Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Four Ways to Manufacture a Coordination Finding: A Control Battery for Mixed-Motive MARL

Murad Farzulla

murad@dissensus.ai

Dissensus https://orcid.org/0009-0002-7164-8704

Research Article

Keywords: multi-agent reinforcement learning, construct validity, experimental artifact, coordination, signed target correlation, shared-state contention; reproducibility

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10375465/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Four Ways to Manufacture a Coordination Finding

A Control Battery for Mixed-Motive MARL

Murad Farzulla[1,*]

[1]Dissensus, London, UK

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

31 August 2026

Abstract

Mixed-motive multi-agent reinforcement-learning sweeps can manufacture findings before analysis. In a four-agent shared-state task, four choices make reward scale look structurally dominant, manufacture a symmetric treatment profile, turn state contention into a general correlation effect, and misattribute attainable-objective costs to learning. The treatment is signed correlation among target coordinates (ideal points), not a manipulation of objective- or reward-function alignment. The battery finds that the original sampler never produced negative target correlation and that stake dominance was degree-one reward scaling. Separable IQL is correlation-invariant and the VDN slope was not detected; on legacy-Gaussian support, slope estimates steepened across four fixed first-k shared-coordinate configurations. Frozen held-out evaluation preserved the target-correlation gradient. Translating the distribution inside the feasible box made it three to four times steeper. An exact finite-horizon oracle attributed 54–63% (95% bootstrap intervals 51–72%) of that gradient to feasibility and 37–46% (28–49%) to a support-dependent frozen-policy residual. The full proposed index $F = \sigma(1+\varepsilon)/(1+\alpha)$ has no valid confirmatory verdict: on a post-diagnostic $\varepsilon = 0$ signed grid, its reduced restriction loses to a degree-one scale-homogeneous alternative by AICc, and the observation-noise-variance proxy interacts with signed target correlation in the opposite direction. These post-diagnostic controls are exploratory demonstrations. The battery separates treatment validity, algebraic scale, environmental scope and benchmark validity before judging learned policies.

Keywords: multi-agent reinforcement learning; construct validity; experimental artifact; coordination; signed target correlation; shared-state contention; reproducibility

1 Introduction

An analysis can be statistically careful and still answer the wrong experimental question. The well-known failure modes in reinforcement learning include too few seeds, unstable baselines and reporting choices that reverse comparisons (Henderson et al., 2018; Agarwal et al., 2021). This paper studies an upstream family: the sampler may fail to deliver the named treatment, the reward may build a scale ranking into the outcome, the environment may supply the interaction attributed to preferences, and the benchmark may mix achievability with policy quality. None of those defects is repaired by a smaller p-value. Each requires a control aimed at the experimental construction.

The case arose in a test of the consent-friction proposal (Farzulla, 2026a). Its candidate coordinates are alignment α, stakes σ and information loss ε, combined in the phenomenological ansatz

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

The formula is not a theorem. It supplies a useful falsification target because its directions, interaction and single-index claim are explicit. The original $5 \times 5 \times 5$ factorial appeared to support a striking story: reward scale explained most raw variation; nominal neutrality was uniquely bad; either signed extreme appeared helpful; and the observation-noise proxy was small. The story was tidy because the design had manufactured its geometry.

The target sampler coupled every agent to a shared latent with weight $|\alpha|$. Its realized cross-agent correlation was $\alpha^2/[\alpha^2 + (1-|\alpha|)^2]$: even in α and never negative. The nominally adversarial cells were positive-correlation twins. Reward scale then multiplied quadratic rewards almost homogeneously, making their raw variance share large by construction. Once the sign was repaired, a positive-target-correlation decline survived, but only when agents controlled common state. Finally, the outcome compared continued training against an unconstrained static ideal while many targets lay outside the state space. That comparison mixed feasibility, reachability, transient motion, exploration and policy performance.

Throughout the empirical sections, the delivered treatment is therefore named narrowly: *signed target-coordinate correlation*, equivalently correlation among agents' ideal points. The general theory uses "alignment" for agreement among objectives or reward functions. Because the experiment changes target coordinates inside a fixed quadratic reward family, it does not directly manipulate or measure that broader construct. Any connection from ρ to objective alignment is an environment-specific instantiation, not an identified equivalence.

Table 1 states the paper's spine. The title's four choices are the sampler, reward scale, interaction structure and measurement construction. Separable and partial-sharing environments are complementary scope controls; frozen feasible-centred evaluation and the exact oracle jointly repair the fourth choice. Three initial findings lose their claimed status. The fourth retains an exploratory target-correlation gradient but loses its original magnitude and attribution.

The contribution is methodological first and empirical second. Methodologically, the paper gives

Table 1: Four design choices, the misleading inference each supported, and the control that changes its status. Every inferential reading remains exploratory; the evidence archive is separately versioned.

Choice	Apparent finding	Targeted control	Exploratory post-control reading
Sampler	Nominal alignment forms a symmetric U; neutrality is worst	Verify realized correlation; rerun a signed DGP	The U is a magnitude-only artifact; signed target correlation is asymmetric
Reward scale	Theoretical stakes structurally dominate	Normalize by the degree-one scale factor	Raw dominance is mechanical; the implemented coefficient scales losses by construction
Environment	Target correlation generally governs the gap	Feasible-centred separable endpoint and paired legacy-Gaussian-at-zero-support partial sharing	The gradient is detected only with shared state in these variants; point estimates grade with overlap
Measurement	Positive target correlation mainly improves learning against a static optimum	Freeze policies, hold out resets, centre targets in-box, use a dynamic oracle	Within these controls a gradient remains; feasibility and a support-dependent policy residual both contribute

a portable sequence: verify the treatment actually realized; divide out an algebraic scale before ranking factors; preserve interaction while varying the proposed environmental channel; and match evaluation support and benchmark dynamics to the learned policy. Empirically, it characterizes one bespoke quadratic-loss resource task. More positively correlated target coordinates lower the frozen gap when agents contend over shared coordinates. At full separation the IQL expectation is invariant to correlation and the VDN slope is not detected; on paired legacy-Gaussian-at-zero support, the four overlap-level point estimates steepen monotonically. With feasible-centred targets, more positive target correlation improves both the attainable common objective and the residual to a matched finite-horizon controller. These claims are conditional on this environment family and two value-based learner/objective bundles.

