta^2$. The per-replication and condition-level levels agree on ordering and differ only in magnitude, the latter having averaged out within-condition replication variance (§3.4). We therefore conduct all inference at the *replication-block* unit, never at the agent level. As a clustering sensitivity check we confirmed this matters: re-fitting an interaction at the *agent* level (four agents per run, $n = 1920$) with cluster-robust standard errors at the declared unit (the replication, $G = 480$) weakens a naively significant agent-level interaction from $p \approx 0.02$ to $p = 0.11$ (IQL) / 0.086 (VDN); cell-clustering ($G = 16$, with $\rho$ constant within cluster) is too few-cluster to be reliable. The four-agents-per-run pseudoreplication this guards against is exactly why we never report agent-level $p$-values.

**Dynamic range.** The dynamic range of the manipulation is large. The best condition ($\alpha = 0.4$, $\sigma = 0.2$, $\varepsilon = 0$: $-0.418 \pm 0.136$) versus the worst ($\alpha = 0$, $\sigma = 1.0$, $\varepsilon = 0.75$: $-4.568 \pm 1.738$) gives Cohen's $d = 3.37$ (Cohen, 1988)—a contrast of low-stakes structured preferences against high-stakes uncorrelated ones. A within-structure stakes contrast at matched alignment magnitude ($|\alpha| = 0.8$, $\sigma = 0.2$, $\varepsilon = 0$: $-0.494 \pm 0.252$ versus $|\alpha| = 0.8$, $\sigma = 1.0$, $\varepsilon = 1.0$: $-3.120 \pm 1.663$) gives Cohen's $d = 2.21$ (for provenance: the low cell is drawn from the $\alpha = +0.8$ arm and the high cell from the $\alpha = -0.8$ arm; the two single-arm pairings give $d = 1.93$ and $d = 2.31$); the pair differs only in reward scale and noise, not in nominal $\alpha$, since the $\pm 0.8$ cells are exchangeable (§3.2). That the absolute worst case is *neutral* alignment, not any nonzero $\alpha$, is the first-stage design's headline effect restated as an effect size.

### 5.7 Cross-bundle check: does the U-shape reappear under VDN with a team reward?

The first-stage U-shape should be visible to any learner/objective bundle that is sensitive to the shared structure the sampler actually varies. We therefore ask a narrow replication question: does the shape reappear when both learner and objective change? This does not isolate non-stationarity or centralization because VDN also replaces individual rewards with a summed team reward (§3.5). We re-run with Value-Decomposition Networks (Sunehag et al., 2018), an additive factorization of a shared team value, on a reduced grid: all five $\alpha$ levels, $\sigma \in \{0.2, 0.6, 1.0\}$, $\varepsilon = 0$, 30 replications per cell, against matched IQL runs.

The trough survives intact. Neutral alignment is the worst case in *every* one of the six $(\sigma, \text{mode})$

cells. On the reward-scale-normalized gap, the mean U-depth—the excess gap at $\alpha = 0$ over the mean of the $|\alpha| = 0.8$ cells—is $+1.473$ under IQL and $+1.589$ under VDN: VDN does not flatten the trough at all—its mean U-depth is, if anything, marginally deeper. The VDN/team-reward bundle lowers the coordination gap at every alignment level (pooled normalized gap IQL versus VDN: 2.39 vs 1.70, 2.17 vs 1.71, 3.94 vs 3.32, 2.19 vs 1.73, 2.55 vs 1.76 at $\alpha = -0.8, -0.4, 0, +0.4, +0.8$; Welch $p < 0.01$ at all five levels)—but the level shift cannot be attributed to centralization separately from the team objective. The bundle's level advantage is marginally *larger* at the extremes ($+0.74$ averaged over $|\alpha| = 0.8$) than at neutral alignment ($+0.63$), so it does not preferentially fill the trough. The $|\alpha|$-symmetric U-shape is thus reproduced across the two learner/objective bundles—but, we emphasize, this is replication of the *artifact*: because the $\pm 0.8$ cells are statistically exchangeable (§3.2), what VDN preserves is the magnitude-of-shared-structure effect, not a signed-alignment ordering. The $\pm 0.8$ symmetry indeed holds only *on average*—individual cells fluctuate in which extreme is worse (e.g. $+0.8$ is slightly worse at $\sigma \in \{0.2, 0.6\}$ but $-0.8$ at $\sigma = 1.0$), exactly as expected for distributionally identical conditions differing only in their stochastic realization. The signed cross-bundle check—does the *signed* shape reappear under VDN/team rewards?—is deferred to §6, where it does.

### 5.8 Robustness: GPU/CPU cross-validation

**A defect, a prediction, and a re-run.** We re-ran the *entire* factorial on a parallel CPU codebase (Python multiprocessing, 22 workers) with matched hyperparameters (identical learning rate, discount, exploration schedule, network size, and 1,000-episode budget) and compared the two implementations across all 125 conditions (Table 15). The first version of this arm was not the MDP-level replication we intended, and finding out why is the most useful thing this subsection does.

The environment's action decoder assumes the Gymnasium `MultiDiscrete` encoding and maps $\{0,1,2\} \mapsto \{-1,0,+1\}$ by subtracting one. Every caller, however, builds its action table with `build_action_map`, which already emits $\{-1,0,+1\}$. Composed, the two produce deltas in $\{-2,-1,0\}$: under that dynamics a resource level can only ever *fall*. The maximal upward action—all four agents requesting $+1$ on every resource—moved the state by exactly zero. The GPU path of record never goes through this decoder (it applies its action table to the state directly), so no other result in this paper is affected; the CPU arm did. Two further defects, the observation-noise variance mapping and the replay next-observation leak disclosed in §8, had been fixed on the GPU side and never propagated to the CPU one.

The defect made the environment *easier*, not harder, and getting that sign right is what identifies it. Agent targets are drawn about zero (§3) while the state is clipped to $[0, C]$, so a state that can only fall converges to a point that is close to *every* agent's target. With the state pinned at 0 the per-agent reward has a closed form, $-\sum_j w_{ij} t_{ij}^2$, and the defective arm's 125 condition means sit on it: mean absolute deviation 0.109 with mean signed difference $-0.029$, against 0.554 and 0.528 for the repaired CPU arm and the GPU arm respectively. The defective arm scored *higher* than the GPU arm in 122 of 125 cells, and repairing it moved CPU reward down by 0.525. A broken decoder that collapses the state to the one point the reward function likes is not a handicap; it is a free ride, and the closed form says so without reference to any other defect.

We then fixed all three defects and re-ran the arm in full (3,750 runs, 32.8 hours). Both rows below are scored against the post-fix GPU arm of record, so only the CPU arm varies:

| CPU arm | offset | Spearman $\rho$ | Pearson $r$ | ICC(3,1) | ICC(A,1) |
|---|---|---|---|---|---|
| before (all three defects) | $+0.499$ | 0.948 | 0.943 | 0.907 | 0.769 |
| after (all three repaired) | $\mathbf{-0.026}$ | **0.964** | **0.963** | **0.962** | **0.962** |

Earlier versions of this paper printed $+0.478$, $\rho = 0.948$, $r = 0.945$, $\mathrm{ICC}(3,1) = 0.913$ for the first row. Those came from pairing the defective CPU arm against the *pre*-fix GPU arm; re-scoring the same CPU data against the GPU arm of record gives the row above. The GPU arm itself moved by only $-0.021$ between the two, so the change is bookkeeping rather than substance—but it is a change in a published number and we flag it as one.

Two ICCs are reported because the consistency form cannot carry the claim at issue. $\mathrm{ICC}(3,1)$ is invariant to a pure column shift: it reads 0.907 on an arm carrying a $+0.499$ offset, which is precisely the disagreement under test. The absolute-agreement form $\mathrm{ICC}(\mathrm{A},1)$ does register it, at 0.769, and it is the move from 0.769 to 0.962 that supports a statement about levels.

**What the re-run identifies, and what it does not.** Three defects were repaired at once, so the collapse of the grid-mean offset is jointly attributable to the set. The data do contain one clean single-defect stratum, and it does not say what the simple story would predict. Defects (ii) and (iii) sit behind an `if cond_epsilon > 0` guard in the runner, and the corresponding `else` branches are byte-identical either side of the commit that fixed them, so at $\varepsilon = 0$ the action-encoding shift is the only live change. In that stratum the repair does not remove the offset—it overshoots, from $+0.260$ ($p = 9 \times 10^{-7}$) to $-0.166$ ($p = 0.005$). Everything that collapsed collapsed at $\varepsilon > 0$, where the other two fixes were active ($+0.559 \to +0.009$, $p = 0.73$). The pre-fix offset was itself strongly $\varepsilon$-dependent (one-way ANOVA $F = 4.94$, $p = 0.001$; $\varepsilon = 0$ against $\varepsilon > 0$, Welch $p = 5 \times 10^{-7}$), and a decoder that pins the state is $\varepsilon$-independent by construction, so the decoder alone never accounted for the whole offset. The pattern is consistent with the replay leak's advantage growing with the noise it conceals, but we have not run that manipulation in isolation and do not claim it.

The residual is correspondingly *not* undifferentiated noise. It is not distinguishable from zero where all three fixes act and is significantly negative in the one stratum that isolates the decoder, with detectable heterogeneity across strata ($F = 3.53$, $p = 0.009$). We do not have a complete explanation for the $\varepsilon = 0$ residual; one candidate is a difference the two runners still carry, namely zero-initialized biases in all three GPU linear layers against PyTorch's default uniform bias initialization in the CPU networks. We tested that candidate under a frozen, matched-seed staged diagnostic over the 25 $\varepsilon = 0$ cells, extending from 5 to 10, 20, and at most 30 replications per cell. At the 30-replication cap (750 runs), zero initialization improved CPU mean reward by 0.0412 (condition-bootstrap 95% CI $[0.0315, 0.0514]$) and moved the matched CPU-minus-GPU offset from $-0.1658$ to $-0.1246$. The primary CI $[-0.2322, -0.0323]$ was neither contained within nor wholly outside the prespecified $\pm 0.10$ equivalence margin. The frozen rule therefore remains inconclusive at its planned maximum: bias initialization is a real, directionally helpful implementation difference, but not an established full explanation of the residual. The protocol and sequential outputs are archived under `results/zero_bias_eps0/`.

The corrected aggregate result is the claim the defect had cost us: the two implementations agree at the level of the decision process, not only in rank order. Over 125 conditions, with independent seed streams and different floating-point accumulation, $\rho = 0.964$, $r = 0.963$, $\mathrm{ICC}(\mathrm{A},1) = 0.962$, and a grid-mean offset of $-0.026$ (90% CI $[-0.065, 0.013]$) that is not distinguishable from zero (paired $t$, $p = 0.27$; Wilcoxon, $p = 0.55$). Paired TOST establishes equivalence at a $\pm 0.10$ margin ($p = 0.0011$), but not at $\pm 0.05$ ($p = 0.156$). Bland–Altman bias is $-0.026$ with limits of agreement of roughly

$\pm 0.52$ and 91.2% of conditions inside them, slightly below the nominal 95%. All figures come from `results/full_factorial_postfix/crossval.py`.

Two notes on what is *not* claimed. Cross-implementation agreement for the withdrawn secondary proxies (§3.4) is not reported: replication of a misdefined quantity is not evidence about the quantity it was named for. And the replay-sampling explanation we previously floated for the offset is withdrawn on its own merits, independently of the decoder: sampling with and without replacement are both unbiased for the same expected gradient and differ only in minibatch variance, so neither predicts a shift of consistent sign.

We record the whole episode, including the part that reflects badly on us. The first draft of this subsection got the decoder's sign backwards—it asserted that a state which can only fall sits *further* from every target, two sentences after reporting that the defective arm scored *higher*—and then used that reasoning to upgrade the finding from plausible to established on the strength of a re-run that repaired three things at once. An adversarial check against the data caught both. Neither error changed the direction of the diagnosis, and the identification that replaced them is stronger than the one they were meant to support; but a confounded manipulation credited to a single cause, with a mechanism that predicts the opposite of the observed sign, is the exact failure this paper is about, committed in the section claiming to apply the paper's method to the paper. That is the more useful datum. A control battery is not a disposition one acquires by writing about it, which is why we report the numbers the strata give rather than the ones the story wanted.

## 6 Results II: The Control Battery

The first stage measured the magnitude of a correlation that never changed sign. Throughout this section $\rho$ denotes the signed cross-agent target correlation—the genuinely signed realization that the framework's $\alpha \in [-1,1]$ was meant to denote—so that "cooperation" ($\rho > 0$) and "opposition" ($\rho < 0$) are now distinct rather than collapsed by the $|\alpha|$-only sampler of §3.2. Throughout this section $\bar{g}$ denotes the per-cell mean reward gap $\Delta R$ of §3.4, and $\bar{g}/\sigma$ its stake-normalization. This section runs the control battery that turns that misleading baseline into the result of record. It begins by rebuilding the data-generating process so the cross-agent correlation is genuinely signed and re-run for both IQL and VDN (§6.1–§6.4): the symmetric U-shape of §5 disappears, leaving a cooperative-side decline. A separable-resource control (§6.5) then tests whether even that surviving effect is a context-free target-correlation law or is bounded to the shared-state environment family—and finds the latter. A paired legacy-Gaussian-at-zero-support parametric partial-sharing control (§6.6) interpolates between the shared and separable endpoints and grades that bound as an overlap response in the fraction of shared coordinates. An $n = 2$ latent-target opposition control (§6.7) reaches sign-verified target $\rho = -1$ while carrying an agent-count confound and feasible correlation $-0.468$; it detects no opposition advantage. Finally, matched target-support and dynamic-benchmark controls (§6.8) freeze policies, use held-out resets, deliver the signed target distribution inside the state box, and compare performance with an exact finite-horizon controller. This catches the fourth artifact; the separable and partial-sharing arms are related scope checks rather than additional members of the title's four design choices. Except for the original factorial hypotheses, these controls were designed after the first-stage failure was diagnosed. Repository history dates the earlier signed and contention arms; the frozen-policy, feasible-support, dynamic-oracle and signed-observation-noise protocols were fixed locally on 13 August 2026 before their full grids were analysed, but they were not preregistered or independently replicated. We report their uncertainty and multiplicity controls as exploratory inference, not confirmatory evidence. The verdict (§6.9) reads off the honest core: more positive signed target-coordinate correlation lowers the gap only in the tested shared-state

variants, while the split between achievability and the frozen-policy residual depends materially on target support.

### 6.1 The data-generating process: bug and fix

**The bug.** The first-stage sampler draws one base vector $b$ *shared across all agents* and sets $\tau_i = \alpha b + (1-|\alpha|)\xi_i$ (3). As (4) makes explicit, the induced cross-agent target correlation is $\alpha^2/(\alpha^2+(1-|\alpha|)^2)$, a function of $|\alpha|$ alone that is *never negative*. We verify this directly (Figure 1, left; 20,000 draws): the realized correlation is $\approx +0.941$ at both $\alpha = -0.8$ and $\alpha = +0.8$, and $+0.307$ / $+0.304$ at $\alpha = -0.4$ / $+0.4$ (theory: 0.941 and 0.308), with only $\alpha = 0$ giving $\rho \approx 0$. The negative-$\alpha$ cells are not anti-aligned; they carry the *same* strong positive correlation as their positive twins. The first-stage "adversarial arm" did not exist, and the apparent U-shape is the signature of a manipulation that varied $|\alpha|$ while holding the sign of the correlation fixed at $+$. The same coupling entangles a second, subtler distortion: the marginal target variance $\mathrm{Var}(\tau_{ij}) = \alpha^2 + (1-|\alpha|)^2$ dips non-monotonically (to 0.52 at $|\alpha| = 0.4$), and the cross-agent target dispersion $\mathrm{Var}(\tau_{ij} - \tau_{kj}) = 2(1-|\alpha|)^2$ shrinks monotonically in $|\alpha|$ regardless of sign—so the first stage varied target spread alongside correlation magnitude. The equicorrelated repair below holds marginal variance at 1 for every $\rho$ and therefore corrects *both* distortions at once; first-stage-versus-signed contrasts are accordingly cross-design comparisons, not a pure sign repair.

**The fix.** We make the cross-agent correlation $\rho$ the control knob directly, and genuinely signed, through two principled designs.

*(A) Equicorrelated continuum (primary).* Per resource, draw the $n$-agent target vector from $\mathcal{N}(0, \Sigma(\rho))$ with

$$\Sigma(\rho) = (1-\rho)\, I_n + \rho\, \mathbf{1}\mathbf{1}^\top, \tag{7}$$

sampled via an eigendecomposition-based matrix square root (stable at the boundary). An equicorrelated matrix is positive semidefinite only for $\rho \geq -1/(n-1)$, so for $n = 4$ the adversarial arm is capped at $\rho = -1/3$. Marginal target variance is held at 1 for every $\rho$, so latent-target magnitudes are comparable across the sweep. This yields a signed continuum in the latent targets, $\rho \in \{-0.33, -0.2, -0.1, 0, 0.2, 0.4, 0.6, 0.8\}$; the box constraint attenuates the corresponding feasible-optimum correlations asymmetrically (§8).

*(B) Two-team block (strong opposition).* To reach opposition beyond the equicorrelated bound, partition the $n = 4$ agents into two teams of two: team A shares $+$base, team B shares $-$base, with strength opp $\in [0,1]$. Within-team correlation is $+\text{opp}^2$ and between-team correlation is $-\text{opp}^2$, reaching $-1.0$ at opp $= 1$ (genuine strong opposition; mean pairwise correlation $-\text{opp}^2/3$).

In both designs the weights $w_{ij} \sim \mathcal{N}(\sigma, 0.05^2)$ clipped to $[0,1]$ are *unchanged*, preserving $\sigma$ as the reward-scale proxy for stakes, and the IQL/VDN learners are reused verbatim—only the target sampler is swapped. A pre-run verification (Figure 1, right) confirms the realized equicorrelated correlation tracks its target essentially exactly $(-0.330, -0.200, -0.100, 0.000, 0.200, 0.400, 0.600, 0.800)$ and the team design spans between-team correlation $0 \to -1.00$. The repaired design *passes the sign gate for its latent targets* that the first-stage design silently failed: the realized latent-target correlation spans negative $\to$ zero $\to$ positive. This does not make the feasible-optimum axis exact; its attenuation is quantified in §8.

Stated as a transferable procedure, that gate has two parts and both are load-bearing. First, compute the cross-agent coupling the sampler actually realizes. Second, locate the estimand the design's own text claims to sweep. Flag on a *mismatch* between the two—never on the shape of the realized correlation alone. A sampler whose realized coupling is always positive, or always negative, is not thereby defective;

it is defective only where it is read as varying the sign. Running the first half without the second produces false positives, as we demonstrate on ourselves in §7.4. The move mirrors the redesign of the StarCraft Multi-Agent Challenge into SMACv2 after a construct-validity failure was demonstrated there—open-loop policies conditioned only on the timestep could solve much of the benchmark, so it was not measuring the closed-loop coordination it was assumed to (Ellis et al., 2023). Here, likewise, the diagnosis is that the instrument did not realize the signed target dependence it claimed. The repair verifies that narrower treatment; it does not establish equivalence with general objective or reward-function alignment.