This distinction between a total gap and a policy residual is load-bearing. The total gap is a valid description of how far the realized team outcome lies below the agents' separate ideals. It is not, without decomposition, a measure of failure to learn. Opposed targets may make a high-reward shared state impossible even for an omniscient controller; random resets and bounded actions impose travel costs before any policy choice can be judged. Conversely, a favourable continuous optimum does not show that decentralized learners find it. The reader-facing result therefore reports both objects: how signed target correlation changes the best attainable trajectory and how the frozen policies perform relative to that trajectory.

The paper retains the failed factorial rather than presenting only the repaired experiment. That choice exposes a common asymmetry in reproducibility practice. Seeds, confidence intervals and hardware are often documented after the environment has already fixed what counts as treatment, interaction and success. Here the most consequential mistakes occurred before estimation and survived a full statistical analysis. Showing the seductive result beside its control makes the diagnostic procedure inspectable and records which earlier claims were withdrawn. The technical supplement preserves the complete chain

so compression cannot turn correction into disappearance; its navigation index assigns stable identifiers to every moved evidence class.

The controls were developed after the first-stage defect was found. Except for the original factorial hypotheses, they were neither preregistered nor independently replicated. Holm adjustment controls named multiplicity families; it does not make the work confirmatory. Non-detections are retained rather than promoted to "null findings": legacy opposition estimates are mixed rather than equivalent; the separable IQL expectation is structurally correlation-invariant, whereas the VDN slope is an empirical non-detection; departure from a linear overlap ramp and the legacy frozen-policy residual are not detected; and a noise penalty is not detected in several opposed cells. The accompanying technical supplement preserves the complete first-stage tables (`TS-FIRST`), implementation chronology (`TS-IMPL`), stronger-opposition diagnostics (`TS-OPP`) and conditional tables (`TS-COND`) removed from this reader-facing narrative.

The functional is useful here because it is unusually falsifiable. It names three manipulable coordinates, fixes their directions, proposes a composition rule, and supplies a numerical adequacy threshold. That does not make the functional privileged, and this paper is not an attempt to rescue it after a failed test. Low-dimensional coordination indices are attractive precisely because they compress heterogeneous situations into a common scale; the same compression also makes them vulnerable to an experimental axis whose operationalization has changed meaning. The reusable object is therefore the battery rather than this index's verdict. A different theory may name different coordinates, but it still has to show that its sampler realizes them, its outcome does not algebraically pre-rank them, its environment does not supply the purported mechanism, and its benchmark scores an attainable decision problem.

Testing our own proposed index matters for provenance as well as rhetoric. The first-stage U-shape would have been convenient evidence for a symmetric refinement, and an earlier version in fact introduced one. Once the realized treatment was calculated, that refinement became a model of the sampler artifact and was withdrawn. The correction is therefore part of the scientific result, not a pre-publication detail to be hidden by presenting only the improved experiment. The full record keeps the defective generator, obsolete secondary outcomes, rerun chronology and negative controls; the reader paper keeps enough of that chain to make every surviving claim traceable without making the failure history the main narrative.

The design also distinguishes four questions that are easy to collapse. Does nominal α track realized signed target dependence? Does that dependence predict the total ideal-to-outcome gap? Does it matter because agents act on shared coordinates? And, conditional on what the dynamics permit, does it predict a trained policy's residual to a matched controller? A single downward curve can answer the second question while leaving the other three unresolved. The successive controls are ordered to prevent that curve from being over-read: treatment validity precedes effect estimation, scope precedes mechanism language, and dynamic benchmarking precedes learning attribution.

The remainder proceeds in the order the design should have been audited. Section 2 locates the battery among RL validity work, mixed-motive preference sweeps and contention models. Section 3 defines

the shared, separable and partially shared MDPs, distinguishes late-training from frozen estimands and states the exploratory analysis families. Results I keeps the manufactured findings as a failure case. Results II verifies the signed treatment, tests the interaction structure, repairs target support, decomposes the dynamic benchmark and crosses signed target correlation with observation noise. The discussion returns only to claims those controls can carry.

2 Related Work

Construct validity in RL evaluation. Deep-RL comparisons are sensitive to seeds, implementation and aggregation (Henderson et al., 2018; Agarwal et al., 2021); theory-motivated mechanism explanations may also fail controlled intervention (Ilyas et al., 2020). In MARL, evaluation is especially difficult because an agent's performance depends on co-players (Lanctot et al., 2017; Balduzzi et al., 2018). Standardized evaluation proposals document wide protocol heterogeneity (Gorsane et al., 2022), while SMACv2 is a direct benchmark precedent: open-loop policies exposed that the original benchmark did not require the closed-loop coordination attributed to it (Ellis et al., 2023). Recent work likewise argues that return alone can conceal the coordination mechanism being measured (Cardei et al., 2026), within a wider construct-validity account of machine-learning benchmarks (Freiesleben and Zezulka, 2025). Our niche is the experiment-design side of the same problem: the treatment generator, reward algebra, state coupling and benchmark jointly determine what a mixed-motive sweep can identify.

Control batteries and algorithmic collusion. The closest empirical genre is the audit of algorithmic pricing. After independent Q-learners appeared to sustain supracompetitive prices (Calvano et al., 2020), subsequent work tested whether exploration, synchronized updating, identical algorithms and long horizons manufactured the interpretation (Abada and Lambin, 2023; den Boer et al., 2026; Murthy and Pandey, 2026). Those controls hold the economic environment largely fixed and vary learning conditions. Ours holds the learner family comparatively stable while changing the target-coordinate DGP, physical state coupling, target support and reference point. The two batteries are complementary: a clean learner comparison cannot repair a sign-blind treatment, and a correct treatment does not cure exploration-induced collusion.

This literature also clarifies why an audit should preserve the failed analysis. Reproduction of a headline curve is not mechanism validation when alternative design features generate the same curve. The relevant standard is intervention on the proposed channel: alter exploration if the claim concerns punishment strategies, alter state coupling if it concerns contention, and alter target support if it concerns learned coordination rather than feasibility. Our controls are organized by that intervention logic rather than by a generic robustness checklist.

Mixed motives and the preference axis. Sequential social dilemmas make cooperation an emergent consequence of incentives (Leibo et al., 2017). Scalar objective couplings, individual-versus-shared reward mixtures and social-value orientations show that preference alignment changes learned behaviour (Tampuu et al., 2017; Durugkar et al., 2020; McKee et al., 2020). Social Value Orientation is important

here because its signed angle distinguishes cooperation from competition by construction; our original shared-anchor magnitude did not. Results that full common interest is not always optimal (Radke et al., 2023) also warn against assuming a universal monotone benefit. Our evidence is narrower: in this shared-state quadratic task, the gap declines as target correlation becomes more positive, and that decline disappears when the state is fully factorized.

The sign issue separates our failure from ordinary attenuation. A noisy signed variable may weaken an estimated effect while retaining the intended ordering. The shared-anchor construction instead maps both $+\alpha$ and $-\alpha$ to the same positive dependence, so the named opposition arm is absent. This is closer to a treatment-validity failure than measurement error. It is also readily preventable: a pre-run table of realized pairwise correlations would have rejected the design before any learner was trained.

Independent learning and value decomposition. Independent Q-learning is transparent but treats adapting co-players as a non-stationary environment (Tan, 1993; Buşoniu et al., 2008; Hernandez-Leal et al., 2017). Its weaknesses are well documented (Matignon et al., 2012; Hernandez-Leal et al., 2019), though independent policy-gradient methods remain competitive on some cooperative benchmarks (de Witt et al., 2020; Papoudakis et al., 2021). Centralized training with decentralized execution addresses part of the problem through VDN, QMIX and related factorizations (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Rashid et al., 2020). Our VDN comparison changes both architecture and training reward: IQL uses individual rewards, whereas VDN uses their sum. It can test whether the shape recurs in a second learner/objective bundle, not whether centralization causes the change.

Opponent-aware methods can exploit co-player learning dynamics directly (Foerster et al., 2018), while actor–critic CTDE methods such as MADDPG and MAPPO change both optimization and information structure (Lowe et al., 2017; Yu et al., 2022). Those methods are relevant external-validity tests, but adding them without crossing reward information would not resolve the confound in the present VDN comparison. We therefore prefer a narrow, transparent second bundle over a broad claim of architecture robustness.

Contention and Cooperative AI. Price-of-anarchy analysis compares equilibrium welfare with a social optimum (Koutsoupias and Papadimitriou, 1999; Roughgarden, 2005; Nisan et al., 2007). Our absolute reward gap is not that ratio: it includes an infeasibility floor and finite-horizon control costs. Common-pool MARL environments model appropriation and depletion directly (Pérolat et al., 2017); our task contains only rivalry in control of one shared state, so "contention" is the precise term. Cooperative AI asks when heterogeneous agents can find mutually beneficial outcomes (Dafoe et al., 2020, 2021). The unresolved measurement gap across these literatures is whether a purported alignment effect comes from a narrow ideal-point manipulation, the interaction structure, or the reference used to score learning. The present battery separates those components in one controlled family rather than asserting a general law.

The resulting research gap is methodological rather than algorithmic. Existing work supplies better

learners, better benchmark protocols and examples of construct failure. What is missing is a compact demonstration that sampler, scale, environment and performance reference can each change the qualitative conclusion of the same mixed-motive sweep, together with a control matched to each choice. That is the role of the battery below.

Price-of-anarchy language helps sharpen what the reported gap is not. Classical inefficiency ratios compare an equilibrium outcome with a feasible social optimum under a specified game. Our first-stage *DeltaR* instead compares a finite-horizon learned trajectory with the sum of agents' separate unconstrained ideals. When target coordinates disagree, that reference is not jointly feasible; when resets are random and actions bounded, even the best controller pays to travel. Calling the difference a learning failure would therefore import an equilibrium interpretation that the numerator and denominator do not support. The dynamic oracle below borrows the benchmarking discipline of the price-of-anarchy tradition while retaining an absolute gap, because the zero-valued ideal makes ratios unstable and because the decomposition, rather than a scalar ratio, is the object needed for attribution.

The distinction also places this study beside work on benchmark validity rather than beside papers proposing a new MARL algorithm. The learner is intentionally ordinary. The intervention is on what the experiment means: first the target generator, then the physical coupling, then the support of the targets, then the evaluation and reference point. A stronger optimizer could reduce the level of the residual while leaving every construct problem intact. Conversely, a transparent learner can reveal a valid directional effect if the treatment, interaction and benchmark have been audited. This is why IQL and VDN are used as two learner/objective bundles rather than as a leaderboard comparison.