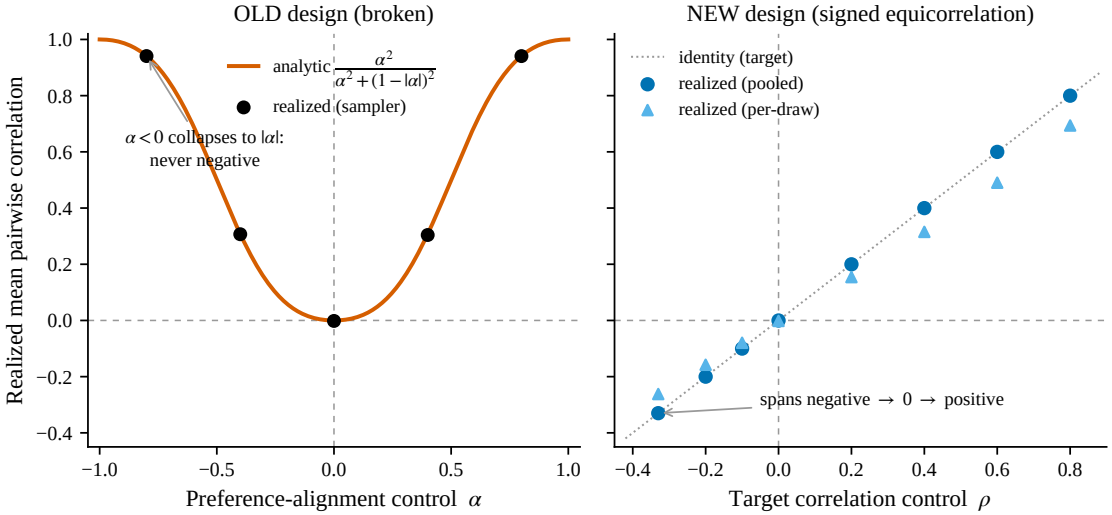


Figure 1: The signed equicorrelation DGP fixes the $|\alpha|$-symmetry artifact of the original sampler. Left: the old design's realized mean pairwise correlation is $\alpha^2/(\alpha^2+(1-|\alpha|)^2)$, symmetric in $|\alpha|$ and never negative. Right: the new control spans negative $\to 0 \to$ positive correlation as targeted.

**Compute.** The runner batches all (replication $\times$ agent) networks and the environment loop into single tensor operations, so the GPU wins despite tiny networks. Measured on an AMD Radeon RX 7900 XTX (ROCm 6.3), the full 1,000-episode run is $\approx 2.65\times$ faster than a single-process CPU baseline; the full equicorrelated sweep (8 correlations $\times 2$ reward scales $\times 30$ replications $\times 1{,}000$ episodes, both IQL and VDN) completes in $\approx 22$ minutes across the two GPUs. Because the preference draws depend only on $(\rho \times \sigma, \text{condition stream}, \text{replication index})$ and not on the learner, IQL and VDN see *identical preference draws* at matched condition streams and replication indices, so the two halves share the same problem instances; we are careful *not* to claim identical network initializations (the archived runs predate the deterministic torch seeding discussed in the data statement, so weight inits were a single realization), only matched preferences. The signed correlation and contention arms in this section fix $\varepsilon = 0$ (the first stage found a real but small effect on its own scale—the full-range clean-evaluation contrast is bounded to 21.5% of the reward-scale anchor—but that estimate was obtained under the first stage's effectively cooperative realized structure. The later reduced signed-noise crossing tests that interaction on feasible-centred held-out outcomes; at $\varepsilon = 0$ alone the multiplicative form M3 is unidentifiable), and use 30 replications per cell and $\sigma \in \{0.6, 1.0\}$ (equicorrelated) / $\sigma = 1.0$ (team).

### 6.2 Legacy-support training-window gap tracks signed target correlation

Table 1 and Figure 2 report the normalized coordination gap $\overline{g}/\sigma$ across the signed continuum (pooled over $\sigma$, $n = 60$ per cell). The curve is a ramp, not a U: a high plateau over the opposition-to-baseline

range ($\rho \leq 0$) and a monotone decline over the cooperative range ($\rho > 0$), bottoming at $\rho = +0.8$ for both learners. There is no second, adversarial arm.

Table 1: Legacy-support training-window mean normalized coordination gap $\bar{g}/\sigma$ by signed cross-agent correlation $\rho$ (95% CI, $n = 60$ per cell, pooled over $\sigma \in \{0.6, 1.0\}$, $\varepsilon = 0$). Worst cell $\rho = -0.10$ for both learners; best $\rho = +0.80$.

| $\rho$ | IQL gap [95% CI] | VDN gap [95% CI] |
|---|---|---|
| $-0.33$ | 3.863 [3.433, 4.292] | 3.196 [2.828, 3.564] |
| $-0.20$ | 4.071 [3.624, 4.518] | 3.362 [2.979, 3.745] |
| $-0.10$ | **4.077** [3.715, 4.438] | **3.380** [3.076, 3.684] |
| 0.00 | 4.016 [3.609, 4.423] | 3.355 [2.998, 3.712] |
| $+0.20$ | 3.650 [3.308, 3.992] | 3.023 [2.727, 3.318] |
| $+0.40$ | 3.394 [3.038, 3.751] | 2.757 [2.457, 3.056] |
| $+0.60$ | 3.336 [2.912, 3.759] | 2.690 [2.340, 3.040] |
| $+0.80$ | **3.309** [2.729, 3.889] | **2.431** [1.976, 2.885] |

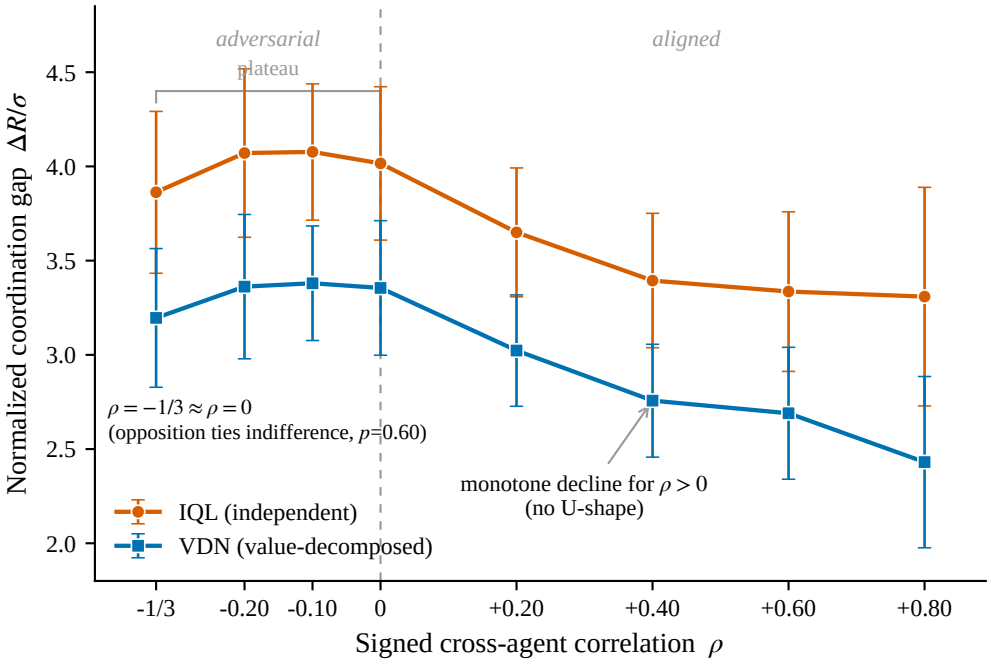


Figure 2: Legacy-support training-window gap tracks cross-agent correlation, with no U-shape: mean normalized gap $\bar{g}/\sigma$ versus signed cross-agent correlation $\rho$ for IQL and VDN (95% CI, $n = 60$ per cell, pooled over $\sigma \in \{0.6, 1.0\}$). Point estimates are approximately level over opposition to baseline and decline over the tested cooperative levels; conditional opposition contrasts remain mixed and are not equivalence results.

**Legacy-support training-window contrasts: pooled opposition is unresolved.** The planned contrasts (Welch's $t$ on the normalized gap, Table 2) confirm the asymmetry. For genuine opposition (the sampled equicorrelated $\rho = -.33$) versus baseline ($\rho = 0$) *no statistically significant difference is detected in this sample*: the difference is $-0.154$ for IQL ($t = -0.52$, $p = 0.60$) and $-0.159$ for VDN ($t = -0.62$, $p = 0.54$). We read this as a failure to detect a difference, not as evidence of equivalence—the per-cell confidence intervals are wide (Table 1), so we do not claim opposition and baseline are

the same, only that the pooled estimate does not detect an opposition advantage here. Cooperation ($\rho = +0.8$), by contrast, beats baseline: $-0.707$ for IQL ($t = -2.00$, $p_{\text{raw}} = 0.048$) and $-0.924$ for VDN ($t = -3.20$, $p_{\text{raw}} = 0.0018$). Under a Holm correction across the four planned contrasts (Table 2), only the VDN cooperative contrast survives ($p_{\text{Holm}} = 0.0071$); the IQL one ($p_{\text{Holm}} = 0.144$) does not, so the single-contrast cooperative benefit rests on VDN, with the IQL direction corroborating rather than independently confirming it. The opposed-endpoint estimates are non-detections, not evidence of zero effect or equivalence; the full-grid slope decreases as target correlation becomes positive. The pooled opposition contrast also masks a reward-scale split. At $\sigma = 0.6$, the opposition gap is 0.644 (IQL) / 0.523 (VDN) below baseline ($p = 0.080$ / 0.098); at $\sigma = 1.0$, it is 0.337 / 0.205 above baseline ($p = 0.466$ / 0.609). The pooled non-detection therefore contains cancellation across reward scales, and the lower-scale arm is a directional signal rather than evidence of equivalence. Direct robust interaction tests do not resolve the heterogeneity: on the full grid the $\rho \times \sigma$ term has HC3 $p = 0.129$ (IQL) / 0.097 (VDN), and on the two endpoint cells the opposition$\times\sigma$ difference-in-differences has $p = 0.099$ / 0.157. None of the four within-$\sigma$ opposition contrasts survives a separate Holm family ($p_{\text{Holm}} = 0.319$ for both lower-scale signals; 0.932 for both higher-scale contrasts). Thus "no detected advantage" is an average over the experiment's equal $\{0.6, 1.0\}$ reward-scale mixture, not an invariant effect. The conditional estimates and interactions are regenerated by `results/adversarial_rerun/analyze.py`.

Table 2: Legacy-support training-window contrasts on the normalized gap (Welch's $t$), with Holm-adjusted $p$-values across the family of four planned contrasts. Opposition shows no detectable advantage over baseline; only cooperation beats it, and only the VDN cooperative contrast survives multiplicity correction.

| **Contrast** | **Learner** | **diff** | $t$ | **df** | $p_{\text{raw}}$ | $p_{\text{Holm}}$ |
|---|---|---|---|---|---|---|
| opposition ($-.33$) − baseline (0) | IQL | $-0.154$ | $-0.52$ | 117.7 | 0.605 | 1.000 |
| cooperation ($+0.8$) − baseline (0) | IQL | $-0.707$ | $-2.00$ | 105.8 | 0.048 | 0.144 |
| opposition ($-.33$) − baseline (0) | VDN | $-0.159$ | $-0.62$ | 117.9 | 0.537 | 1.000 |
| cooperation ($+0.8$) − baseline (0) | VDN | $-0.924$ | $-3.20$ | 111.7 | **0.00178** | **0.0071** |

Holm–Bonferroni step-down over the $K = 4$ planned contrasts (sorted raw $p$: $0.00178, 0.048, 0.537, 0.605$; multipliers $4, 3, 2, 1$ with monotone enforcement). Only the VDN cooperation-beats-baseline contrast survives at $\alpha_{\text{fw}} = 0.05$; the IQL cooperative contrast ($p_{\text{Holm}} = 0.144$) does not. The opposition contrasts are n.s. raw and adjusted. The descriptive signal is the negative target-correlation slope, corroborated by the variance decomposition (§6.2), not any single marginal contrast.

**The sign now carries the variance.** Decomposing the correlation effect on the normalized gap, the signed-linear term ($\rho$) and the quadratic term ($\rho + \rho^2$) explain nearly the same variance—signed-linear $R^2 = 0.029$ versus quadratic $R^2 = 0.029$ for IQL, 0.050 versus 0.053 for VDN—so the *sign-share* (the fraction of explained variance attributable to the linear, signed term) is 99.3% for IQL and 94.6% for VDN, with a *negative* slope ($-0.755$, $p = 1.7 \times 10^{-4}$, IQL; $-0.838$, $p = 7.3 \times 10^{-7}$, VDN): more correlation, smaller gap. On the eight-point cell-mean curve the linear fit attains $R^2 = 0.816$ (IQL) and 0.863 (VDN). This is the exact inverse of the first-stage design, where 0% of the alignment variance was sign and the relationship was $|\alpha|$-symmetric. The DGP repair flips the structure from symmetric to genuinely signed. At equal nominal $|\rho| = 0.2$, the point estimates run in the predicted direction: $\bar{g}/\sigma$ is 4.07 at $\rho = -0.2$ versus 3.65 at $\rho = +0.2$ (IQL), and 3.36 versus 3.02 (VDN). But those pairwise comparisons are non-significant ($p = 0.14$ IQL, $p = 0.16$ VDN), and the feasible magnitudes are not exactly matched ($+0.155$ versus $-0.138$; §8). The directional conclusion rests on the overall slope, not

on this pair.

**Corrected H1 verdict.** The sign-blind first-stage axis cannot adjudicate H1. On the repaired proxy, the overall target-coordinate $\rho$ slope is negative, but this is not a direct measure of objective- or reward-function alignment. Under legacy training-window scoring the $\rho = -.33$ versus 0 endpoint is statistically unresolved. Frozen legacy evaluation comes from a later deterministic rerun, not the same trained policies, and is a cross-run sensitivity: its four estimates are penalty-signed but only IQL at $\sigma = 1$ survives Holm. Under feasible-centred support the four-agent $\rho = -.33$ endpoint is detectably worse than 0 at both reward scales for both learners. Section 6.8 shows why these statements cannot be merged: the target support changes both the delivered treatment and the split between continuous feasibility and residual frozen-policy performance. Every slope in this section therefore carries that qualification.

**Reward-scale dominance is again mechanical.** The new data reproduce the degree-1 scaling result of §5.6. In a $\rho \times \sigma$ ANOVA on the *raw* gap, reward scale carries $\eta^2 = 0.235$ (IQL) / 0.225 (VDN); on the reward-scale-normalized gap the same term collapses to $\eta^2 = 0.0002$ / 0.0004, while the correlation term holds at $\eta^2 = 0.036$ / 0.058. "Stakes dominate" is once more a property of the raw scale, not of structure—identical to the first-stage finding and to the prior targeted analysis.

### 6.3 No detected advantage for strong opposition (and a directional penalty for independent learners)

The equicorrelated continuum caps opposition at $\rho = -1/3$. The two-team block design (§6.1) reaches genuine strong opposition (between-team correlation down to $-1.0$). Table 3 and Figure 3 report the raw gap ($\sigma = 1$, $n = 30$ per cell). No test detects an advantage for opposition. For VDN the minimum gap is at baseline (opp $= 0$, gap 2.838), and opposition lifts it modestly with no trend (slope $+0.097$, $p_{\text{trend}} = 0.77$; Spearman $\rho_s = -0.059$, $p = 0.43$). For IQL the gap is directionally worst at full opposition (opp $= 1.0$, gap 4.541 versus 3.434 at baseline), but we report this as a *non-significant directional trend*, not a monotone effect: the series is non-monotone ($3.43 \to 3.94 \to 3.86 \to 3.85 \to 4.38 \to 4.54$), the monotonic test is a non-detection (Spearman $\rho_s = +0.058$, $p = 0.44$), and the baseline-versus-full-opposition contrast is only marginal (Welch $t = 1.79$, $p = 0.081$). The one significant statistic is the parametric linear-slope test ($+0.979$, $p_{\text{trend}} = 0.048$), which by itself we treat as suggestive rather than conclusive. The honest reading: no advantage for opposition is detected, while the IQL point estimate at full opposition is worse than baseline but uncertain. The magnitude-only design's "neutral is uniquely worst" claim was never a valid signed result; this diagnostic does not restore it.

### 6.4 The signed shape reappears under VDN and team rewards

The VDN/team-reward bundle reproduces the signed-correlation shape and shifts its level. The gap-versus-$\rho$ curves for IQL and VDN are nearly identical in shape (Pearson curve correlation 0.976, $p = 3.5 \times 10^{-5}$; Spearman 1.000). Because the bundle changes architecture and objective together, that agreement cannot be attributed specifically to centralization. It shifts the level: VDN lowers the normalized gap relative to IQL by a mean 0.690 (95% CI $[0.652, 0.729]$; paired $t = 35.2$ over $n = 480$ matched runs), and the per-$\rho$ reduction ranges from $+0.627$ to $+0.878$, every level significant ($p < 0.05$). The reduction is roughly constant across $\rho$—the VDN/team-reward bundle does not selectively rescue the opposition region or manufacture a U. The value-decomposition-robustness claim the first-stage VDN slice could only make about the *artifact* (§5.7) now holds for the signed *shape*: the gap falls as target correlation becomes positive and no opposition advantage is detected under either IQL/individual rewards

Table 3: Two-team block design: mean raw coordination gap $\bar{g}$ versus opposition strength (between-team correlation $= -\text{opp}^2$; $\sigma = 1$, $n = 30$ per cell, 95% CI).

| opp | between-corr | IQL gap [95% CI] | VDN gap [95% CI] |
|---|---|---|---|
| 0.00 | 0.00 | 3.434 [3.001, 3.867] | **2.838** [2.480, 3.196] |
| 0.30 | −0.09 | 3.939 [3.495, 4.383] | 3.370 [2.970, 3.769] |
| 0.50 | −0.25 | 3.861 [3.145, 4.577] | 3.141 [2.575, 3.706] |
| 0.70 | −0.49 | 3.853 [3.009, 4.697] | 3.014 [2.435, 3.592] |
| 0.85 | −0.72 | 4.376 [3.283, 5.470] | 3.224 [2.502, 3.947] |
| 1.00 | −1.00 | **4.541** [3.355, 5.727] | 3.027 [2.316, 3.738] |

IQL linear slope +0.979 ($p_{\text{trend}} = 0.048$), Spearman $\rho_s = +0.058$ ($p = 0.44$, n.s.), opp=0 vs opp=1 Welch $t = 1.79$ ($p = 0.081$). VDN slope +0.097 ($p = 0.77$), minimum at baseline.