Finally, the word "control" has different force across the battery. The separable design is a negative endpoint: for independent learners its flatness follows from marginal invariance, so it bounds the domain but cannot by itself establish a mechanism. Partial sharing is the informative intervention because agents still interact while the fraction of jointly moved coordinates varies. Likewise, frozen evaluation removes updating and action noise but cannot by itself remove the infeasibility floor; feasible-centred support and the dynamic oracle complete that repair. Reading the controls as a linked sequence avoids asking any one of them to identify more than it can.

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* α (agreement among agents' objectives or reward functions,

$\alpha \in [-1,1]$), *stake* σ (how much agents care, $\sigma \geq 0$), and *entropy* ε. The experiment implements ε 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: ρ is correlation among target coordinates within one fixed quadratic reward family, and σ 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 σ, so doubling every agent's stake doubles friction ($F \propto \sigma$, degree-1 in σ); (D2) *entropy monotonicity*—F is nondecreasing in ε, with $\varepsilon = 0$ (perfect information) the friction-minimizing floor; (D3) *alignment monotonicity*—F is nonincreasing in α ($\partial F/\partial \alpha \leq 0$); (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 α attains its α-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 σ 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 α-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-α shape it was built to capture was the sampler artifact (§7.5).

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-variance 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 α directionally: positive α is cooperative, negative α 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 ρ directly and holding marginal target variance fixed. It does not solve the second. Accordingly, the repaired experiment tests the directional prediction only for this target-coordinate instantiation; claims about general alignment remain theoretical scope conditions rather than measured results.

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 $\mathcal{S} = [0, C]^m$ (resource levels, capacity $C = 10$), $n = 4$ agents, $m = 3$ shared resources, a per-agent action space $\mathcal{A}_i = \{-1, 0, +1\}^m$ (request a decrease, maintain, or request an increase for each resource, so $|\mathcal{A}_i| = 3^m = 27$), 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 σ sets the weight magnitudes: each $w_{ij} \sim \mathcal{N}(\sigma, 0.05^2)$, clipped to $[0,1]$, so every weight is approximately σ and the reward (2) is, to leading order, homogeneous of degree one in σ ($R_i \approx \sigma \sum_j u(s_j - \tau_{ij})$); we use this in the effect-size analysis (§5.2). The implemented information-loss proxy ε is the observation-noise variance parameter (Definition 3.3). Nominal α 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,

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

ranging from 0 at $\alpha = 0$ to ≈ 0.94 at $|\alpha| = 0.8$. Two consequences matter for interpretation. First, (4) depends only on $|\alpha|$: the design varies the *strength* of shared preference 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-α cells are not anti-aligned but carry the same shared structure as their positive counterparts. Second, the degree-1 homogeneity in σ means the reward gap scales mechanically with stakes—a fact we exploit, and partly correct for, in §5.2.

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 ε corresponds to less accurate information about the true state. The code uses standard deviation $\sqrt{\varepsilon}$, so ε indexes covariance, not amplitude; 0 is the degenerate no-noise boundary. It is a variance-level proxy, not a direct measure of entropy or general information deficit.

3.3 Factorial design

We manipulate the three factors in a full $5 \times 5 \times 5$ factorial 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. The final 100 episodes of the original factorial retain parameter updates, condition-specific resets and the 1% action-exploration floor. We therefore call this a late-training outcome, not learned-policy evaluation. The repaired control battery freezes every trained policy and scores a common bank of 100 held-out reset seeds under clean greedy, matched-observation-noise, and matched-noise-plus-1%-exploration protocols. The late-training outcome remains in the record as a distinct estimand rather than being silently replaced.

3.4 Dependent variables and reference points

The analysis uses one outcome: the reward gap $\Delta R = R^* - \bar{R}$, where $\bar{R}$ is the mean agent payoff averaged over the declared outcome window. Three secondary proxies in earlier drafts—sample-relative "Pareto inefficiency," an agent-aggregated policy-variance statistic and an impossible-valued convergence-time measure—are withdrawn because their implementations did not match their names. They are documented in full in the technical supplement at `TS-PROXY` and are not used to corroborate any result here.

Because every reward is a negated quadratic, each agent's unconstrained ideal is zero:

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

This is a constant ideal, not an attainable team reward. The achievable continuous single-state optimum is

$$R^{\text{cont}} \equiv R^{\text{joint}} = \max_{s \in [0,C]^m} \frac{1}{n} \sum_i R_i(s) \leq R^* = 0. \tag{6}$$

When targets differ, one shared state cannot satisfy all of them. With legacy targets centred at zero, roughly half of target coordinates are below the feasible box and increase the gap even for a joint controller. Moreover, R^{joint} ignores random initial states, bounded integer actions and rewards accrued while travelling. We therefore report ΔR as a total coordination gap and reserve policy interpretation for the frozen finite-horizon decomposition in Section 6.4. Reward-scale-normalized analyses use $\Delta R/\sigma$ to remove the leading degree-one reward scale, while retaining raw values where the scale itself is the estimand.

The two references answer different questions and neither is discarded. R^* keeps the observed gap an exact affine transform of mean reward, making the original factorial transparent. The dynamic controller

asks how much of that gap remains after respecting the environment agents actually face. A condition can therefore have a large total gap and a small policy residual if its ideal points are mutually infeasible; it can also have a favourable attainable optimum and a large residual if learners fail to reach it. Treating the latter as a component rather than a replacement prevents the repaired analysis from changing the dependent variable after seeing the result.

The late-training, clean frozen, matched-noise and restored-exploration outcomes are likewise separate estimands. Their comparisons diagnose particular channels. Late training versus restored exploration includes parameter-update and reset-stream differences; matched noise versus clean evaluation isolates the evaluation-time information channel; restored exploration versus matched noise isolates action randomization with parameters fixed. We do not subtract these quantities and call the remainder a universal learning cost, because learner objectives and decentralization remain different from the joint oracle.

3.5 Signed, support and interaction controls

The repaired signed sampler draws the n targets for each resource from $\mathcal{N}(\mu, \Sigma(\rho))$, where

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

For $n = 4$, positive semidefiniteness bounds the equicorrelated arm at $\rho \geq -1/3$; the sampled grid is $\{-.33, -.2, -.1, 0, .2, .4, .6, .8\}$ with $\sigma \in \{.6, 1\}$ and 30 replication blocks. The endpoint $-.33$ is interior but near the bound: $\Sigma(-.33)$ has smallest eigenvalue .01, whereas the exact $-1/3$ matrix is singular. A two-team block reaches stronger between-team opposition, while an $n = 2$ diagnostic reaches latent-target $\rho = -1$ at the cost of changing agent count. Legacy support uses $\mu = 0$; the support control uses $\mu = 5$ and rejects the vanishingly rare draw outside $[0, 10]$. Both use the same covariance and paired latent problems. A 20,000-draw gate reproduces every requested correlation within .001 and places all checked feasible-centred targets inside the box.

Translation is chosen instead of clipping. Coordinate-wise clipping changes the covariance nonlinearly and attenuates negative dependence more strongly than positive dependence. Adding five to the same Gaussian leaves covariance and latent ordering unchanged while placing essentially all mass in the box; paired hashes verify that feasible and legacy rows differ by this translation rather than by redrawn tasks or networks. The support comparison is therefore a controlled change to whether preferred states are deliverable, not a comparison of unrelated problem samples.

The strong-opposition diagnostics serve a different purpose. The team block can deliver negative between-team correlation beyond the equicorrelation bound but makes teammates cooperative. The $n = 2$ design can reach latent -1 without that within-team structure, but reducing agent count changes interference and the feasible box attenuates the treatment. We use these arms to check for an opposition advantage or singularity, not to estimate the main four-agent slope.

Definition 3.4 (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.

Definition 3.5 (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 3.4 and $k = m$ is Definition 3.1 at the level of state, observation, transition, action and reward maps. Because coordinate subsets are not counterbalanced, the overlap fraction is not identified independently of coordinate position.

The endpoint construction is exact at the level of the formal maps. Separately, a deterministic short-run smoke gate compares the reported training reward, three frozen-evaluation rewards, convergence time and agent rewards across implementations for both IQL and VDN. In its two- replication, five-episode check, $k = 0$ matches the separable runner and $k = m$ the shared runner to 10^{-6}. The gate establishes numerical equality of those reported endpoint summaries; it does not instrument every internal target, network, transition or reset. Only the two intermediate overlap arms require new training in the full follow-up, and inference clusters paired problem blocks by $(\rho, \sigma, \text{replication})$. See `TS-ENDPOINT`.

3.6 Frozen evaluation and analysis status

Each trained replication is frozen and evaluated for 100 episodes on a held-out reset bank under three protocols. The primary outcome is clean greedy evaluation without observation noise. Matched-noise evaluation restores the training condition's observation noise, while a third protocol also restores the 1% action-exploration floor. No protocol updates parameters. The signed support, separable and partial-sharing controls have $\varepsilon = 0$, so their matched-noise columns equal clean evaluation by construction; the restored-exploration columns remain distinct. Only the signed-noise crossing can distinguish clean from matched-noise scoring. Condition, problem-instance and latent-problem hashes make pairing auditable at row level.

The replication block, not an episode or timestep, is the inferential unit. Frozen evaluation averages the common 100-reset bank within a trained replication before regression or contrast. Repeated timesteps are therefore never treated as independent observations. Partial-sharing standard errors cluster the same latent problem across overlap; support-change tests pair the translated and legacy versions of that problem. HC3 uncertainty is retained for the declared slopes because variance differs materially across stakes and opposition cells.

Only the initial H1–H4 factorial hypotheses and its $R^2 > .7$ adequacy rule preceded estimation. The

signed redesign and later controls were created after diagnosis. We designate a four-contrast exploratory family for the legacy signed endpoints and apply Holm–Bonferroni correction; the feasible-support opposition family, learner interactions and signed-noise interactions have their own explicitly named Holm families. Other model comparisons and sensitivity analyses remain exploratory. Estimates, confidence intervals and non-detections are interpreted as controlled demonstrations requiring independent replication.

Multiplicity families follow the question rather than collecting every reported diagnostic into one omnibus correction. The four legacy endpoint contrasts answer two opposition/cooperation questions across two learners. The four feasible opposition contrasts answer the reviewer-raised within-stake question. The two learner interactions form the signed-noise family. Conditional cell tables and shape sensitivities remain descriptive or carry their own stated adjustment. This partition was fixed in the repair ledger before final regeneration, but it remains post-diagnostic and exploratory.

3.7 Local provenance and regeneration

The local scientific record separates raw outcomes, analysis-ready tables and manuscript prose. The repaired first-stage ANOVAs, interactions, functional comparisons and effect-size partitions are regenerated together from the post-fix factorial rather than copied from earlier tables. The Reviewer 3 follow-up analysis derives frozen outcomes, oracle components, within-stake contrasts, support-change slopes and their multiplicity adjustments from row-level CSVs. Separate scripts assemble the paired partial endpoints, analyze the signed-noise crossing, verify the dynamic oracle and smoke-test the partial-sharing endpoints. Manuscript numbers were checked against those generated tables during the repair, but the scripts and local CSVs remain the evidential source of record.