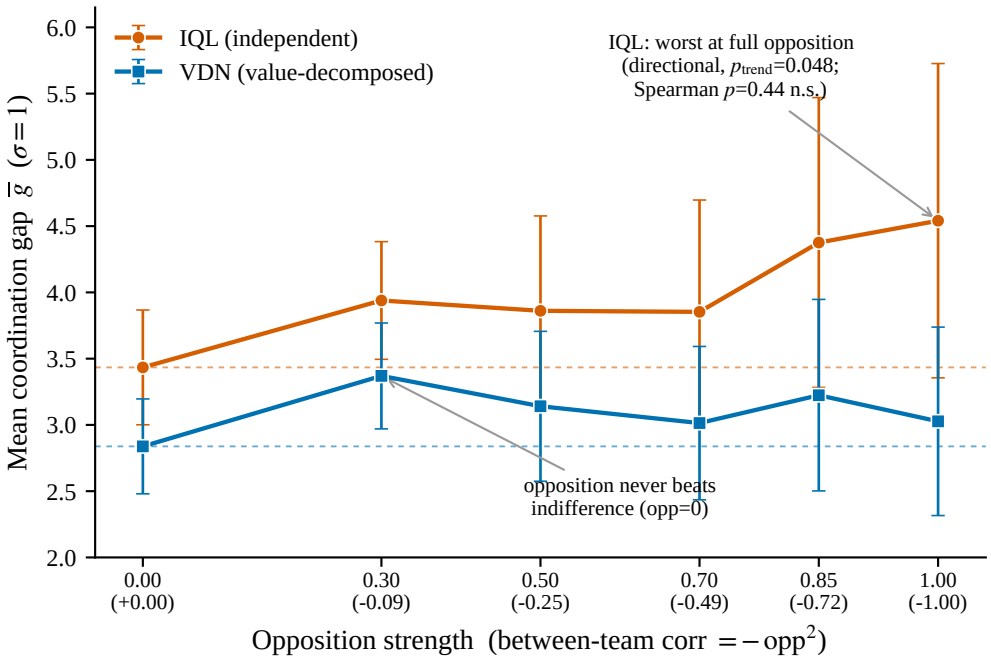


Figure 3: In the two-team block design, no test detects an advantage for opposition. Mean coordination gap $\bar{g}$ versus opposition strength (between-team correlation $= -\text{opp}^2$; $\sigma = 1$, $n = 30$ per cell, 95% CI); IQL worsens directionally toward full opposition (parametric $p_{\text{trend}} = 0.048$, but Spearman $p = 0.44$, n.s.), while VDN's minimum is at baseline.

or VDN/team rewards (but, we stress, not yet under policy-gradient or other non-value-based learners; §8).

### 6.5 Negative endpoint: no slope is detected after removing the shared state

The signed sweep estimates that, in our environment, more positive target correlation lowers the coordination gap. But *why*? Two readings are observationally equivalent in the shared-pool data. Under a *preference-correlation* reading the effect is about how aligned agents' objectives are, full stop. Under a *shared-state-contention* reading it is about agents being *physically forced* to reconcile divergent targets in one common resource state: anti-aligned targets make agents pull the shared pool in opposite directions, and that contention—not the correlation as such—is what raises the gap. A word on what to call this. Our environment shares with a common-pool resource the feature that does the work here—one state, jointly controlled, which cannot simultaneously satisfy divergent ideal points—and it is that rivalry in *control* which the separable twin removes. But it lacks the features that make a common-pool resource a distinctive institutional problem: nothing is appropriated, extracted, depleted or replenished, there is no stock that shrinks with use and no withdrawal that forecloses another agent's withdrawal. Rewards are quadratic losses on a shared position, not returns on a consumed good. We therefore describe the environment as a *shared-state contention* task and treat the commons literature (Ostrom, 1990; Hardin, 1968) as an analogy for the rivalry structure rather than a claim that we have modeled a commons; readers should not import tragedy-of-the-commons mechanisms into our results. The MARL setting that *does* implement appropriation directly—emergent defection and conflict over a depletable pool—is Pérolat et al. (2017) (with the related Gathering/Wolfpack sequential social dilemmas of Leibo et al., 2017), and the difference is worth keeping in view when comparing our findings to theirs. Where Pérolat et al. (2017) show that a learned tragedy and reward inequality *emerge* in a common-pool game, our separable-resource twin supplies a negative endpoint: the target-correlation effect is detected with shared-state coupling and not detected in the fully factorized twin. This does not distinguish physical contention from every possible form of multi-agent interaction; it distinguishes the shared-pool environment from the separable one tested here.

**Design.** We re-run the signed equicorrelated sweep in a *separable*-resource variant of the environment: each agent acts on its *own* independent resource pool (its $\pm 1$ actions update only its own state, with no summation into a shared state) and is rewarded by its own pool against its own target. Everything else is held fixed—the same signed equicorrelated DGP, the same $\rho$ grid $\{-.33, -0.2, -0.1, 0, 0.2, 0.4, 0.6, 0.8\}$, the same $\sigma \in \{0.6, 1.0\}$, the same 30 replication indices, the same IQL/VDN networks, optimizers, $\varepsilon_{\text{exp}}$-schedule, replay, and matched condition streams (only four marked environment-dynamics lines differ). For IQL, the marginal target distribution is invariant to $\rho$ and each agent's training problem factorizes, so the expected curve is flat by construction. VDN retains additive team-reward and loss coupling during training, so finite-sample learning need not be strictly invariant even though the underlying control problem is separable. The arm is therefore a negative endpoint and scope check, not unique mechanism identification.

**Definition 6.1** (Separable-resource MDP)**.** The state is $S_t = (s_{1t}, \ldots, s_{nt}) \in [0, C]^{n \times m}$, with independent resets $s_{i0} \sim \text{Uniform}([0, C]^m)$. Agent $i$ observes only $\tilde{s}_{it} = s_{it} + v_{it}$, retains the action $a_{it} \in \{-1, 0, +1\}^m$, and moves only its own pool: $s_{i,t+1} = \text{clip}(s_{it} + a_{it}, 0, C)$. Its reward is $R_i(s_{i,t+1}) = -\sum_j w_{ij}(s_{ij,t+1} - \tau_{ij})^2$. IQL trains on $R_i$; VDN trains its additive joint value on $R_{\text{team}} = \sum_i R_i$, as in the shared arm. Targets, weights, horizon, clipping, exploration, replay and network sizes otherwise match Definition 3.1 and the signed design.

**Result: the separable slopes are not detected.** Table 4 and Figure 4 report the normalized gap $\overline{g}/\sigma$ across $\rho$ for shared versus separable pools. In the shared pool the gap tracks $\rho$ strongly and significantly for both learners (slope $-0.755$, $p = 1.7 \times 10^{-4}$, IQL; $-0.838$, $p = 7.3 \times 10^{-7}$, VDN; one-way ANOVA over $\rho$ significant for both). With separable pools the point slopes are near zero and not detected (IQL $-0.029$, $p = 0.87$; VDN $-0.053$, $p = 0.76$), the one-way ANOVA over $\rho$ does not detect a difference (IQL $F = 0.90$, $p = 0.51$; VDN $F = 0.89$, $p = 0.52$), and neither the cooperation-versus-baseline contrast (IQL $p = 0.67$, VDN $p = 0.71$) nor the opposition-versus-baseline contrast (IQL $p = 0.24$, VDN $p = 0.38$) approaches significance. The shared and separable gap-versus-$\rho$ curves are essentially unrelated (shape correlation $+0.22$ IQL, $+0.15$ VDN). Tellingly, the separable gap for IQL sits at roughly the shared-pool *baseline* level (3.90 against 4.02 at $\rho = 0$): the positive-target-correlation dividend—the drop *below* that baseline as $\rho$ rises—is precisely what is not detected once the pool is no longer shared. For VDN the level comparison does not hold and we do not make it: its separable gap (4.05) sits above *every* shared-pool cell it has, including its own worst. Moving to the separable design therefore raises VDN's level more than IQL's, which is the opposite of what a team-reward learner's advantage in the shared pool would suggest, and a further reason to read that advantage as an objective effect rather than an architectural one (§6.4). The mechanism is intuitive: with independent pools each agent steers its own state to its own target irrespective of what the others want, so $\rho$ has nothing to bind.

Table 4: Separable-resource control: normalized gap $\overline{g}/\sigma$ by signed correlation $\rho$, shared pool versus separable pools (pooled over $\sigma \in \{0.6, 1.0\}$, $n = 60$ per cell). IQL is structurally $\rho$-invariant; the VDN separable slope is not detected.

| | **Shared pool** | | **Separable pools** | |
|---|---|---|---|---|
| $\rho$ | **IQL** | **VDN** | **IQL** | **VDN** |
| $-.33$ | 3.86 | 3.20 | 3.87 | 4.05 |
| $-0.20$ | 4.07 | 3.36 | 3.97 | 4.13 |
| $-0.10$ | 4.08 | 3.38 | 3.94 | 4.09 |
| 0.00 | 4.02 | 3.35 | 4.09 | 4.22 |
| $+0.20$ | 3.65 | 3.02 | 3.52 | 3.66 |
| $+0.40$ | 3.39 | 2.76 | 3.97 | 4.15 |
| $+0.60$ | 3.34 | 2.69 | 3.85 | 3.97 |
| $+0.80$ | 3.31 | 2.43 | 3.96 | 4.11 |
| slope($\rho$) | $-0.755$ | $-0.838$ | $-0.029$ | $-0.053$ |
| $p$(slope) | $1.7\times10^{-4}$ | $7.3\times10^{-7}$ | 0.87 | 0.76 |

**Scope, stated honestly.** This control is a *negative endpoint and scope check*, not a unique mechanism identification. For IQL, $\rho$-invariance is definitional because there is no common state, reward or learner through which cross-agent target correlation can affect a marginal control problem. VDN preserves additive team-reward coupling in its loss, making the non-detected slope a finite-training check, but the underlying MDP still has an additive separable optimum. Together they bound the detected effect to the tested state coupling; equivalence is not established. This does not exclude reward coupling, observation coupling, communication or other interaction structures as channels in other environments. An interaction-preserving non-contention arm would be needed for that stronger claim. One further caveat about the legacy control's construction: the separable gap has an irreducible, $\rho$-invariant floor (targets $\sim \mathcal{N}(0,1)$ are half-unreachable under the $[0, C]$ clamp at every $\rho$), so the separable result con-

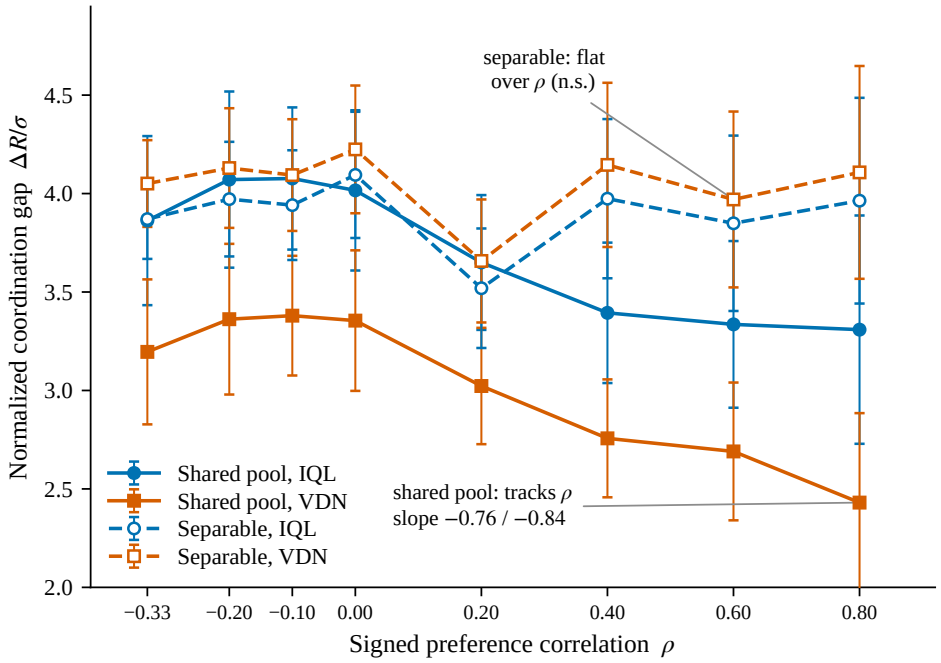


Figure 4: The signed-correlation slope is detected only in the shared-state variants tested here. Normalized gap $\bar{g}/\sigma$ versus $\rho$: shared-pool curves (solid) decline into cooperation. The separable IQL expectation is $\rho$-invariant; the VDN slope is not detected and equivalence is not established. The positive-target-correlation dividend is not detected at the separable endpoint. Absolute levels are not compared across learners.

cerns the $\rho$-invariant IQL expectation and the non-detected VDN slope, not the absolute level. The centred in-box follow-up removes that floor and is reported separately below. The legacy shape is consistent with a shared-state reading, but does not identify a unique causal mechanism. A second, design-level point: this control contrasts only the *two extremes* of resource coupling—fully shared and fully separable—establishing a strong $\rho$-slope under full sharing, structural invariance for separable IQL and a non-detected separable VDN slope. It does not, on its own, show smooth scaling with the degree of state coupling. The next control (§6.6) supplies a parametric *partial-sharing* sweep that varies the fraction of resources fed into the common pool and finds monotone steepening in the four gap-slope point estimates. This grades sensitivity to state coupling within the construction; it is not a causal mediation result.

**Target-support control.** The signed legacy design draws targets from $\mathcal{N}(0, \Sigma(\rho))$ while the state is confined to $[0, 10]^m$, so roughly half of target coordinates are below the lower wall. The construct-validity follow-up uses the translation- equivalent design $\mathcal{N}(5, \Sigma(\rho))$ and rejects the vanishingly rare out-of-box draw. A 20,000-draw gate at every $\rho$ level puts all checked targets in $[0, 10]$ and reproduces the requested correlations to within .001. The legacy and feasible-centred rows share the same underlying Gaussian, weight and replication streams; condition and evaluation seeds plus row-level problem-instance and latent-problem hashes verify matching after subtracting the translation. This is an exploratory support-sensitivity analysis, not a second confirmatory experiment. The clean frozen separable slopes remain small and are not detected after translation: $+0.004$ (95% CI $[-.018, .026]$, HC3 $p = .718$) IQL and $+0.020$ ($[-.019, .058]$, $p = .318$) VDN. The paired support-change slopes are also unresolved after

Holm correction ($p = .393$ and $.448$). Thus centring targets strengthens the signed effect only in the shared environment; it does not manufacture a $\rho$ gradient in the separable endpoint.

For auditability, the paired support model is formed after a one-to-one pivot on the block key

$$b = (\rho, \sigma, \text{replication}).$$

It sets

$$D_b = g_b^{\text{feasible}} - g_b^{\text{legacy}}, \qquad D = X\beta + u, \qquad X = [\mathbf{1}, \rho],$$

with HC3 covariance. The support hashes are checked before forming $D$; no unpaired level regression is labelled as paired.

### 6.6 Partial sharing on legacy-Gaussian-at-zero support: the slope grades with the fraction of shared coordinates

The separable control of §6.5 contrasts only the two endpoints of resource coupling—a fully shared pool against fully separable pools. It leaves open whether the gap-versus-$\rho$ slope is an all-or-nothing property of *any* shared state, or whether it is *graded* in the shared-coordinate fraction. We probe this with a parametric *partial-sharing* control that interpolates between the two endpoints and asks whether the slope scales monotonically with the degree of overlap. This paired series retains the legacy-Gaussian-at-zero target support; it must not be numerically conflated with the separate feasible-centred endpoint control.

**Overlap construction.** With $m = 3$ resources, an overlap level $w = k/m$ makes the first $k$ ordered coordinate positions a *common pool*—all four agents act on it and their per-step requests are summed into one shared state (contention)—and the remaining $m - k$ resources *private*, each agent acting only on its own copy (separable). We sweep the four reachable levels $w \in \{0, 1/3, 2/3, 1\}$ ($k = 0, 1, 2, 3$). The preference distribution and training protocol are held fixed across overlap: every level uses the same signed equicorrelated DGP with legacy-Gaussian-at-zero target support, the same $\rho$ grid, $\sigma \in \{0.6, 1.0\}$, $\varepsilon = 0$, 30 replications, and the same IQL/VDN networks, optimizers, $\varepsilon_{\text{exp}}$-schedule and replay. Within the partial sweep, one base seed is assigned to each $(\rho, \sigma)$ cell and reused at all four overlap levels. Preferences, replication blocks and initial network weights are therefore paired over $w$; only environment coupling changes. The same seed schedule is used by the standalone shared and separable runners, so the two endpoint cells are reused rather than retrained.

**Definition 6.2** (Partial-sharing MDP)**.** For overlap $k \in \{0, 1, 2, 3\}$, the implementation assigns the first $k$ ordered coordinate positions to the shared block. Write the state as one shared block $x_t \in [0, C]^k$ and private blocks $y_{it} \in [0, C]^{m-k}$. Shared and private coordinates reset independently and uniformly. Agent $i$ observes the length-$m$ concatenation $\tilde{o}_{it} = (x_t, y_{it}) + \nu_{it}$—not the other agents' private blocks—and retains $a_{it} \in \{-1, 0, +1\}^m$. Shared coordinates follow $x_{t+1} = \text{clip}(x_t + \sum_i a_{it}^{\text{sh}}, 0, C)$; private coordinates follow $y_{i,t+1} = \text{clip}(y_{it} + a_{it}^{\text{pr}}, 0, C)$. Rewards concatenate the two post-state blocks against the same agent-specific target, $R_i = -\sum_{j \le k} w_{ij}(x_{j,t+1} - \tau_{ij})^2 - \sum_{j > k} w_{ij}(y_{ij,t+1} - \tau_{ij})^2$. IQL uses $R_i$ and VDN uses $\sum_i R_i$. Thus $k = 0$ is Definition 6.1 and $k = m$ is Definition 3.1 at the level of state, observation, transition, action and reward maps.

**Validity check: the endpoints reduce to the existing runners.** The formal construction collapses exactly onto the two endpoint MDP maps. A deterministic short-run smoke test separately runs the stan-

dalone and partial implementations at $k = 0$ and $k = m$ for both learners and matches their reported training reward, three frozen-evaluation rewards, convergence time and agent-reward summaries to $10^{-6}$. It does not separately instrument internal draws, networks, transitions or reset banks. The full follow-up therefore trains only the two genuinely new intermediate arms ($k = 1,2$) and assembles $k = 0,3$ from their already-scored endpoint files. This removes the former independent- redraw ambiguity and makes inference over $w$ a paired problem-block design; standard errors for the continuous $\rho \times w$ interaction are clustered by $(\rho, \sigma, \text{replication})$.

**Result: a paired frozen-policy overlap response.** Table 5 and Figure 5 report HC3 per-overlap slopes of the clean held-out normalized gap. The point estimates steepen monotonically with the shared-coordinate fraction under the fixed first-$k$ coordinate ordering: from $-0.219$ at the separable IQL endpoint to $-0.566$, $-0.807$ and $-1.156$, and from $-0.125$ to $-0.594$, $-0.895$ and $-1.162$ for VDN. The separable slopes are not detected; every level with at least one shared coordinate is negative at $p < .005$. The primary paired interaction clusters by the $(\rho, \sigma, \text{replication})$ problem block: $\rho \times w = -0.916$ (95% CI $[-1.304, -0.528]$, $p = 3.7 \times 10^{-6}$) for IQL and $-1.023$ ($[-1.260, -0.787]$, $p = 2.0 \times 10^{-17}$) for VDN. Categorical slope modulation is also detected ($p = 3.6 \times 10^{-6}$ / $3.1 \times 10^{-16}$), while departures of those four slopes from a linear ramp are not ($p = .890$ / $.191$). Thus the paired held-out design supports graded modulation and does not support an all-or-nothing threshold; four overlap levels still do not establish the exact functional shape. Because overlap also changes transition gain, reachable-lattice, clipping, travel geometry and credit-assignment conditions, this is an interaction and scope result, not a formal causal mediation estimate.

Table 5: Paired legacy-Gaussian-at-zero-support partial-sharing overlap response: per-overlap $\rho$-slope of the clean frozen normalized gap (HC3 95% CI), with 100 held-out resets per trained replication ($m = 3$, 30 replications, $\sigma \in \{0.6, 1.0\}$, $\varepsilon = 0$; $n = 1920$ per learner). The endpoint gate makes $w = 0$ and $w = 1$ exactly the separable and shared runners. The separate problem-block-clustered $\rho \times w$ tests are $-0.916$ ($p = 3.7 \times 10^{-6}$) IQL and $-1.023$ ($p = 2.0 \times 10^{-17}$) VDN.

| | **IQL** | | | **VDN** | | |
|---|---|---|---|---|---|---|
| overlap $w$ ($k$) | slope | 95% CI | $p$ | slope | 95% CI | $p$ |
| 0 ($k$=0, separable) | $-0.219$ | $[-0.563, +0.126]$ | .213 | $-0.125$ | $[-0.496, +0.246]$ | .509 |
| $1/3$ ($k$=1) | $-0.566$ | $[-0.956, -0.176]$ | .0045 | $-0.594$ | $[-0.949, -0.240]$ | .0010 |
| $2/3$ ($k$=2) | $-0.807$ | $[-1.297, -0.318]$ | .0012 | $-0.895$ | $[-1.263, -0.528]$ | $1.8 \times 10^{-6}$ |
| 1 ($k$=3, shared) | $-1.156$ | $[-1.695, -0.618]$ | $2.6 \times 10^{-5}$ | $-1.162$ | $[-1.515, -0.808]$ | $1.2 \times 10^{-10}$ |

**The overlap response is in the slope, not a difficulty level.** One confound deserves a direct check: more shared resources could raise multi-agent credit-assignment difficulty independent of $\rho$, inflating the baseline gap and masquerading as a steeper overlap interaction. It does not occur here. With problem-block clustering, the $\rho = 0$ frozen baseline instead declines with overlap: slope $-0.390$ (95% CI $[-0.659, -0.121]$, $p = .0045$) IQL and $-0.849$ ($[-1.151, -0.546]$, $p = 3.7 \times 10^{-8}$) VDN. An unpaired HC3 sensitivity is less precise for IQL ($p = .119$) but retains the negative sign; VDN remains detected ($p = .0018$). This is opposite the proposed baseline-inflation confound. Moreover, an additive baseline shift across $w$ cannot create a $\rho \times w$ coefficient. The one piece this design cannot rule out is a difficulty $\times \rho$ *interaction*—harder credit assignment at high $w$ amplifying $\rho$-sensitivity—since raising $w$ moves contention and credit-assignment difficulty together. It also changes transition gain and, on the affected coordinates, reachable-lattice, clipping and travel geometry. The control therefore identifies

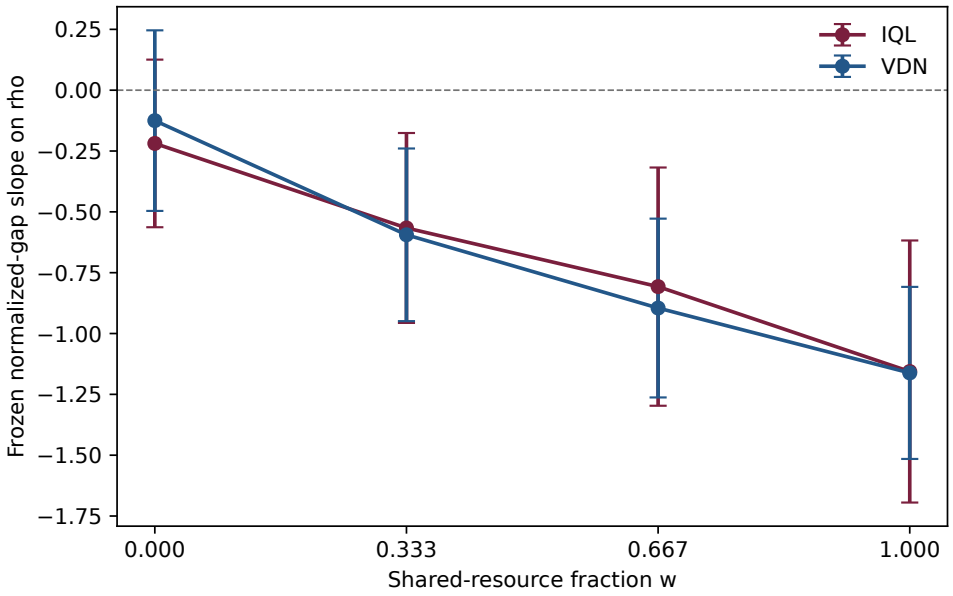


Figure 5: Paired legacy-Gaussian-at-zero-support partial-sharing overlap response on the frozen held-out outcome. The normalized-gap slope point estimates descend monotonically over the four tested fixed first-$k$ shared-coordinate configurations; error bars are HC3 95% CIs. The problem-block-clustered linear interactions are reported in the text and the four categorical slopes do not depart detectably from a linear ramp.

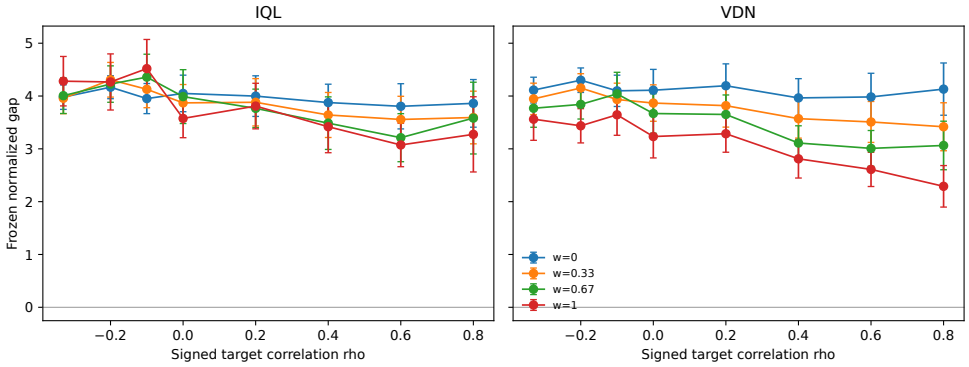


Figure 6: Clean frozen normalized gap versus $\rho$ on paired legacy-Gaussian-at-zero target support, one curve per overlap level and learner. At the separable endpoint IQL is structurally $\rho$-invariant and VDN is not detected; the slope point estimates steepen monotonically as shared coordinates are added.

modulation with the shared-coordinate fraction in these fixed-order variants, not a unique psychological, algorithmic or dynamical mechanism.

**Relationship to the earlier unpaired surface analysis.** Before the endpoint files were reassembled into the paired problem-block follow-up, the four-level analysis gave HC3 linear $\rho \times w$ estimates of $-0.765$ (95% CI $[-1.348, -0.183]$, $p = .0100$) for IQL and $-0.731$ ($[-1.260, -0.203]$, $p = .00670$) for VDN. In that older sensitivity, the three categorical slope deviations were not jointly detected at .05 ($p = .0633/.0533$), nor was departure from a linear ramp ($p = .8896/.8869$); the better categorical full-surface fit came from nonlinear baseline shifts. Those non-detections remain part of the forensic record. The paired, problem-clustered follow-up above is the primary analysis because it preserves problem identity across overlap and reuses the verified endpoints.

**Caveats, stated honestly.** Every partially or fully shared slope is now detected, so the old suggestion of a majority-shared threshold is withdrawn. The four estimates do not depart detectably from a linear ramp ($p = .890$ IQL; .191 VDN), but failure to detect curvature with only four overlap levels is not proof of exact linearity. A finer grid (larger $m$, hence finer $k/m$ steps) is needed to distinguish smooth curvature from a linear response. Within those bounds the paired upgrade is clean: the overlap modulation is no longer a two-point contrast or an artifact of independently redrawn endpoints, but a graded change in the mean frozen-policy slope.

### 6.7 Control: at latent-target $\rho = -1$, opposition shows no detectable advantage

The equicorrelated continuum caps opposition at $\rho = -1/3$ (the positive-semidefinite bound for $n = 4$), and the two-team block design that reaches stronger opposition (§6.3) confounds strong *between*-team opposition with strong *within*-team cooperation. Neither gives a latent-target point at $\rho = -1$. We obtain one by reducing to $n = 2$: with a single pair, the signed equicorrelated sampler reaches full $\rho = -1$ (verified $-1.0000$) with *no* within-team-cooperation confound and *no* PSD cap. This is the cleanest signed *target* correlation the design admits—though not, as an earlier version of this paragraph claimed, the cleanest adversarial test outright: the box constraint attenuates it more than anywhere else in the paper, quantified below.

**Design.** `n2`: the same shared-pool environment, monkeypatched only to $n = 2$ agents and the signed equicorrelated sampler, over $\rho \in \{-1.0, -0.6, -0.2, 0, 0.4, 0.8\}$, $\sigma \in \{0.6, 1.0\}$, IQL and VDN, 30 replications, 1,000 episodes. Contrasts are all *within* the $n = 2$ sweep ($n = 60$ per cell, pooled over $\sigma$); $n = 2$ and $n = 4$ absolute gaps are not directly comparable and we do not compare them. We stress that reducing to $n = 2$ to reach latent-target $\rho = -1$ *confounds agent count with the opposition condition*: two agents face less multi-agent interference than four—fewer simultaneous claims on the shared pool, fewer collisions, an easier exploration and credit-assignment problem—and that reduction in interference could independently flatten the gap or soften any opposition penalty, quite apart from the change in $\rho$. We therefore use the $n = 2$ sweep only for its *internal* opposition-versus-baseline-versus-cooperation ordering at latent-target $\rho = -1$, and do *not* compare its shape, slope, or levels against the $n = 4$ equicorrelated baseline: the non-comparability we already flagged for absolute levels extends to the relative and shape claims.

One further qualification, and it bites hardest exactly here. "Clean" and "unconfounded" refer to the *latent* targets: the sampler realizes $\rho = -1.0000$ between them, verified. But the agents contend over a state confined to $[0, C]^m$, and the correlation of their *feasible* optima is attenuated by that box (§8), most

severely at the opposed end. Measured directly, the $n = 2$ nominal $\rho = -1$ realizes a feasible correlation of $-0.468$: a retention of 0.47, against 0.66 at the $n = 4$ opposed end and 0.93 at the cooperative one (`results/adversarial_rerun/feasible_correlation.py`). The strongest opposition the paper can specify is therefore the most heavily attenuated treatment it administers. So $\rho = -1$ names the strongest opposition our design can specify, not the strongest an agent pair can experience; the non-detection reported below concerns a treatment weaker than its label. This does not affect the ordering we read off, which is internal to the sweep, but it does mean the $n = 2$ control bounds opposition less tightly than the shorthand "latent-target $\rho = -1$" suggests on its own.

**Result: opposition shows no advantage over (or fragilely loses to) baseline; cooperation beats it only at strong alignment.** Table 6 and Figure 7 report the gap. At latent-target $\rho = -1$ opposition does *not* beat baseline for either learner. For VDN there is *no statistically significant difference* from baseline (diff $+0.345$, 95% CI $[-0.42, +1.11]$, Welch $t = 0.90$, $p = 0.37$; Mann–Whitney $p = 0.85$); the interval is wide and straddles zero, so this is a non-detection rather than a demonstration of equivalence. For IQL opposition is directionally *worse* but only fragilely so: the pooled Welch test is significant on the mean (diff $+1.270$, 95% CI $[+0.05, +2.49]$, $t = 2.07$, $p = 0.041$)—opposition is the *worse* arm—but the result is tail-driven: the rank-based Mann–Whitney comparison is not detected ($p = 0.24$) and neither $\sigma$ level is significant alone ($p = 0.24$ at $\sigma = 0.6$, $p = 0.09$ at $\sigma = 1.0$); the opposition arm has a heavy right tail (sd 4.27 versus 2.08 at baseline), so a handful of badly coordinating runs inflate the mean more than the median. The honest reading is that no consistent advantage for opposition is detected; the IQL mean instead points toward a larger gap. Cooperation ($\rho = +0.8$), by contrast, beats baseline for both learners (IQL $-1.083$, $p = 0.0045$; VDN $-0.677$, $p = 0.037$; these $n = 2$ contrasts, like the team contrasts, are reported raw—only the four equicorrelated planned contrasts of Table 2 carry the Holm correction—so the VDN value is marginal). The cooperative benefit is moreover confined to the extreme: intermediate cooperation ($\rho = +0.4$) does *not* follow the gradient—it ties baseline for IQL (4.04 vs. 4.04) and is slightly worse for VDN (3.47 vs. 3.28), so the $n = 2$ positive-target-correlation effect is a strong-alignment phenomenon ($\rho = +0.8$), not a monotone trend across the cooperative range (within this $n = 2$ design). One wrinkle: the single worst point is $\rho = -0.6$ rather than $-1$ for both learners, so the opposition arm *plateaus/peaks* in the opposition region rather than rising strictly monotonically to $-1$. In gap terms the opposition-region point estimates generally sit above baseline and the strong-cooperation point sits below it, but only the $\rho = +0.8$ cooperative drop is significant for both learners; the opposition comparison is fragile for IQL and not detected for VDN or under the IQL rank test.

### 6.8 Control: dynamic benchmarking separates feasibility from frozen-policy performance

The static quantity $R^{\text{joint}}$ in Equation (6) is not an appropriate learning benchmark. It is the continuous single-state optimum, while a controller starts from a random reset, can move only on the reset-specific integer-reachable lattice, and pays a finite- horizon travelling transient. The archived outcome also averages the final 100 *training* episodes with ongoing parameter updates and a 1% action-exploration floor. Calling $R^{\text{joint}} - \bar{R}$ a "learning shortfall" therefore bundles discretization, transient, exploration and policy suboptimality.

**Matched frozen protocol and exact oracle.** For each trained replication we freeze all parameters and evaluate greedy policies for 100 episodes on a held-out reset bank. We also score the same starts with matched observation noise and with the 1% action-exploration floor restored. Because the signed support, separable and partial-sharing controls fix $\varepsilon = 0$, matched-noise and clean scores are identical there

Table 6: $n = 2$ latent-target strong-opposition control: normalized gap $\overline{g}/\sigma$ by signed target correlation $\rho$ (verified $\rho = -1.0000$ reachable, but feasible-optimum correlation $-0.468$; pooled over $\sigma \in \{0.6, 1.0\}$, $n = 60$ per cell). Opposition ($\rho = -1$) shows no detectable advantage over (VDN) or fragilely loses to (IQL) baseline ($\rho = 0$); only strong cooperation ($\rho = +0.8$) lowers the gap—intermediate $\rho = +0.4$ does not.

| $\rho$ | IQL gap | VDN gap |
|---|---|---|
| $-1.00$ | 5.31 | 3.62 |
| $-0.60$ | **5.39** | **3.97** |
| $-0.20$ | 4.93 | 3.81 |
| 0.00 | 4.04 | 3.28 |
| $+0.40$ | 4.04 | 3.47 |
| $+0.80$ | **2.96** | **2.60** |

Bold: worst (max) and best (min) cell per learner. Headline contrasts (pooled, $n = 60$): opposition−baseline IQL $+1.270$ (Welch $p = 0.041$ pooled-only; MWU $p = 0.24$, n.s.), VDN $+0.345$ (Welch $p = 0.37$; MWU $p = 0.85$); cooperation−baseline IQL $-1.083$ ($p = 0.0045$), VDN $-0.677$ ($p = 0.037$).

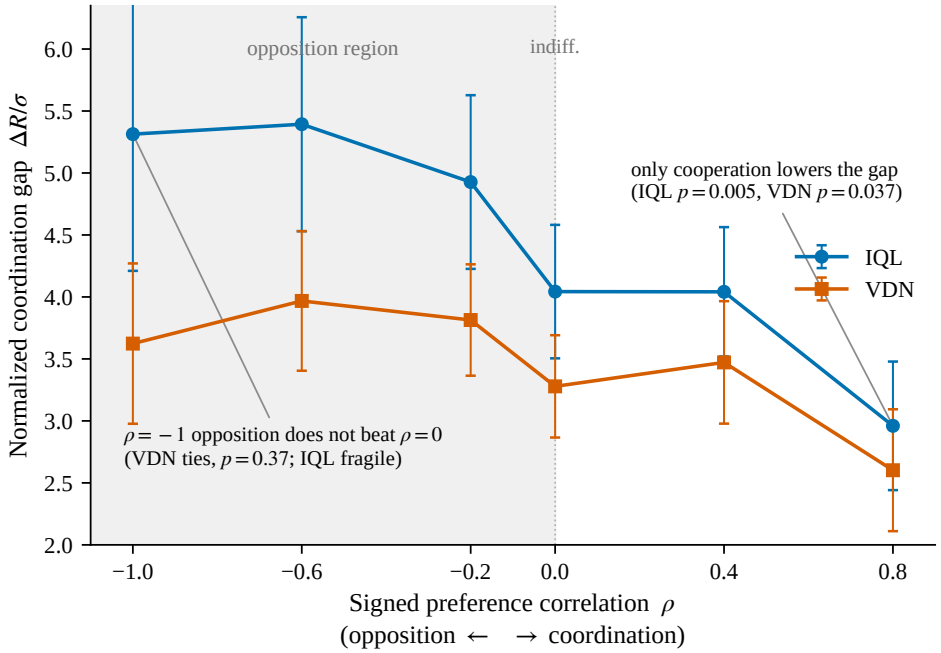


Figure 7: At latent-target $\rho = -1$ ($n = 2$, no PSD cap or within-team-cooperation confound, but an agent-count confound and feasible correlation $-0.468$), opposition shows no detectable advantage over baseline; only strong cooperation reduces the gap. Normalized gap $\overline{g}/\sigma$ versus $\rho$ for IQL and VDN (95% CI, $n = 60$ per cell). The opposition region ($\rho < 0$) is shaded; baseline ($\rho = 0$) is marked. Opposition shows no detectable advantage over baseline for VDN and is only fragilely worse for IQL; the opposition arm peaks near $\rho = -0.6$ rather than at $-1$.

by construction; their restored-exploration scores remain distinct. The reduced signed-noise crossing is the only follow-up in which matched observation noise differs from clean scoring. For the clean greedy outcome $R_T^{\pi}$, an exact $T = 100$ dynamic oracle yields

$$R^* - \mathbb{E}R_T^{\pi} = \underbrace{R^* - R^{\text{cont}}}_{\Phi \text{ continuous feasibility}} + \underbrace{R^{\text{cont}} - \mathbb{E}R^{\text{lat}}}_{Q \text{ reachable-lattice cost}} + \underbrace{\mathbb{E}(R^{\text{lat}} - R_T^{\text{dyn}})}_{T_c \text{ travel transient}} + \underbrace{\mathbb{E}(R_T^{\text{dyn}} - R_T^{\pi})}_{P \text{ frozen-policy residual}} . \quad (8)$$

$R^*$ and $R^{\text{cont}}$ are deterministic conditional on a target-weight instance; $R^{\text{cont}}$ uses its clipped reward-scale-weighted target mean, while $R^{\text{lat}}$ is the best state on the exact reset-specific 21-point lattice per resource; and $R_T^{\text{dyn}}$ is computed by backward induction under the same transition constraints and starts. The full three- resource dynamic program agrees with the separable per-resource implementation to $3.98 \times 10^{-13}$, and a separate implementation passes 54/54 two-step brute-force checks, including integer and boundary starts. The identity above holds per replication to numerical tolerance; expectations average the held-out reset bank (and any evaluation randomness) for that fixed instance. $P$ is the right residual for learned-policy performance against joint control, but we do not call it pure learning friction: for IQL it also contains decentralization and the difference between individual training rewards and the oracle's team objective.