Pairing is asserted only where it is encoded. Each follow-up row records learner, support, ρ, σ, replication index, reset and initialization seeds, plus a digest of the latent problem instance. Support comparisons require the same digest after removing the five-unit translation. Partial-sharing comparisons reuse the same problem block across k. Signed-noise pairs reuse the clean/noisy problem and initialization. The archived legacy IQL/VDN runs share preference draws but not provably identical network initialization, so their cross-learner level comparison is paired only at the problem level. These distinctions prevent a visually matched row order from being mistaken for experimental pairing.

The paired-difference model matrices are explicit. For the support control, each verified block $b = (\rho, \sigma, \text{replication})$ supplies $D_b = g_b^{\text{feasible}} - g_b^{\text{legacy}}$ and the reported slope comes from $D = X\beta + u$ with $X = [\mathbf{1}, \rho]$ and HC3 covariance. For signed noise, each verified block supplies $D_b = g_b^{\varepsilon=1} - g_b^{\varepsilon=0}$; the pooled interaction model uses $X = [\mathbf{1}, \rho, \mathbf{1}\{\sigma = 1\}]$ and reads the coefficient on ρ, while the reward-scale-specific models use $X = [\mathbf{1}, \rho]$. The separate $X = [\mathbf{1}, \rho, \mathbf{1}\{\sigma = 1\}, \rho\mathbf{1}\{\sigma = 1\}]$ sensitivity tests whether those two slopes differ. Thus the paired claims arise from within-block differences, not from an unpaired level regression with a pairing label attached.

Regeneration and availability are separate gates. Zenodo v3.0.0 archives the checked August code and raw outputs under immutable record DOI 10.5281/zenodo.22004215; the GitHub default branch

is synchronized to the release line. File paths and script names identify the exact analysis surface, while the record DOI, inventories and hashes identify the immutable public payload. This distinction prevents a mutable working tree from being treated as the archived evidence version. The 24 August problem-block bootstrap and protocol-summary delta postdates v3; the arXiv source package carries its exact script, derived CSVs and provenance README as ancillary files, while repository/Zenodo synchronization remains a separate release task.

3.8 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 documented at `TS-IMPL`), 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 ε_{exp}-greedy exploration rate (distinct from the entropy/observation-noise candidate coordinate ε) 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. 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 (§6.2).

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_{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 ε_{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 compare the canonical single index with alternatives that either combine the factors differently or retain them separately. The first-stage fit addresses the full H4 expression but inherits the invalid signed axis; the repaired signed fit is an exploratory comparison of its $\varepsilon = 0$ restriction. We compare four initial 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 σ alone. The signed refit adds homogeneous M5, $Y = \sigma(\beta_0 + \beta_\rho \rho)$, its free-intercept M6 sensitivity, and quadratic/cubic canonical-index curves M7/M8. 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 (the signed comparison reported below). M4 is deliberately a *linear* main-effects model, not a saturated dummy-coded ANOVA. M5 is fit without an intercept because degree-one homogeneity requires a zero reward gap at zero reward scale; M6 tests that restriction. M3 is explicitly 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 $\alpha_{\text{fw}} = 0.05$ (the conditional contrast family reported below). 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: How the first-stage design manufactured its findings

5.1 A symmetric profile that was not signed target correlation

The first-stage condition means form a clear U over the nominal α axis: the neutral cell has the largest gap, while $\alpha = -.8$ and $+.8$ look similarly favourable. That pattern is real for the sampled data and invalid as evidence about opposition. Equation (4) shows that the two extreme arms have the same positive correlation, approximately .94, while $\alpha = 0$ is the only unstructured point. The signed-linear share is zero by construction. We retain this failed result because it is the object the control battery diagnoses, not because it estimates H1.

The failure is more than a wrong label on the horizontal axis. The marginal target variance dips at intermediate $|\alpha|$, and between-agent target dispersion shrinks as $|\alpha|$ grows. The factorial therefore changes correlation magnitude and dispersion together. A symmetric outcome is not surprising once the generator is written in realized rather than nominal coordinates. The signed repair holds marginal variance constant and varies ρ directly, so it repairs both the missing sign and the changing spread. Comparisons between the two sweeps are cross-design diagnostics rather than a pure relabeling exercise.

The descriptive scale of the failed result explains why it was persuasive. Across the post-fix 3,750 replications, mean reward ranges from -4.568 in the worst cell ($\alpha = 0$, $\sigma = 1$, $\varepsilon = .75$) to $-.418$ in the best ($\alpha = .4$, $\sigma = .2$, $\varepsilon = 0$), a roughly 10.94-fold difference in the zero-referenced gap. The nominally neutral target arm has mean reward -2.526, compared with $-1.659, -1.457, -1.468, -1.615$ at $\alpha = -.8, -.4, +.4, +.8$. In a per-replication ANOVA the nominal α main effect is large ($F = 185.44$, $\eta^2 = .0963$, $p < .001$), and its interaction with stakes is detected ($F = 10.34$, $\eta^2 = .0215$, $p < .001$). Those are accurate statistics for the generated cells. They do not validate the signed interpretation, because the exchangeability of $\pm\alpha$ is built into the target distribution.

This is an important difference between statistical and construct replication. The U reappears under the VDN/team-reward bundle, with a normalized U-depth of 1.589 versus 1.473 under IQL, and the repaired GPU and CPU implementations reproduce the condition ordering. Both facts make the profile more reproducible while doing nothing to restore the missing adversarial treatment. A finding can be stable across seeds, implementations and learner bundles and still be about $|\alpha|$ when the paper names signed α. The sign gate, not another replication of the same generator, is what changes its evidential status.