**The target-support choice changes the answer.** Table 7 reports HC3 slopes of the reward-scale-normalized components on $\rho$ ($n = 480$ per learner and support). Under the legacy $\mathcal{N}(0, \Sigma(\rho))$ targets, the clean frozen gap declines as target correlation becomes positive ($-1.156$ IQL; $-1.162$ VDN), but the residual $P$ is not detected: $+0.136$ (95% CI $[-.228, .500]$, $p = .463$) IQL and $+0.130$ ($[-.001, .261]$, $p = .051$) VDN. These are non-detections, not equivalence results. Continuous feasibility supplies more than the entire gradient and the other components offset it. Under the centred in-box design, however, the clean frozen gradient steepens to $-3.602$ IQL and $-4.210$ VDN. Continuous feasibility contributes $-2.276$ for both, while $P$ contributes $-1.323$ IQL and $-1.931$ VDN. Thus feasibility explains about 63% of the IQL gradient (95% matched-problem-block-bootstrap interval 56–72%) and 54% of the VDN gradient (51–57%), with residual frozen-policy performance explaining about 37% (28–44%) and 46% (43–49%), respectively. These are ratios of jointly resampled component and total slopes, not constrained mixture proportions. The bootstrap uses 10,000 draws, seed 20260824, resampling replication/problem blocks within each $(\rho, \sigma)$ cell and reusing the draw across supports, learners, numerator and denominator. Lattice and transient slopes nearly cancel. The paired feasible-minus- legacy change in the frozen gap slope is $-2.446$ IQL and $-3.049$ VDN (HC3, Holm-adjusted $p < 10^{-8}$ for both). The boundary-mismatched design materially attenuated the signed effect and left its estimated policy component undetected. It also changes the opposition comparison across outcome protocols: on frozen legacy evaluation the four stake-by-learner point estimates are penalty-signed, but only IQL at $\sigma = 1$ survives the four-test Holm family (difference $+1.016$, adjusted $p = .034$). That isolated result does not replicate across the other stake or learner cells; the feasible-centred design below supplies the consistent comparison.

**Current exploratory reading.** The supported exploratory claim is no longer that positive target correlation lowers the gap *only* by making a common state feasible. Rather, both target supports show a positive-target-correlation decline in the frozen gap, and the exactly feasible support shows two channels: positively correlated targets make a high-reward shared state more attainable and reduce the learned policies' residual to a matched finite-horizon joint controller. The size and even presence of the latter

Table 7: Exact finite-horizon decomposition of the clean frozen-policy gap. Entries are HC3 slopes of each reward-scale-normalized component on $\rho$; component slopes sum to the frozen-gap slope. The legacy support makes the estimated gradient feasibility-dominated, with $P$ not detected. With every target inside the state box, both continuous feasibility and the residual to the exact dynamic oracle improve as target correlation becomes positive.

| Support | Learner | Gap | $\Phi$ | $Q$ | $T_c$ | $P$ |
|---|---|---|---|---|---|---|
| Legacy Gaussian-at-zero | IQL | $-1.156$ | $-1.336$ | $+0.031$ | $+0.012$ | $+0.136$ |
| Legacy Gaussian-at-zero | VDN | $-1.162$ | $-1.336$ | $+0.031$ | $+0.012$ | $+0.130$ |
| Feasible centred | IQL | $-3.602$ | $-2.276$ | $+0.057$ | $-0.059$ | $-1.323$ |
| Feasible centred | VDN | $-4.210$ | $-2.276$ | $+0.057$ | $-0.059$ | $-1.931$ |

channel are support-dependent. At the tested four-agent endpoint the feasible-centred control detects a consistent contrast: $\rho = -.33$ versus 0 raises the frozen gap at both reward-scale levels and for both learners, with all four within-$\sigma$ contrasts surviving Holm correction. The earlier training-window pooled non-detections and the mixed frozen legacy result were therefore not evidence that opposition is harmless; they were bounded to a treatment whose ideal points were often outside the environment. This is the fourth construct lesson: benchmark and target support must be audited together before a gap is interpreted as learned coordination.

### 6.9 Verdict: the honest core, and its scope

Across the full control battery, four apparent findings enter and none leaves intact: three lose their original status, while the fourth keeps its effect but loses its attribution. First, the *dominance* (ranking) of the reward-scale proxy in the raw $\eta^2$ partition is mechanical and scale-dependent—the ranking inverts under reward-scale normalization—a degree-1 scaling artifact (§5.6, reproduced here in §6.2); the *direction* H2 asserts (higher stakes raise friction) is supported throughout, it is only the "stakes are the dominant factor" ranking that is an artifact of the $\sigma$ grid. Second, the symmetric *U-shape* of §5 is a sign-blind data-generating-process artifact: with a genuinely signed correlation it disappears entirely, along with its phantom adversarial arm. Third, the apparent general effect of target correlation is bounded to shared-state contention (§6.5, §6.6). The fully separable IQL endpoint is structurally $\rho$-invariant because each agent's marginal problem is $\rho$-invariant; the VDN slope is not detected despite retaining team-reward coupling, and equivalence is not established. The stronger evidence is the paired legacy-Gaussian-at-zero-support, interaction-preserving partial-sharing series, whose linear $\rho \times w$ term detects graded modulation without resolving the exact surface shape. Fourth, a narrow same-signed estimate remains—*more positive target correlation lowers the frozen-policy gap*—but target support and the benchmark determine how it should be read. With legacy out-of-box targets continuous feasibility contributes more than the point-estimated gradient, while $P$ is not detected ($+.136$, $p = .463$, IQL; $+.130$, $p = .051$, VDN); absence or equivalence is not established. Opposition–baseline is unresolved. With feasible-centred targets the slope is three to four times larger, the four-agent $\rho = -.33$ endpoint is worse than 0 at both reward scales after multiplicity control, and an exact finite-horizon decomposition assigns the target-correlation gradient to both continuous feasibility and a smaller frozen-policy residual (§6.8). Thus the old pooled non-detection and the old "entirely achievability" attribution are each support-specific, not portable conclusions. The shape is reproduced by a second learner/objective bundle (curve correlation 0.976), but that comparison does not isolate architecture from the switch to a team reward. Neutral preferences is not the worst case. No signed design supports an opposition advantage or the divergence D5 predicts; the feasible-centred control instead detects an opposition penalty. More

positive target correlation in a shared pool yields both the most attainable objective and, conditionally on in-box targets, the smallest residual to the matched dynamic controller.

## 7 Discussion

### 7.1 What the control battery supports locally

The paper is organized around a single discipline: do not report an apparent coordination finding until targeted controls have probed whether it is an artifact of the experimental setup. Applying that discipline turned a tidy positive result into a narrower, explicitly conditional one—four claims entered the battery and none left intact, three as outright artifacts and the fourth with its effect intact but its magnitude and attribution made conditional (§6.9). What is worth drawing out here is not the list again but what the fourth case implies. The signed target-correlation gradient is detected; the error lived in pairing out-of-box ideal points and a late-training outcome with a static reference. Frozen held-out evaluation, an in-box target support and an exact transition-matched oracle were all needed to locate the reported channels.

For the consent-friction functional this is a mixed verdict rather than a clean vindication or refutation. Its directional alignment content receives only proxy-specific support: the coordination gap decreases as target-coordinate correlation becomes positive, which is consistent with a monotone $1/(1+\alpha)$-style denominator in this environment but is not a direct measure of objective or reward-function alignment. The dynamic decomposition (§6.8) shows a feasibility channel under both target supports and a residual frozen-policy channel only when the signed targets are centred inside the box. This supports the denominator's direction within this environment, not a support-invariant learning law. Beyond this, three of its commitments must be weakened. The predicted *divergence toward opposition* ($\alpha \to -1$, desideratum D5) is not observed: at latent-target $\rho = -1$ the gap is finite (5.31 versus 4.04 at baseline for IQL) rather than divergent (§6.7). The single-ratio *multiplicative combination* over-specifies how the factors interact: it retains degree-one reward scaling but imposes a one-dimensional coupling among scale, target correlation and noise that the separately estimated effects do not follow. And the target-correlation effect carries an unstated *scope condition*—it is a property of shared-state contention, not of objective agreement in the abstract. The honest summary is neither "the equation was wrong" nor "the test was broken": positive target correlation lowers the gap within shared contention; reward scale enters mechanically (degree-1, by construction, not discovered); and observation noise raises the mean frozen gap under feasible support, but with a positive $\rho \times \varepsilon$ interaction—the reverse of the canonical ratio's implied cross-effect. The specific functional form, its adversarial divergence, and its single-ratio multiplicative combination are not what a controlled test supports; and the target-correlation gradient combines achievability with a support-dependent frozen-policy component (§6.8).

**What the case study feeds back to the formal framework.** Stated as instructions, the verdict is fivefold. (i) *Retain only the proxy-specific direction, not the denominator*: the training-window target-correlation $\rho$ slopes are negative ($-0.755$ IQL, $-0.838$ VDN), the held-out slopes remain negative under both target supports, and M4 retains a $\rho$ term; the composite's $R^2$ rise is not itself evidence for the ratio because the reward-scale-only model fits better (§5.3). The exact decomposition separates a common feasibility channel from a support-dependent frozen-policy residual (§6.8). (ii) *Drop the symmetric refinement* $1/(1+\alpha^2)$: the even-in-$\alpha$ shape it was built to capture was the artifact. (iii) *Weaken the divergence-toward-opposition claim* (D5): no blow-up as $\alpha \to -1$ is observed in any signed design. (iv) *Attach a shared-state interaction scope condition*: the post-control target-correlation estimate is

bounded to variants with jointly controlled coordinates, which cannot simultaneously satisfy divergent ideal points, rather than objective alignment in the abstract. The overlap series bundles physical contention with transition and credit-assignment changes. (The environment is not subtractable and we do not claim it is; see §6.5.) (v) *Treat the single-ratio multiplicative combination cautiously*: although it retains degree-1 reward scaling, its imposed coupling of scale, target correlation and noise is not supported (§5.6); on the signed data the independent-effects model still beats the composite by $\Delta$AICc $\approx$ 18–22, and now non-circularly (§5.3). The signed functional-form refit itself remains an $\varepsilon = 0$ comparison, so M3 is degenerate there. The reduced signed-noise crossing below separately tests the numerator and its interaction; it supports a positive mean noise penalty but contradicts the ratio's qualitative cross-effect. Stake-homogeneity (D1) remains imposed by the reward algebra rather than empirically confirmed. One clarification about the implemented information channel is worth stating explicitly: our observation noise is drawn independently per agent (Definition 3.3), so $\varepsilon$ does not merely blur a shared picture of the state—it makes the agents *disagree* about it. This is an agent-specific observation-noise proxy, not a direct manipulation or estimate of entropy; the experiment cannot transport the result to every information-asymmetry measure.

**Observation noise: positive mean penalty, reversed signed interaction.** The post-fix first-stage factorial finds a small clean-evaluation penalty: its $\varepsilon = 0$ to 1 reward contrast is 0.305, and an independent-sample TOST bounds it to 21.5% of that design's reward-scale anchor. That bound is design-specific. We therefore crossed $\varepsilon \in \{0,1\}$ with $\rho \in \{-.33, 0, .8\}$ and $\sigma \in \{.6, 1\}$ under feasible-centred support, pairing latent problems and initializations across noise levels, freezing each policy, and scoring 100 held-out resets. Table 8 reports the clean-policy estimand. Training with observation noise raises the normalized frozen gap on average by 2.232 IQL and 0.714 VDN. More important for the proposed composition, the $\rho \times \varepsilon$ slope is *positive*: $+1.559$ IQL and $+0.962$ VDN. The canonical ratio implies the opposite qualitative interaction because $\partial[\sigma(1+\varepsilon)/(1+\rho)]/\partial\varepsilon$ falls as $\rho$ rises. Thus the noise penalty is positive on average, but the multiplicative target-correlation–noise claim does not.

The paired-noise model is equally explicit. After a one-to-one merge on the block key

$$b = (\text{mode}, \text{support}, \rho, \sigma, \text{replication})$$

and a latent-problem-hash check, it sets $D_b = g_b^{\varepsilon=1} - g_b^{\varepsilon=0}$. The pooled interaction uses

$$X = [\mathbf{1}, \rho, \mathbf{1}\{\sigma = 1\}],$$

with HC3 covariance and reads the $\rho$ coefficient; reward-scale-specific fits use $X = [\mathbf{1}, \rho]$. A separate model with

$$X = [\mathbf{1}, \rho, \mathbf{1}\{\sigma = 1\}, \rho\mathbf{1}\{\sigma = 1\}]$$

tests the difference between reward-scale-specific slopes.

The average does not license a universal cell-level claim. After Holm correction over all 12 $\rho \times \sigma \times$ learner cells, neither VDN opposition cell is detected and IQL opposition is detected only at $\sigma = 1$; the $\rho = 0$ and $\rho = .8$ cells carry most of the mean penalty. Within-$\sigma$ interaction slopes are positive and detected for both learners at $\sigma = .6$ and for VDN at $\sigma = 1$, while IQL at $\sigma = 1$ is not; the between-reward-scale slope differences are not detected ($p = .275/.324$), so this is not evidence of reward-scale moderation. Matched-noise evaluation raises the average penalties further (2.792 IQL; 1.818 VDN), while restoring the 1% action- exploration floor changes them little (2.722; 1.752). These protocol

Table 8: Reduced feasible-centred signed-noise control on the clean frozen normalized gap (30 paired replications per $\rho \times \sigma$ cell; $n = 180$ pairs per learner). Overall is the equal-weight average over the declared three-$\rho$, two-reward-scale mixture. Interaction uncertainty is HC3; Holm adjustment covers the two learner interactions.

| | Mean noise penalty | | | $\rho \times \varepsilon$ | | |
|---|---|---|---|---|---|---|
| Learner | estimate | 95% CI | $p_{\text{Holm}}$ | estimate | 95% CI | $p_{\text{Holm}}$ |
| IQL | 2.232 | $[1.635, 2.828]$ | $1.1 \times 10^{-11}$ | 1.559 | $[0.329, 2.789]$ | .013 |
| VDN | 0.714 | $[0.505, 0.923]$ | $2.1 \times 10^{-10}$ | 0.962 | $[0.574, 1.350]$ | $2.4 \times 10^{-6}$ |

contrasts keep learning, measurement noise and exploration distinct.

The other result—opposition versus baseline (§6.2, §6.7)—is not treated as equivalence: legacy contrasts are mixed, intervals are wide, and no equivalence bound was prespecified. Relatedly, on the first-stage re-run, the composite's single-index claim can be tested without fitting anything at all: $F = \sigma(1+\varepsilon)/(1+\alpha)$ asserts that the outcome is *some* monotone function of $F$, so any two factorial cells with equal $F$ must behave alike. They do not. At the six-way tie $F = 1.0$ the cells differ sharply (one-way $p \approx 4 \times 10^{-45}$ on the training-gap DV, $1 \times 10^{-32}$ on the clean eval), and across the factorial no monotone $f(F)$ can account for more than $R^2 = 0.55$ of the between-cell variance on either DV (0.548 training, 0.533 clean eval).

That version of the test, however, compares cells across the first-stage alignment axis—the axis §5.3 disowns, since its realized correlation is even in $|\alpha|$ and a monotone $1/(1+\alpha)$ denominator was guaranteed to misfit it. Rejecting the composite there would be circular, so we re-ran the test *within* each $\alpha$ level, where $F$ is monotone in $\sigma(1+\varepsilon)$ and the broken axis is never crossed. The conclusion survives on that narrower ground: of 25 within-$\alpha$ tie sets, 14 have raw $p < 0.05$, 10 survive Holm family-wise correction, and 12 survive Benjamini–Hochberg false-discovery-rate control (smallest raw $p = 9.2 \times 10^{-10}$, largest $\eta^2 = 0.48$; `results/iso_f_fan/within_alpha.py`). We report the within-$\alpha$ version as the result of record and the full fan as an illustration, since only the former is independent of the design flaw the paper is about. This is the model-free counterpart of §5.3's comparative maintained-model evidence, and it does not depend on which functional form one fits. Both computations are in the release-candidate code (`results/equivalence/`, `results/iso_f_fan/`).

### 7.2 The shared-state-bounded effect and the price of anarchy

The price-of-anarchy literature bounds the equilibrium cost of selfish play for fixed game classes (Koutsoupias and Papadimitriou, 1999; Roughgarden, 2005; Nisan et al., 2007). Our reward gap is at best a loose cousin of that cost: $\Delta R$ is an absolute, $\sigma$-scaled gap rather than a welfare fraction (with $R^* = 0$ the natural normalizer is degenerate), and the exact decomposition of §6.8 separates continuous feasibility—incurred even by a perfectly coordinated planner—from reachable-lattice cost, transient cost and the frozen-policy residual. We therefore draw the PoA connection as motivation, not as a mapping. PoA is an equilibrium concept while our measure is a *learning* outcome on a parameterized family of games. The separable-resource control sharpens what the signed sweep is really measuring. It is not that "coordination loss is governed by the signed alignment of objectives" in general; it is that, *when agents must share one resource state*, positively correlated targets give learners a shared structure to lock onto and a more nearly jointly feasible optimum, while opposed targets create contention for that one state with no offsetting structure. Remove the shared state—give each agent its own pool—and the IQL expectation is structurally $\rho$-invariant while the VDN slope is not detected despite team-reward cou-

pling; equivalence is not established (§6.5). The partial-sharing series provides the nontrivial control by preserving interaction while varying the contended fraction. The operative factor is thus the interaction of signed target correlation *with* shared-state contention—rivalry in control of one state, which is the feature our environment shares with the appropriation setting of Pérolat et al. (2017) without sharing its depletion dynamics—not target correlation alone; and because the effect is reproduced by the VDN/team-reward bundle where it is present (§6.4; curve correlation 0.976), it is not confined to the original IQL/individual-reward bundle. Because learner architecture and reward objective change together, this comparison cannot isolate centralization or independent-learner non-stationarity. Two qualifications bound the lesson. First, the exact decomposition (§6.8) shows that more positive target correlation improves the continuous optimum under both target supports, while the frozen-policy residual improves only under feasible-centred targets; the latter is not a support-invariant learning law. Second, the lesson is bounded to the tested shared-state variants; the overlap series changes a transition bundle rather than physical contention alone, and analogues under reward or communication coupling remain open.

### 7.3 Implications for Cooperative AI

The Cooperative AI agenda asks how agents with partially conflicting objectives can nonetheless reach mutually beneficial outcomes (Dafoe et al., 2020, 2021). One possible lever is greater objective alignment, and our surviving effect bears on that lever with three qualifications. First, only ideal-point correlation is manipulated. More positive target correlation improves the coordination gap, and no signed test finds that the opposed endpoint helps. Under legacy target support opposition is mixed across reward scales and learners; under feasible-centred support the four-agent $\rho = -.33$ endpoint is detectably worse than 0. We flag this as a correction to our own first-stage reading, which the sign-blind design had led toward the opposite, seductive conclusion that "baseline, not conflict, is the worst case"—a conclusion that does not survive a genuinely signed test. Second, the support is *scope-conditioned*: the target-correlation dividend is detected only in the tested variants where agents share a contested resource. The data reward higher target correlation *under shared-state contention*; they do not test interventions on arbitrary objectives. Where agents act on separable resources, the target-correlation slope is structurally flat for IQL and not detected for VDN. Third, the mechanism is not singular. The exact dynamic decomposition (§6.8) shows that target agreement makes common ground more attainable under both supports and, with in-box targets, also reduces the learned policies' residual to a matched controller. The cautionary half of the lesson is for experimenters on the agenda: a target sweep that couples agents through a shared latent can silently fail to test opposition at all, and a shared-pool environment can make a target-correlation effect appear context-free—both worth checking before generalizing to objective alignment.

### 7.4 The methodological cautionary tale

The instructive part of this paper is how a symmetric U-shape can be manufactured by a design choice that looks innocuous. Coupling agents' preferences through a *shared latent* ($\tau_i = \alpha b + (1-|\alpha|)\xi_i$) feels like a natural "alignment knob," and its sign $\alpha$ reads like a cooperative-versus-adversarial dial. But the realized cross-agent correlation it induces, $\alpha^2/(\alpha^2 + (1-|\alpha|)^2)$, is even in $\alpha$ and never negative: the dial only turns the *magnitude* of an always-positive correlation. Any outcome measured against it is then forced to be symmetric in $\alpha$, and—because the only unstructured cell is $\alpha = 0$—will generically display a U with neutral at the bottom. The trap is that nothing in the *analysis* is wrong; the bug is upstream, in the data-generating process, and is invisible unless one computes the realized correlation. (Nor is the sign the whole of it: the same coupling silently shrinks the between-agent target dispersion,

$2(1-|\alpha|)^2$, as $|\alpha|$ grows—a second, entangled driver of the U that a sign-only reading of the bug would miss; §6.1.) We were ourselves misled, to the point of proposing a symmetric quadratic refinement $F^{(2)} = \sigma(1+\varepsilon)/(1+\alpha^2)$ to "match" the trough. We now *withdraw* that refinement: the even-in-$\alpha$ shape it was built to capture was the artifact, and the genuinely signed data restore a signed, asymmetric ordering that the original $1/(1+\alpha)$-style denominator already encodes. We are careful not to over-generalize from a single case. We do *not* claim that shared-latent coupling is a pervasive trap throughout mixed-motive MARL—we have not surveyed the literature for it, and many designs realize signed correlation correctly by construction. Social Value Orientation is the canonical example of getting it right: SVO parameterizes an agent's preference as a *signed angle* over its own and others' payoffs (McKee et al., 2020), so cooperative and competitive orientations are genuinely distinct and the magnitude-only trap cannot arise by construction. Our shared-anchor sampler, by contrast, encodes only a magnitude; naming SVO as the design that gets it right makes the trap precise—it is specific to designs that couple preferences through an *unsigned* shared latent, not to signed-orientation parameterizations. What we can say is local and concrete: in *our* setup the bug was real, invisible in the analysis, and caught only by computing the realized correlation; and the fix—verify that the realized cross-agent correlation actually spans the intended sign before reading any U-shape as evidence about opposition—is cheap. We offer it as a check worth running, not as a diagnosis of others' experiments. The same modest framing applies to the other three design choices whose artifacts the battery exposes: degree-1 reward scaling making the reward coefficient make stakes look dominant, a shared-pool environment making contention look like a general target-correlation effect, and an unconstrained per-agent reference point making achievability look like learning. Each was an artifact *of this setup* that a targeted control exposed; whether analogous artifacts lurk elsewhere is for those experiments to check, not for us to assert. The discipline these checks implement is the one Patterson et al. (2024) codify for RL experiments generally, and the concern is the environment-side sibling of the analysis-side sensitivities documented by Henderson et al. (2018). The first three artifacts sit upstream of analysis in the data-generating process and environment; the fourth is a measurement choice that requires changing the reference point in a re-analysis of the same runs.

**A calibration datum: the check's first false positive was ours.** The sign-verification check has been run once without reference to the design's stated estimand—by us, on our own experiment—and it produced a false positive. On 30 July 2026 we ran its first half—compute the realized coupling—against our own two-team block control and flagged the design as an additional sign-blind artifact: the block sampler's mean pairwise correlation is $-\text{opp}^2/3$, negative for every opp $\neq 0$ and symmetric in the sign of opp, so it too fails to track its nominal label. We retracted the finding the same day (the retraction is in the repository history, commit `4dd7750`). The arithmetic was correct—§6.1 states $-\text{opp}^2/3$ itself—but the inference was not, because the team design never claims to sweep a signed $\rho$: opp $\in [0,1]$ is a strength knob and the paper reads it as one. There was no additional artifact. This is the only calibration evidence the battery has, and it points squarely at the failure mode the two-part statement in §6.1 is meant to remove—an auditor who computes realized coupling without locating the design's own stated estimand will flag correct designs. We report it because a paper about auditing one's own experiments ought to report the audit's own false positives, and because it bounds how much weight a reader should place on the check run that way—which is how an outside auditor, with no access to what a design's authors meant, would have to run it.

### 7.5 Stability of outcomes versus strategies

That 99.2% of runs reach reward stability while only 0.88% reach policy stability is *consistent with* the framework's claim that stable coordination need not mean consensus. We read it as suggestive only. Appendix A.1 sets out why: the measurement is an arg max over a fixed probe set, sampled every 200 episodes and averaged across agents, and it cannot separate genuine strategic adaptation from slow drift in a largely settled policy. So we claim only the weaker reading—stability observed at the level of *outcomes*, not established at the level of *strategies*.

The point worth carrying forward is not about our environment. A familiar MARL pathology is that independent learners fail to converge in policy space, and that failure is usually reported as fatal. Our data are at least compatible with a different reading: that policy-space non-convergence and stable collective behavior can coexist, and that which one a study reports depends on the level at which it chose to measure stability. That is a claim about measurement choice, which is this paper's subject, and it does not require us to settle what the underlying policy motion actually is.

## 8 Limitations and Future Work

**Three implementation defects found in our own code, and what they did.** A paper about manufactured findings should say what its own instrument got wrong. Two defects in the factorial runner were found after the first-stage analysis and before this version. (i) Observation noise was scaled by $\varepsilon$ rather than $\sqrt{\varepsilon}$, so the realized covariance was $\varepsilon^2 I$ where Definition 3.3 specifies $\varepsilon I$; the nominal $\varepsilon$ grid therefore labeled a different parameter-to-variance mapping than the one stated. (ii) The replay buffer stored a *clean* next observation alongside a noisy current observation, so temporal difference targets bootstrapped on state information the acting policy never saw—a privilege not present in the stated decision process. Both are fixed at HEAD, and the entire $5 \times 5 \times 5$ factorial was re-run. Both (i) and (ii) affected the results of record, which is why that re-run was necessary; they were then found to be present in the CPU cross-validation runner as well, where they had never been propagated. A third defect, in the environment's action decoder, affected *only* the CPU cross-validation arm—the GPU path of record does not go through that decoder—and is disclosed in §5.8. All three are repaired and the CPU arm has been re-run in full. Every factorial quantity reported in this paper—the main-effect and interaction ANOVAs, the model comparison, and the effect-size partitions—is regenerated from that re-run by `results/gpu_factorial_postfix/regenerate_tables.py`, which is the single source for all of them; an earlier version of this section claimed the same thing while three of those tables were in fact still fit on the pre-fix run, which is precisely the defect this paper is about and is why they now come from one script rather than from hand-transcription. The fixes demonstrably bit, and in the expected place: across the 125 matched condition cells the mean absolute change is 0.0174 at $\varepsilon = 0$ against 0.0329 at $\varepsilon > 0$ (Mann–Whitney $p = 0.006$; `fix_impact.py`), so the cells that should have moved moved more. The comparison is condition-level because the pre-fix run survives only as condition means. What they did *not* do is change any conclusion—the condition-level partition shifts by under 0.006 on every term (Table 10). We report this as the outcome it is, a robustness check we did not choose to run. The legacy signed and contention arms are untouched because they run at $\varepsilon = 0$; the reduced signed-noise follow-up uses the repaired observation path and is reported separately above.

Several limitations qualify and bound these findings. First—and most important for the result of record—the post-control target-correlation estimate is detected *only* with shared-state contention in these variants. At the fully separable endpoint (§6.5), $\rho$-invariance is built into the IQL arm, whose agents share neither state nor reward and have $\rho$-invariant marginal targets. VDN retains team-reward cou-

pling during training, so its non-detected slope is an empirical check rather than equivalence, but the environment's optimal control problem remains additive and separable. The endpoint is not sufficient mechanism evidence. The paired legacy-Gaussian-at-zero-support partial-sharing control is the informative test because it preserves interaction while varying the fraction of jointly controlled coordinates; its HC3 $\rho \times w$ term detects graded slope modulation (§6.6). We therefore bound the empirical claim to shared-state contention where agents' $\pm 1$ actions enter a common state (Definition 3.1), while leaving other forms of coupling untested.

That interpolation has an additional ordered-coordinate limitation. At each $k$, the runner shares the first $k$ coordinate positions rather than counterbalancing or randomizing subsets of the same size. The four-level result therefore identifies modulation across these fixed first-$k$ configurations; it does not identify the overlap fraction independently of coordinate position. A counterbalanced subset-invariance control would require new training.

Second, the equicorrelated adversarial arm is capped at $\rho = -1/(n-1) = -1/3$ for $n = 4$, the positive-semidefinite bound on an equicorrelation matrix (7). For $n = 4$, stronger opposition is only reachable via the non-equicorrelated team-block design; latent-target $\rho = -1$ is reachable by the equicorrelated sampler only after reducing to $n = 2$. The designs agree (opposition $\leq$ baseline), but the cleanest continuous arm cannot pass $-1/3$ by construction. The bound tightens, not loosens, with population size: $-1/(n-1) \to 0$ as $n$ grows, so mutual opposition becomes *less* reachable the more agents there are, and $\rho = -1$ is available only at $n = 2$. This is why the latent-target opposition control (§6.7) is run at $n = 2$ rather than by scaling the equicorrelated arm up, and it is a structural feature of equicorrelation rather than a limitation of our grid: many agents cannot all be pairwise maximally opposed.

The sampled endpoint is $\rho = -.33$, not the exact singular boundary $-1/3$. $\Sigma(-.33)$ has smallest eigenvalue .01, so the arm is interior but near-boundary. The already-trained feasible-centred cells at $-.20$ and $-.10$ retain the same penalty-signed point estimate relative to $\rho = 0$ in all eight learner-by-reward-scale comparisons; only two survive Holm correction over that sensitivity family. Directional continuity is therefore visible, but the individual interior contrasts are not uniformly detected.

**The environment delivers less opposition than the label claims.** A sharper version of the same limitation applies to the whole signed axis, and it extends the realized-correlation control one step further than §6 takes it. That control verifies that the sampler realizes the nominal correlation of the *targets*. But an agent's reward is a weighted quadratic loss on a state confined to $\mathscr{S} = [0, C]^m$, so the state it actually prefers is not $\tau_i$ but $\text{clip}(\tau_i, 0, C)$—and since targets are marginally $\mathscr{N}(0,1)$ with $C = 10$, the lower bound binds about half the time while the upper never does. Measuring correlation on the feasible optima instead of the latent targets (`results/adversarial_rerun/feasible_correlation.py`):

| nominal $\rho$ | $-0.33$ | $-0.20$ | $+0.20$ | $+0.40$ | $+0.60$ | $+0.80$ |
|---|---|---|---|---|---|---|
| latent corr | $-0.330$ | $-0.201$ | $+0.200$ | $+0.399$ | $+0.602$ | $+0.797$ |
| feasible corr | $-0.217$ | $-0.138$ | $+0.155$ | $+0.335$ | $+0.530$ | $+0.743$ |
| retained | 0.66 | 0.68 | 0.78 | 0.84 | 0.88 | 0.93 |

The box constraint attenuates the axis asymmetrically: the cooperative end is delivered nearly intact (0.93 at $\rho = +.8$), while the opposed end loses a third (0.66 at $\rho = -.33$). Monotonicity survives but linearity in the nominal label does not, and the legacy opposition non-detection was obtained under a weaker feasible treatment than its latent label advertised. We therefore ran the missing control rather than

leaving it as future work: translate the same signed Gaussian to the centre of $[0, C]^m$, keep its covariance and all learning settings fixed, and evaluate frozen policies on the same held-out reset bank. Every target is then feasible and realized correlations match the nominal grid. The frozen slope becomes three to four times steeper, and opposition is worse than baseline for the four-agent $\rho = -.33$ versus 0 contrast at both reward scales and for both learners after Holm correction (§6.8). The legacy non-detection is thus support-bounded, not evidence that opposition is harmless. We treat target support and benchmark construction as the two pieces of the title's fourth measurement choice, not as an extra manufactured finding. Translation is a support intervention rather than a pure feasibility-only manipulation: it also changes reset-to-target distance and boundary geometry.

Third, the legacy-support training-window effect of correlation is modest in size ($\eta^2 \approx 0.036$–$0.058$ normalized), and per-replication variance at $\sigma = 1$ is large, so several single-pair contrasts are marginal: the cooperation-beats-baseline test is $p = 0.048$ (IQL), and the team opposition-hurts trend rests on one parametric statistic ($p_{\text{trend}} = 0.048$) while the monotonic and pairwise tests are non-detections (§6.3). The $n = 2$ latent-target opposition control inherits the same fragility on the IQL side: its opposition-versus-baseline effect is Welch-significant only when pooled over $\sigma$ ($p = 0.041$), with a rank-test non-detection ($p = 0.24$) and a heavy-tailed opposition distribution. We therefore report that no test detects an opposition advantage; this is neither equivalence nor a consistently detected penalty. Under a Holm correction across the four planned contrasts of Table 2, the VDN cooperation-beats-baseline contrast ($p_{\text{Holm}} = 0.0071$) survives while its IQL counterpart ($p_{\text{Holm}} = 0.144$) does not, so the only corrected endpoint detection is VDN; the negative full-grid slopes are separate exploratory evidence (§6.2) rather than the single IQL contrast. These limitations remain true for that archived estimand. They do not override the separate feasible-centred held-out control, whose within-$\sigma$ four-agent endpoint penalties survive Holm correction for both learners; instead the discrepancy is evidence that the estimand is target-support dependent.

No prospective power or minimum-detectable-effect calculation governed these post-diagnostic controls. Accordingly, the separable VDN slope, departure from a linear partial-sharing ramp, legacy opposition contrasts and $\rho \times \sigma$ interactions remain non-detections rather than evidence of absence or equivalence; the designs were not calibrated to exclude small effects. Reported intervals and achieved sample sizes provide descriptive sensitivity. A post-hoc power calculation would not convert that record into prospective design justification.

Fourth, IQL is a deliberately naive choice—coordination failure is what we measure, not what we optimize away—and the VDN comparison preserves the shape (curve correlation 0.976) while jointly changing the learner architecture and reward objective. It therefore spans one additive value-decomposition/team-reward bundle rather than isolating either ingredient. Auxiliary repository analyses do include MAPPO (Yu et al., 2022) on magnitude and signed grids, but we do not use that arm as headline evidence here: its per-replication advantage normalization can alter reward-scale sensitivity, no normalization ablation isolates that contribution, and the available comparison does not isolate centralization. We did not run the monotonic-mixing QMIX method (Rashid et al., 2018). The signed grid is also moderately coarse (eight $\rho$ levels, two $\sigma$ levels, $\varepsilon = 0$), so finer structure between grid points may be missed. Multiplying every reward by a constant is also not exactly equivalent to changing Adam's learning rate because moment normalization and the optimizer's $\varepsilon$ term break that equivalence. With the optimizer fixed, the study identifies reward-scale sensitivity of this training bundle, not an optimizer-invariant law or an isolated effective-step-size mechanism.

Fifth, every quantitative claim here is bounded to the specific environment family we tested, and should be read with those conditions attached: $n = 4$ agents ($n = 2$ for the latent-target opposition con-

trol), $m = 3$ continuous shared resources on $[0, C]$, a quadratic-loss reward (2), and the signed equicorrelated / team-block / separable preference designs above. These are a single bespoke environment and its separable twin, not a recognized benchmark such as the Multi-Agent Particle environments, the StarCraft Multi-Agent Challenge, or Melting Pot (Leibo et al., 2021). We therefore speak of an empirical *characterization* of this environment family rather than a general law: "more positive target correlation lowers the measured gap under shared-state contention" is the local result here, under continuous shared resources and quadratic loss, and its generality to other reward shapes, discrete or networked resources, larger $n$, and standard benchmarks is open. Finally, the cooperative benchmark $R^* = 0$ (5) is the unconstrained per-agent optimum, strictly above the achievable joint optimum $R^{\text{joint}}$ (6) wherever targets diverge, so multiplicative gap *ratios* inherit a scale dependence and an infeasibility floor; we lead level summaries with standardized and reward-scale-normalized quantities accordingly, and the signed/contention contrasts compare across $\rho$, where that floor varies monotonically with alignment. Because the floor itself is larger toward opposition, a non-detection on the total gap is not conservative evidence about learning; the decomposition is required to separate feasibility from learned shortfall.