5.2 Reward scale dominates only before normalization

On the 125 condition means, raw main-effect shares are approximately .74 for reward scale, .18 for nominal α and .015 for observation noise. This ranking follows from reward homogeneity: to leading order $R_i = \sigma \sum_j u(s_j - \tau_{ij})$, so multiplying σ expands both the mean gap and its dispersion. On the reward-scale-normalized gap, the ordering reverses: nominal α carries approximately .82, noise .064

and reward scale .004. The per-replication partition changes the magnitudes because it retains within-condition variance, but not the diagnosis. Neither partition is a structural ranking; raw and normalized views answer different scale questions.

This distinction separates a supported direction from an unsupported ranking. Higher σ raises raw losses in every design because the reward definition makes it do so. The experiment does not independently discover that theoretical stakes matter, and it cannot infer that reward scale is more important than target dependence from a grid whose units make the former multiply the outcome. Reporting both scales preserves the descriptive data without converting a dimensional choice into theory.

Normalization is nearly, but not perfectly, exact because the weight sampler is not a pure scale family at the top of the grid. It draws $w_{ij} \sim \mathcal{N}(\sigma, .05^2)$ and clips to $[0,1]$. At $\sigma = 1$, approximately half the proposed weights are clipped, lowering their mean slightly and creating a boundary atom; below .8 clipping is negligible. This is why the condition-level reward-scale share falls to about .004 rather than literally zero after division by σ. The residual does not rescue a dominance claim. It records that the top grid point is an imperfect rescaling and motivates reporting normalized effects within stake as well as pooled over the declared mixture.

The two variance partitions also operate at different inferential levels. The condition-level partition is a saturated description of 125 cell means and carries no residual degrees of freedom; it is useful for asking how the chosen grid apportions between-cell variation, not for testing a population effect. Inferential models use the 30 replication blocks per cell. Agents and timesteps inside one block are not independent observations. A diagnostic agent-level regression illustrates the cost of ignoring that structure: clustering the four agents back to their replication weakens a nominal interaction from about .02 to .11 IQL and .086 VDN. All reported tests therefore aggregate evaluation episodes within a trained replication and use that replication as the unit.

Observation noise in the first-stage factorial is similarly a descriptive starting point rather than evidence about genuine opposition. Its per-replication main effect is modest ($F = 15.76$, $\eta^2 = .0082$, $p < .001$), moving mean reward from -1.549 at $\varepsilon = 0$ to -1.861 at 1. The sign-blind $\alpha \times \varepsilon$ interaction is not detected ($F = 1.15$, $p = .302$); the $\sigma \times \varepsilon$ interaction is small ($F = 1.83$, $\eta^2 = .0038$, $p = .023$), and the three-way interaction is not detected ($F = .94$, $p = .622$). Those results motivate, but do not answer, the signed feasible-centred noise crossing in Section 6.5.

5.3 The full H4 claim is not confirmatorily adjudicated; its reduced restriction loses

On the sign-blind factorial, an independent-effects model greatly outperforms the canonical single-index model; equal-F cells also have sharply different outcomes. That grid crosses all three factors but cannot test the signed denominator it names. The repaired signed grid delivers the target-correlation sign but fixes $\varepsilon = 0$ and was designed after diagnosis. The full H4 claim therefore receives no valid confirmatory verdict from either design. As an exploratory reduced-grid calibration, the canonical $\sigma/(1+\rho)$ restriction reaches $R^2 = .666$ for IQL and .717 for VDN; the latter's numerical crossing of the old .7 threshold is not an H4 pass. A linear model retaining ρ and reward scale σ separately reaches .935 and .925

and beats the canonical model by strict ΔAICc $= 22.5$ and 17.7; reward scale alone also beats it. More informatively, the degree-one homogeneous response $\sigma(\beta_0 + \beta_\rho \rho)$ reaches .959 and .958 and improves on the canonical model by ΔAICc $= 33.45$ and 30.51. Giving this form a free intercept does not improve AICc, and quadratic or cubic response curves in the canonical index do not rescue it. The exploratory comparison therefore favours a scale-homogeneous linear response in target correlation, not the reciprocal denominator. Neither result identifies objective- or reward-function alignment: ρ is target-coordinate correlation in this task.

Unequal-cell-precision and replication-row sensitivities preserve the ordering: M5 beats M1 by 27.60–50.79 AICc units and M4 by 6.42–8.54. The replication likelihood is descriptive working independence despite matched problem blocks, and criteria are compared only within learner and estimator. Equal-condition OLS remains canonical because it weights the declared cells equally; the historical H4 threshold remains a reduced-grid diagnostic, not confirmatory reuse.

Taken together, these exploratory controls weigh against the complete composite as a maintained single-predictor description in this environment; this is not a confirmatory H4 verdict and does not withdraw every coordinate. The implemented reward coefficient scales losses by construction. Signed target correlation has a conditional direction under shared contention. Observation noise requires the repaired signed crossing reported below. The symmetric refinement $1/(1+\alpha^2)$ is withdrawn because it was designed to fit the sampler artifact. At $\varepsilon = 0$, the original multiplicative competitor is degenerate, so the signed model comparison does not itself test the entropy numerator. Section 6.5 supplies that separate test and contradicts its interaction direction.

The adverse single-index evidence does not depend on choosing M4. Equation (1) implies that cells sharing the same F must share the same expected outcome up to a common one-dimensional response function. In the sign-blind grid they do not: a six-way $F = 1$ tie is sharply heterogeneous, and no monotone transform of F explains more than about .55 of the between-cell variance. Because ties that cross α inherit the invalid signed axis, a narrower model-free check holds α fixed and compares only equal-F combinations of reward scale and noise. Fourteen of 25 within-α tie sets have raw $p < .05$, ten survive Holm family-wise correction and twelve survive Benjamini–Hochberg control. This is still exploratory and uses the first-stage design, but it shows that collapsing reward scale and noise into the same scalar fails even without crossing the broken target-correlation manipulation.