Future work should: add the advantage-normalization control needed to bring the auxiliary MAPPO arm into the reported battery, extend to monotonic-mixing CTDE methods and to standard mixed-motive benchmarks; refine the partial-sharing overlap response (§6.6) on a finer overlap grid—a larger $m$ giving finer $k/m$ steps—to resolve curvature that the present four levels cannot distinguish from a linear ramp; extend the opposition control to $n > 2$ teams and to graded between-group structures; refine the $\rho$ grid; and connect the learned, shared-state-bounded coordination gap to analytic price-of-anarchy bounds for this game family.

## 9 Conclusion

This paper's result of record is a control battery. Four apparent coordination findings entered it and none left intact. First, the sampler manufactured a symmetric *U-shape*: coupling agents through a shared latent makes the realized correlation $\alpha^2/(\alpha^2 + (1 - |\alpha|)^2)$, even in $\alpha$ and never negative, so $\alpha = \pm 0.8$ were identical positive-correlation arms and no adversarial arm existed. Second, the reward scale manufactured *stakes-dominance* through degree-1 homogeneity ($\eta^2 \approx 0.23$ raw, $\approx 0.0003$ reward-scale-normalized). Third, the shared-state environment made an effect of contention look like a general effect of target correlation: the fully separable IQL expectation is structurally $\rho$-invariant and the VDN slope is not detected despite team-reward coupling, without establishing equivalence, while an interaction-preserving partial-sharing sweep on paired legacy-Gaussian-at-zero support detects graded $\rho \times w$ modulation. Fourth, target support and the performance reference manufactured an attribution rather than an effect. Legacy targets centred at the state boundary attenuated the signed treatment, and a static single-state optimum plus continued-training outcome did not identify learned-policy quality.

That fourth case is the one we would most want a reader to take away because it required three controls together. Frozen policies on held-out resets remove late updating and residual exploration; translating the same Gaussian into the feasible box delivers the intended treatment; and an exact finite-horizon controller separates continuous feasibility, reachable-lattice cost, travel transient and the policy residual. Under legacy support, continuous feasibility contributes more than the point-estimated frozen slope, while the residual is not detected ($+.136$, $p = .463$, IQL; $+.130$, $p = .051$, VDN); absence or equivalence is not established. Under feasible-centred support, the frozen slope is three to four times steeper, feasibility supplies about 54–63%, and the residual supplies about 37–46%. The methodological takeaway is offered as checks worth running on similar setups, not as a diagnosis of the field: verify realized treatment, normalize algebraic scale, preserve interaction while varying the proposed mechanism,

freeze and hold out evaluation, and match the benchmark to the actual dynamics and target support.

The battery's pre-stated target—the case study that gave it claims fixed in advance—was the consent-friction functional $F = \sigma(1+\varepsilon)/(1+\alpha)$. Its surviving empirical core is narrow, and once correctly scoped and decomposed it is smaller still. In a four-agent shared-state resource game with quadratic loss, more positive target-coordinate correlation lowers both the archived training-window gap and the held-out frozen-policy gap. This experiment does not directly measure objective- or reward-function alignment. The legacy training-window contrasts are non-detections and the agent-count-confounded $n = 2$ arm does not establish an advantage; frozen legacy evaluation is mixed, with a Holm-significant penalty only for IQL at $\sigma = 1$. Under feasible-centred support, the four-agent $\rho = -.33$ endpoint is worse than 0 at both reward scales and for both learners after Holm correction. The result is shared-state-bounded: the fully separable IQL expectation is structurally $\rho$-invariant and the VDN slope is not detected despite team-reward coupling, without establishing equivalence, while the interaction-preserving partial-sharing sweep on paired legacy-Gaussian-at-zero support detects graded slope modulation ($\rho \times w$ problem-block-clustered $p = 3.7 \times 10^{-6}$ IQL, $2.0 \times 10^{-17}$ VDN). In this paired follow-up, the problem-block-clustered categorical test also detects modulation ($p = 3.6 \times 10^{-6}$ IQL; $3.1 \times 10^{-16}$ VDN), while departure from a linear ramp is not detected ($p = .890/.191$); four levels still do not establish exact linearity. The shape is reproduced by the VDN/team-reward bundle (curve correlation 0.976), a comparison that does not isolate architecture from objective. The exact finite-horizon decomposition identifies a feasibility channel under both supports and a sizeable residual frozen-policy channel only under feasible-centred targets. Observation noise also raises the clean frozen gap on average in the reduced feasible-centred crossing, but its penalty grows as target correlation becomes positive for both learners. That positive $\rho \times \varepsilon$ interaction is the reverse of the proposed ratio's qualitative cross-effect, and the cell-wise penalty is not detected across most opposed cells. The first-stage small-effect bound is therefore not transported to this support or mixture. (The first-stage M4-over-M1 comparison, $\Delta\text{AIC} = 153.6$, is entangled with the sign-blind alignment axis and is retained only as part of the failure record. The result of record is the signed-grid refit: M1's $R^2$ is 0.67–0.72, while the degree-one homogeneous M5 reaches $0.959/0.958$ and wins by $\Delta\text{AICc} = 33.45$ IQL and 30.51 VDN. M4 and reward scale alone also beat M1, while flexible polynomials in the canonical index do not rescue it. The pre-stated $R^2 > .7$ rule cannot receive a confirmatory verdict here: the reduced restriction reaches .666 for IQL and .717 for VDN, but the signed grid fixes $\varepsilon = 0$ and is post-diagnostic. Taken together with the signed-noise result, these exploratory comparisons reject the implemented ratio as a maintained single-index predictor in this environment; they do not separately validate its $1/(1+\rho)$ denominator or imply that the smaller signed-grid margin is merely a sample-size rescaling of the first-stage statistic.) For the formal framework the verdict is usable but narrow: retain objective alignment only as a theoretical candidate and treat target correlation as an environment-specific proxy under shared-state contention; drop the symmetric quadratic refinement and the adversarial-divergence prediction, retain only a scope-specific positive mean observation-noise direction, and reject the single ratio's target-correlation–noise interaction in this environment. The broader contribution is a worked example of catching one's own artifacts and reporting the corrections, and their scope, as the finding.

Table 9: First-stage sign-blind MARL factorial: design parameters.

| **Parameter** | **Symbol** | **Levels** | **Interpretation** |
|---|---|---|---|
| Alignment | $\alpha$ | $\{-0.8, -0.4, 0, 0.4, 0.8\}$ | Nominal; realized $\rho$ is even and nonnegative |
| Stakes | $\sigma$ | $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ | Weight magnitude on resources |
| Observation-noise proxy | $\varepsilon$ | $\{0, 0.25, 0.5, 0.75, 1.0\}$ | Gaussian noise variance parameter |
| *Design summary* | | | |
| Conditions | | $5 \times 5 \times 5 = 125$ (full factorial) | |
| Replications | | 30 per condition (3,750 total) | |
| Episodes | | 1,000 per replication | |
| Agents | | $n = 4$ per environment (IQL) | |
| Resources | | $m = 3$ shared, continuous, capacity $C = 10$ | |

Table 10: Three-way factorial ANOVA on mean reward, at two levels of aggregation. *Condition-level* $\eta^2$ (headline, $n = 125$ condition means) is the same aggregation as the $R^2$/AIC model comparison (Table 11), with within-replication noise removed. *Per-replication* $\eta^2$ ($n = 3{,}750$, with $F$, $p$) retains the within-condition replication variance. Because $R^* = 0$ is constant, the reward gap is $-1\times$ mean reward and yields identical $F$ and $\eta^2$ at each level. Computed on the *post-fix* factorial (`results/gpu_factorial_postfix`, regenerated by `main_effects_table.py`), i.e. after the observation-noise scaling and replay next-observation corrections of §8. The pre-fix run gave stakes 0.741, alignment 0.179, observation-noise proxy 0.015: the corrections move the condition-level partition by less than 0.006 on every term, which is why we report the change as a robustness check rather than a correction to any claim.

| | | **Per-replication** ($n = 3{,}750$) | | | **Cond.-level** | |
|---|---|---|---|---|---|---|
| **Source** | **df** | $F$ | $\eta^2$ | $p$ | $\eta^2$ ($n = 125$) | **Direction** |
| Alignment ($\alpha$) | 4 | 185.44*** | 0.0963 | $< 0.001$ | **0.182** | U-shaped: $\alpha = 0$ worst |
| Stakes ($\sigma$) | 4 | 749.66*** | 0.3894 | $< 0.001$ | **0.736** | Higher $\sigma \Rightarrow$ more friction |
| Observation noise ($\varepsilon$) | 4 | 15.76*** | 0.0082 | $< 0.001$ | **0.015** | Higher $\varepsilon \Rightarrow$ more friction |
| *Interaction effects* | | | | | | |
| $\alpha \times \sigma$ | 16 | 10.34*** | 0.0215 | $< 0.001$ | 0.041 | |
| $\alpha \times \varepsilon$ | 16 | 1.15 | 0.0024 | 0.302 | 0.005 | |
| $\sigma \times \varepsilon$ | 16 | 1.83* | 0.0038 | 0.023 | 0.007 | |
| $\alpha \times \sigma \times \varepsilon$ | 64 | 0.94 | 0.0078 | 0.622 | 0.015 | |

Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (per-replication ANOVA). **df**: numerator degrees of freedom; per-replication residual df = 3625. $\eta^2$ is classical ($\mathrm{SS}_{\mathrm{effect}}/\mathrm{SS}_{\mathrm{total}}$), the proportion of total variance explained, at each aggregation level. The condition-level main effects sum to $\eta^2 = 0.933$, the interactions to the remaining 0.067. The per-replication main and two-way effects sum to $\eta^2 \approx 0.52$; the remainder there is within-condition replication variance, which the condition-level aggregation removes. On the reward-scale-normalized gap the condition-level ranking *inverts* (alignment 0.82, observation noise 0.064, stakes 0.004; §5.6), which is why we treat neither ranking as a structural fact.

Table 11: Model comparison on the *first-stage sign-blind* factorial: competing friction specifications. DV: condition-level reward gap ($n = 125$). The $|\alpha|$-symmetric sampler makes this comparison circular as evidence about signed target correlation; Table 12 repairs that validity problem by re-estimating the forms on the signed target-correlation design. Computed on the *post-fix* run (`results/gpu_factorial_postfix`), the same source as Table 10, which is what makes the two comparable; the pre-fix run gave $\Delta\text{AIC} = 156.6$.

| **Model** | $R^2$ | **AIC** | **BIC** | **RMSE** |
|---|---|---|---|---|
| M1: Friction $\sigma(1+\varepsilon)/(1+\alpha)$ | 0.1098 | 327.3 | 332.9 | 0.8818 |
| M2: Additive $\sigma+\varepsilon-\alpha$ | 0.1581 | 320.3 | 326.0 | 0.8575 |
| M3: Multiplicative $\sigma\cdot\varepsilon\cdot(1-\alpha)$ | 0.1863 | 316.1 | 321.7 | 0.8431 |
| M4: Independent $\alpha+\sigma+\varepsilon$ | **0.7478** | **173.7** | 185.0 | **0.4694** |

M1–M3 are single-predictor models (2 parameters each); M4 uses three independent predictors (4 parameters). Bold: best on that criterion. $\Delta\text{AIC}$ (M4 vs. M1) = 153.6.

Table 12: Functional-form re-estimation on the *signed* equicorrelation data (§6; DV: condition-level reward gap, 16 $(\rho,\sigma)$ conditions per learner, $\varepsilon = 0$). M3 $= \sigma\varepsilon(1-\rho)$ is identically zero and not identifiable. Because $n = 16$ we report strict small-sample AICc (Burnham and Anderson, 2002). M5 enforces degree-one reward-scale homogeneity without the reciprocal denominator; M6 adds a free intercept; M7/M8 allow quadratic/cubic response to the canonical index. The signed and first-stage information-criterion differences use different data and are not compared on a rescaled common metric. M1's $R^2$ rises from 0.110 to 0.67–0.72. M0 ($\sigma$ alone, discarding $\rho$) is the atheoretical baseline for the single-index class, carried at the same $K = 3$ as M1 and M2; it beats the canonical composite on both learners.

| **Learner** | **Model** | $R^2$ | **AICc** | **$\Delta$AICc vs. M5** |
|---|---|---|---|---|
| IQL | M0: Reward scale $\sigma$ | 0.842 | 17.9 | 21.5 |
| | M1: Friction $\sigma/(1+\rho)$ | 0.666 | 29.8 | 33.4 |
| | M4: Independent $\rho+\sigma$ | 0.935 | 7.4 | 11.0 |
| | M5: $\sigma(\beta_0+\beta_\rho\rho)$ | **0.959** | **−3.6** | — |
| | M6: intercept $+\sigma+\sigma\rho$ | 0.959 | −0.1 | 3.5 |
| | M7: $F+F^2$ | 0.667 | 33.4 | 37.0 |
| VDN | M0: Reward scale $\sigma$ | 0.777 | 18.6 | 26.7 |
| | M1: Friction $\sigma/(1+\rho)$ | 0.717 | 22.4 | 30.5 |
| | M4: Independent $\rho+\sigma$ | 0.925 | 4.7 | 12.8 |
| | M5: $\sigma(\beta_0+\beta_\rho\rho)$ | **0.958** | **−8.1** | — |
| | M6: intercept $+\sigma+\sigma\rho$ | 0.959 | −4.8 | 3.3 |
| | M7: $F+F^2$ | 0.720 | 25.9 | 34.0 |

M5 is the zero-intercept homogeneous model; M6 tests that intercept restriction. M8, omitted for compactness, adds $F^3$ and is also worse than M5. M3 is not identifiable at $\varepsilon = 0$. AICc uses the strict Burnham & Anderson parameter count ($K$ = mean parameters $+\,\sigma^2$); the residual-variance convention adds a constant that cancels in every $\Delta$AICc. Bold: best on that criterion.

Table 13: Sensitivity of the signed target-correlation model comparison to unequal cell precision. The equal-condition OLS rows are canonical and are the only rows assessed against the pre-stated H4 $R^2$ threshold. WLS uses inverse estimated variance of each cell mean; replication-row OLS is a descriptive working-independence sensitivity. AICc differences are comparable only within a learner and estimator.

| Learner | Estimator | $R^2_{\mathrm{M1}}$ | $R^2_{\mathrm{M5}}$ | $\Delta\mathrm{AICc}_{\mathrm{M1-M5}}$ |
|---|---|---|---|---|
| IQL | Equal-condition OLS | .666 | .959 | 33.447 |
| IQL | Inverse-variance WLS | .504 | .926 | 30.368 |
| IQL | Replication-row sensitivity | .186 | .268 | 50.791 |
| VDN | Equal-condition OLS | .717 | .958 | 30.512 |
| VDN | Inverse-variance WLS | .545 | .919 | 27.598 |
| VDN | Replication-row sensitivity | .207 | .277 | 44.179 |

Cell replication variances span max/min ratios 6.404 (IQL) and 5.506 (VDN). WLS weights are $1/(s_c^2/n_c)$, required to be finite and positive and held fixed across models. M5 also beats M4 by $10.99/12.81$ AICc units in equal-condition OLS, $7.42/8.54$ under WLS, and $6.42/8.24$ in the replication-row sensitivity (IQL/VDN). The 480 replication rows retain matched problem blocks across learner runs and are not 480 unqualified independent experimental units. The reproducible source is `results/adversarial_rerun/functional_form_refit.py`; its complete mapped output is `functional_form_refit_robustness.csv`.

Table 14: Pairwise and three-way interaction effects.

| **Interaction** | $F$ | $\eta^2$ | $p$ | **df** |
|---|---|---|---|---|
| *DV: mean reward* | | | | |
| $\alpha \times \sigma$ | 10.34*** | 0.0215 | $< 0.001$ | 16 |
| $\alpha \times \varepsilon$ | 1.15 | 0.0024 | 0.302 | 16 |
| $\sigma \times \varepsilon$ | 1.83* | 0.0038 | 0.023 | 16 |
| $\alpha \times \sigma \times \varepsilon$ | 0.94 | 0.0078 | 0.622 | 64 |
| *DV: Gini coefficient (agent inequality)* | | | | |
| $\alpha \times \sigma$ | 2.54*** | 0.0059 | $< 0.001$ | 16 |
| $\alpha \times \varepsilon$ | 3.10*** | 0.0072 | $< 0.001$ | 16 |
| $\sigma \times \varepsilon$ | 0.84 | 0.0020 | 0.639 | 16 |
| $\alpha \times \sigma \times \varepsilon$ | 0.90 | 0.0084 | 0.701 | 64 |

Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 15: Cross-implementation validation: GPU vs. CPU agreement on the reward gap, computed on all 125 matched condition means.

| **Outcome measure** | **Spearman $\rho$** | **Pearson $r$** | **ICC(3,1)** |
|---|---|---|---|
| Reward gap, repaired CPU arm | **0.964** | **0.963** | **0.962** |
| *before the decoder fix* | 0.948 | 0.943 | 0.907 |

Both implementations run the full $5 \times 5 \times 5$ factorial (125 conditions, 30 replications each, 1,000 episodes); correlations are over the 125 matched condition means. GPU: PyTorch-vectorized environments (Paszke et al., 2019) on AMD Radeon RX 7900 XTX (ROCm 6.3). CPU: Python multiprocessing (22 workers), re-run in full after the action-decoder, observation-noise and replay defects were repaired (§5.8). The second row is the same comparison against the defective arm, retained so the effect of the repair is visible: it removes a systematic $+0.499$ level offset, leaving Bland–Altman bias $-0.026$ with 91.2% of conditions inside the limits of agreement, and improves all three agreement statistics. The validation claim therefore rests on level as well as rank. The three secondary proxies that earlier versions cross-validated here are withdrawn (§3.4). Regenerated by `results/full_factorial_postfix/crossval.py`.

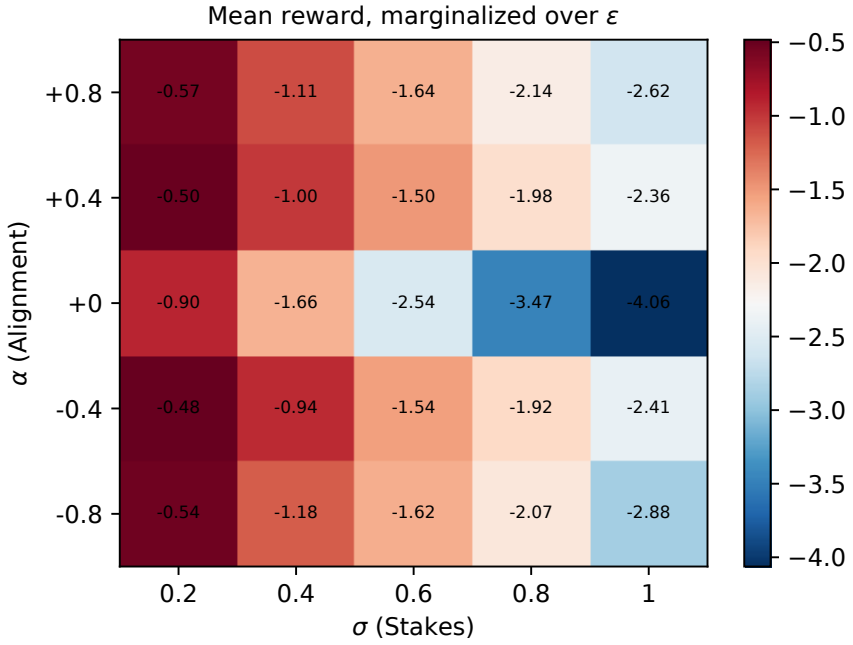


Figure 8: Pairwise interaction heatmaps for mean reward. Left: alignment ($\alpha$) vs. reward scale ($\sigma$), marginalized over observation noise. The U-shaped nominal-$\alpha$ effect is visible: neutral alignment ($\alpha = 0$) produces the highest friction at every stakes level, while both extremes of $\alpha$ (strong shared structure, large $|\alpha|$) reduce it. Right panels: remaining pairwise interactions. This is the first-stage sign-blind factorial: the U-shape is a known-artifactual property of the sampler, not a signed-alignment result.

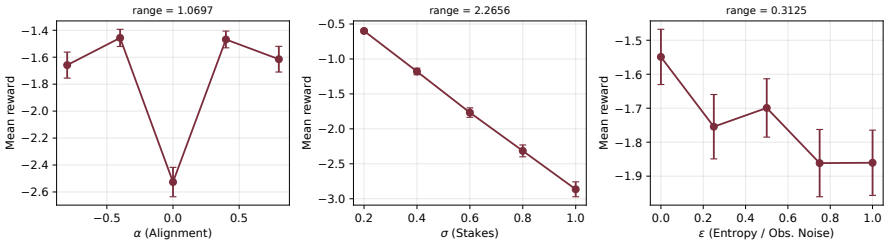


Figure 9: Main effects of alignment ($\alpha$), stakes ($\sigma$), and observation-noise variance parameter ($\varepsilon$) on mean reward (equivalently, on the reward gap up to sign, since $R^* = 0$), with 95% confidence intervals from 30 replications per condition. By condition-level raw effect size ($n = 125$), stakes has the largest effect ($\eta^2 = 0.74$), followed by alignment ($\eta^2 = 0.18$), with the noise proxy weak ($\eta^2 = 0.015$); a reward-scale-normalized partition inverts this ranking (nominal $\alpha$ 0.82, §5.6), so neither is read as a structural fact. The per-replication mean-reward ANOVA ($n = 3{,}750$) shows the same raw ordering at finer grain ($\eta^2 = 0.389, 0.096, 0.008$). Within this sign-blind first-stage design, the alignment profile is clearly U-shaped rather than monotonic; this is descriptive of the known-artifactual sweep, not a signed headline.

## A First-Stage Descriptive Results (Sign-Blind Design)

The two descriptions below are faithful properties of the *first-stage* sign-blind factorial (§5), in which the cross-agent correlation depends only on $|\alpha|$ (4). Both are $|\alpha|$-governed and neither is re-estimated under the signed DGP; we report them as first-stage descriptives, not as claims about signed target correlation.

### A.1 Dynamic equilibria

A sharp divergence separates policy-level and outcome-level convergence. The two convergence fractions we quote are *point values at a single declared pair of thresholds*, and we treat the thresholds as a reporting choice rather than a natural constant. Under a policy-stability criterion—per-agent greedy action-frequency vectors over 128 fixed probe states, with the exploration mixture folded in, sampled every 200 episodes and changing by less than $\delta_\pi = 10^{-3}$ in $L_2$ norm between samples—only **0.88%** of replications (33 of 3,750) converge within the 1,000-episode budget; under a reward-stability criterion (late-training mean reward staying within $\delta_R = 10\%$ of its final value), **99.2%** do. Both are recomputed on the post-fix run by `a1_stability_recompute.py`. Two caveats on the estimand, neither of which we had stated. The criterion is not applied to the joint policy directly: it thresholds the *mean over the four agents* of per-agent censored durations, so a replication can count as converged when as few as one agent's probe sequence stops changing—the same across-agent aggregation for which the policy-variance proxy is withdrawn in §3.4. And "consecutive" means 200 episodes apart, the policy sampling interval. The specific 0.88%/99.2% figures are tied to that single declared pair $(\delta_\pi, \delta_R) = (10^{-3}, 10\%)$; we did not sweep the tolerances, so the exact percentages should be read as conditional on that choice. Because the two fractions sit at opposite extremes ($< 1\%$ versus $> 99\%$), we expect the *qualitative gap*—near-zero policy convergence beside near-universal reward convergence—to survive any nearby threshold choice, but we state that as an expectation rather than a computed robustness result, and we lean only on the qualitative gap, not the exact percentages. The joint policy almost never stabilizes under our policy-stability criterion (33 of 3,750), yet the aggregate reward reaches a stable attractor. We are careful about what this divergence does and does not show. An earlier version of this paper attributed the near-zero policy-stability to the $\varepsilon_{\text{exp}}$-greedy floor of 0.01, reasoning that residual exploration mechanically prevents empirical action frequencies from reaching a fixed point. That reasoning is wrong, and reading the measurement code is what shows it: the stored policy vector is a greedy arg max histogram over 128 *fixed* probe states, blended as $(1-\varepsilon_{\text{exp}})\,\hat{\pi} + \varepsilon_{\text{exp}}/|\mathscr{A}|$. Nothing is sampled. At the floor both coefficients are constant across measurements, so the mixture term cancels in the difference between consecutive vectors, and two measurements with the same greedy mapping are *identical*. The instability is therefore a real property of the learned arg max over the probe set and not an artifact of the exploration floor—which makes the finding stronger than the retreat we had prepared for it, and removes the excuse we had given ourselves for not investigating further. What it still cannot do is distinguish strategically meaningful cycling from drift in a arg max that is measured only every 200 episodes and only on a fixed probe set. The reward-stable-but-policy-unstable pattern is thus *consistent with* the framework's picture that stable coordination outcomes can arise through ongoing mutual adaptation rather than convergence to fixed-point agreement—but it is *not evidence for* genuine strategic cycling: we did not anneal $\varepsilon_{\text{exp}} \to 0$, nor analyze the autocorrelation/trajectory structure of the joint policy, so we make no claim that individual strategies are cycling in any strategically meaningful sense, and we leave the $\varepsilon_{\text{exp}} \to 0$ ablation and a policy-trajectory autocorrelation analysis to future work. We accordingly avoid the strong "dynamic-equilibrium" reading and report only the qualitative descriptive fact (near-zero policy convergence, near-universal reward convergence) without attributing it to emergent cycling. Low-stakes conditions ($\sigma \leq 0.4$) reach reward stability within 100–200 episodes; high-stakes conditions

($\sigma \geq 0.8$) require 300–500. Late-training reward stability is uniform across conditions ($\sim$ 0.04–0.06), with neutral alignment marginally less stable (0.056 vs. 0.043–0.050).

### A.2 Friction and the distribution of reward

Beyond aggregate performance, preference structure shapes the *distribution* of rewards across agents (Table A1, Figure A1). As a transformation-dependent descriptive, the Gini coefficient of absolute agent rewards (condition means ranging 0.061–0.282) peaks at neutral alignment (0.233) and is lowest at the alignment extremes (0.080 at $\alpha = -0.8$, 0.081 at $+0.8$). One caveat on that statistic, since the repository briefly carried two incompatible versions of it: the Gini reported here and in Table 14 is taken over |reward|, which is the definition `regenerate_tables.py` and `make_figures.py` both use. A min-shifted variant used by an earlier analysis script has the *opposite* shape in $\alpha$ and a hard floor of 0.25 at $n = 4$; it is not used anywhere in this paper, and readers comparing against the repository should check which is which. The agent reward *variance* tells the same story more dramatically but is scale-dependent: it peaks at neutral alignment (2.148) and collapses to $\sim 0.08$ at $|\alpha| = 0.8$—a roughly $27\times$ ratio, which we report as a variance ratio rather than a structural constant. At $|\alpha| = 0.8$, none of the 25 conditions crosses the unadjusted $p < 0.05$ best-vs-worst-agent screen, against 8–12% at moderate and neutral alignment. These are descriptive screen rates, not multiplicity-controlled inference; zero crossings do not demonstrate the absence of exploitation.

The pattern is again governed by $|\alpha|$, not by sign. When $|\alpha|$ is large, agents' targets share a common anchor (§3.2), so the achievable outcomes are symmetric across agents whether the anchor is positive or negative. At $\alpha = 0$ the targets are independent draws, which create arbitrary asymmetries—some agents happen to prefer states that are easier to reach under the capacity constraints, gaining a structural advantage with no shared structure to offset it. Agent variance also rises with stakes (from 0.080 at $\sigma = 0.2$ to 1.428 at $\sigma = 1.0$), and the worst dispersion occurs at $(\alpha = 0, \sigma = 1.0)$. Shared preference structure is thus associated with both higher aggregate performance and lower reward dispersion—the two move together rather than trading off. The reward-dispersion pattern here arises under shared-state contention, without appropriation or depletion. That distinguishes it from the common-pool appropriation dynamics documented by Pérolat et al. (2017) and targeted by inequity-averse intrinsic motivation (Hughes et al., 2018); here it is associated purely with preference *structure* (the $|\alpha|$ anchor), with no intrinsic-motivation machinery, and—because it is $|\alpha|$-governed—we report it as a first-stage descriptive of the sign-blind design, not as a claim about signed target correlation.

Table A1: Agent reward inequality by alignment level (first-stage sign-blind design).

| **Alignment** | **Agent reward variance** | **Unadjusted $p < 0.05$ screens** |
|---|---|---|
| $\alpha = -0.8$ | 0.084 | 0% of conditions |
| $\alpha = -0.4$ | 0.495 | 12% |
| $\alpha = 0.0$ | **2.148** | 8% |
| $\alpha = +0.4$ | 0.479 | 12% |
| $\alpha = +0.8$ | 0.078 | 0% |

Neutral alignment exhibits $\approx 27\times$ the agent reward variance of the extremes. The final column is the share of conditions crossing an unadjusted best-vs-worst-agent reward-dispersion screen ($t$-test, $p < 0.05$); it is descriptive and not multiplicity-corrected.

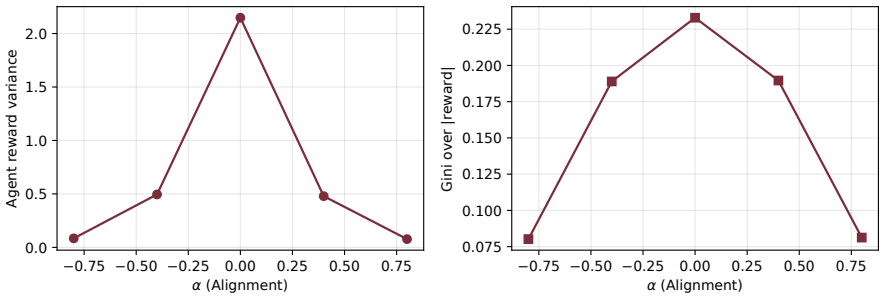


Figure A1: Agent reward inequality as a function of alignment ($\alpha$) on the sign-blind first-stage axis, regenerated from the post-fix factorial by `make_figures.py`. Left: agent reward variance peaks at 2.148 for neutral alignment, dropping to $\sim 0.08$ at the extremes—an $\approx 27\times$ lower variance (a scale-dependent ratio). Right: the Gini over |reward| shows the same inverted-U pattern, peaking at 0.233; it is conditional on that absolute-value transformation. Shared preference structure, at either anchor sign, is associated with lower reward dispersion across agents.

## Declarations

**Conflict of Interest.** The author declares no competing interests.

**Funding.** This research received no external funding.

**Data and Code Availability.** A public repository exists at `https://github.com/dissensus-ai/friction-marl`, and a versioned Zenodo record exists under concept DOI 10.5281/zenodo.20636467 (CC BY 4.0), a concept DOI that resolves to the latest published version. The immutable v3.0.0 record DOI 10.5281/zenodo.22004215 archives the checked August source and raw outputs: post-fix GPU and CPU factorials, frozen-policy evaluations, feasible-target and dynamic-oracle controls, signed observation-noise crossing, paired partial sharing and the earlier February/June battery. Its seven public files match the frozen inventories and `SHA256SUMS`; the record DOI, rather than the mutable repository alone, identifies the replication payload. The 24 August problem-block bootstrap and matched-protocol reporting delta postdates v3; its exact script, derived CSVs and provenance README accompany the arXiv source as ancillary files, while repository/Zenodo synchronization remains a separate release task.

**Reproducibility note.** In the February factorial runners, only the NumPy streams for preferences and observation noise were seeded; the global PyTorch RNG governed network initialization, $\varepsilon_{\text{exp}}$-greedy exploration, and replay sampling. Those February CSVs therefore cannot be regenerated bit for bit. Deterministic PyTorch seeding was added on 28 June 2026 in commit `03594128b14611d01176acd16644cf09202eb1d3`. The 29 July post-fix GPU run and 6 August repaired CPU run were produced after that change, but their schedules differ: the GPU runner uses one deterministic seed stream per condition and batches 30 separately indexed stochastic replications, whereas the CPU runner assigns an explicit deterministic seed to each replication. The replication is the inferential unit in both paths; the distinction here is how each exact realization is regenerated from the archived configuration and code.

**Acknowledgements.** The author thanks Claude (Anthropic; Claude Opus and Claude Fable 5 models) for research and drafting assistance throughout the project (see the AI Assistance statement below), ChatGPT (OpenAI; GPT-5.6 "Sol") for adversarial cross-review of the manuscript and its statistical claims, and the Reviewer3 (reviewer3.com), Thesify, and OpenAIReview (ChicagoHAI) automated review platforms for pre-submission checks.

**AI Assistance.** Claude (Anthropic; Claude Opus and Claude Fable 5 models) was used as a research collaborator for literature synthesis, analysis and simulation code development, LaTeX preparation, and iterative refinement of the manuscript. ChatGPT (OpenAI; GPT-5.6 "Sol") was used for adversarial review and cross-checking of the reported results. Citation verification drew on scite, among other automated research tools. All intellectual claims, statistical results, and errors remain the author's sole responsibility.

## References

Ibrahim Abada and Xavier Lambin. Artificial intelligence: Can seemingly collusive outcomes be avoided? *Management Science*, 69(9):5042–5065, 2023. doi: 10.1287/mnsc.2022.4623.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, Cambridge, MA, 2024.

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. In *International Conference on Learning Representations (ICLR)*, 2018.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edition, 2002.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, 2008. doi: 10.1109/TSMCC.2007.913919.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020. doi: 10.1257/aer.20190623.

M. Cardei, Matthew Landers, and Afsaneh Doryab. Coordination matters: Evaluation of cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2605.06557*, 2026.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI, 2020.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021. doi: 10.1038/d41586-021-01170-0.

Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the StarCraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020. doi: 10.48550/arXiv.2011.09533.

Arnoud V. den Boer, Janusz M. Meylahn, and Maarten Pieter Schinkel. Artificial collusion: Examining supracompetitive pricing by Q-learning algorithms. *Management Science*, 2026. doi: 10.1287/mnsc.2024.08557. Published online 9 June 2026.

Ishan Durugkar, Elad Liebman, and Peter Stone. Balancing individual preferences and shared objectives in multiagent reinforcement learning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2505–2511, 2020. doi: 10.24963/ijcai.2020/347.

Benjamin Ellis, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, and Shimon Whiteson. SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Murad Farzulla. The axiom of consent: Friction dynamics in multi-agent coordination. *arXiv preprint arXiv:2601.06692*, 2026a. doi: 10.48550/arXiv.2601.06692. Authorization rule and candidate friction ansatz; the public arXiv version predates the current local reader edition.

Murad Farzulla. The Replicator-Optimization Mechanism: A Scale-Relative Formalism for Persistence-Conditioned Dynamics with Application to Consent-Based Metaethics. *arXiv preprint arXiv:2601.06363*, 2026b. doi: 10.48550/arXiv.2601.06363. Conditional persistence dynamics, kernel and scale formalism, and formal limits; the public arXiv version predates the current local reader edition.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 122–130, 2018.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. Towards a standardised performance evaluation protocol for cooperative MARL. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi: 10.1126/science.162.3859.1243.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019. doi: 10.1007/s10458-019-09421-1.

Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Mądry. A closer look at deep policy gradients. In *International Conference on Learning Representations (ICLR)*, 2020.

Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 1563 of *Lecture Notes in Computer Science*, pages 404–413. Springer, 1999. doi: 10.1007/3-540-49116-3_38.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 464–473, 2017.

Joel Z. Leibo, Edgar A. Duéñez-Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pages 6187–6199, 2021.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML)*, pages 157–163, 1994. doi: 10.1016/B978-1-55860-335-6.50027-1.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012. doi: 10.1017/S0269888912000057.

Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duéñez-Guzmán, Edward Hughes, and Joel Z. Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 869–877, 2020.

Shree Murthy and Rohan Pandey. Failure modes of deep multi-agent RL in asynchronous pricing: Reproducible triggers, trace diagnostics, and a partial fix. *arXiv preprint arXiv:2606.09884*, 2026.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951. doi: 10.2307/1969529.

Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press, Cambridge, UK, 1990. ISBN 9780521405997. doi: 10.1017/CBO9780511807763.

Christos Papadimitriou. Algorithms, games, and the internet. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 749–753. ACM, 2001. doi: 10.1145/380752.380883.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks)*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25, 2024.

Julien Pérolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

David Radke, Kate Larson, and Tim Brecht. The importance of credo in multiagent learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2243–2252. ACM, 2023.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4295–4304, 2018.

Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 10199–10210, 2020.

Tim Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, Cambridge, MA, 2005.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5887–5896, 2019.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2085–2087, 2018.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLoS ONE*, 12(4):e0172395, 2017. doi: 10.1371/journal.pone.0172395.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning (ICML)*, pages 330–337, 1993.

Callum Rhys Tilbury, Claude Formanek, Louise Beyers, Jonathan P. Shock, and Arnu Pretorius. Coordination failure in cooperative offline MARL. In *ICML 2024 Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET)*, 2024.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698.

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 (NeurIPS), Datasets and Benchmarks Track*, volume 35, pages 24611–24624, 2022.