Implementation repair record. The displayed first-stage values use the repaired GPU factorial. Its observation-noise and replay paths were corrected and rerun; a separate CPU action-decoder defect was also repaired. The post-fix CPU/GPU condition means agree closely, but simultaneous repairs prevent attribution of the remaining offset to any one fix. The complete defect chronology, pre-fix rows, agreement statistics and portability changes are retained at `TS-IMPL`. Implementation agreement is not used as evidence that the original sampler realized signed target correlation.

6 Results II: The Control Battery

The repaired analyses use ρ for genuine signed cross-agent target correlation and $\overline{g}/\sigma$ for the reward-scale-normalized gap. They proceed from treatment verification to shared-state scope, paired overlap, support repair, dynamic benchmarking and signed noise. Except for H1–H4, every inferential result in this section is exploratory.

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-α 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 ρ 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 ρ 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

$$\Sigma(\rho) = (1-\rho)\, I_n + \rho\, \mathbf{1}\mathbf{1}^\top, \tag{8}$$

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 bounded below by $\rho = -1/3$. Marginal target variance is held at 1 for every ρ, 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 $\mathrm{opp} \in [0,1]$. Within-team correlation is $+\mathrm{opp}^2$ and between-team correlation is $-\mathrm{opp}^2$, reaching -1.0 at $\mathrm{opp} = 1$ (genuine strong opposition; mean pairwise correlation $-\mathrm{opp}^2/3$).

In both designs the weights $w_{ij} \sim \mathcal{N}(\sigma, 0.05^2)$ clipped to $[0,1]$ are *unchanged*, preserving σ as the implemented reward-scale proxy for stakes, and the IQL/VDN learners are reused verbatim—only the

preference sampler is swapped. A pre-run verification (Figure 1, right) confirms that realized equicorrelated correlation tracks its target essentially exactly—from $-.330$ at requested $-.33$, through zero, to .800—and that 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 → zero → 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 §7.5. 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 is verified for that narrower treatment; it does not establish equivalence with general objective or reward-function alignment.

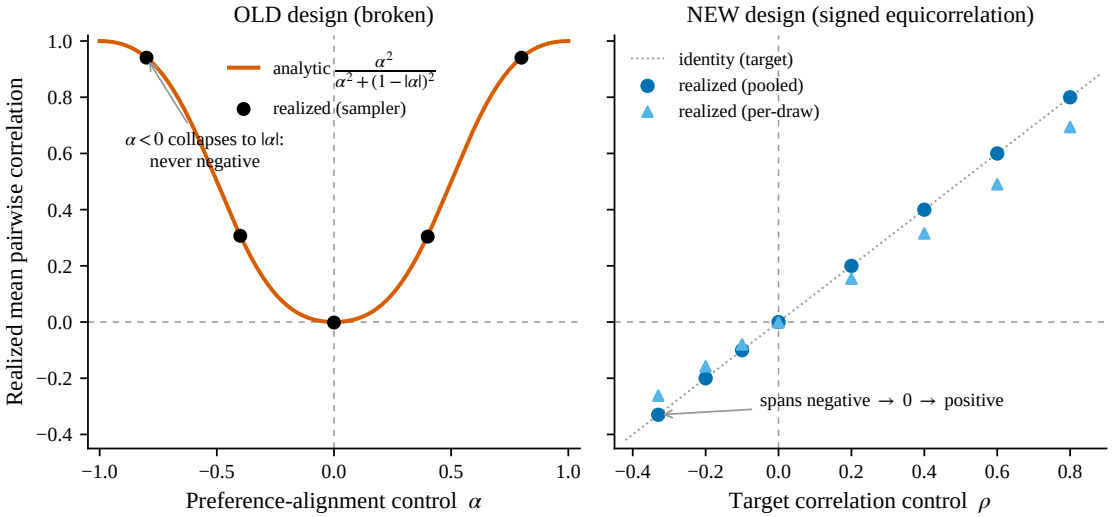


Figure 1: The signed equicorrelation DGP fixes the $|\alpha|$-symmetry artifact of the original sampler. Left: the old design's realized mean pairwise correlation is $\alpha^2/(\alpha^2+(1-|\alpha|)^2)$, symmetric in $|\alpha|$ and never negative. Right: the new control spans negative → 0 → positive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

The signed sweep retains the original training budget: eight ρ levels, two reward-scale levels, 30 replications and 1,000 episodes for each of IQL and VDN. The batched runner evaluates all replication-by-agent networks in tensor operations; on the recorded two-GPU setup the complete equicorrelated sweep took about 22 minutes. Runtime is not part of the inferential claim, but the batch layout matters for provenance. Preference draws depend on condition stream and replication index rather than learner,

so matched IQL and VDN rows receive the same targets and weights. The archived legacy runs predate deterministic PyTorch seeding, however, so they do not guarantee identical initial network weights. Later feasible, partial-sharing and signed-noise controls record condition and evaluation seeds plus problem-instance and latent-problem hashes and use those identifiers, rather than row order, to verify pairing.

The legacy signed and contention arms fix $\varepsilon = 0$ to isolate the repaired target-correlation axis. That choice prevents the observation bug from affecting those archived runs, but it also means they cannot establish how noise behaves under genuine opposition. The reduced signed-noise design is a new crossing, not an extrapolation from the first factorial. Similarly, stake normalization removes the leading degree-one scale but does not change which σ cells were sampled. Every pooled contrast is consequently labelled as the equal-weight average over the exact $.6/1$ mixture.

6.2 Signed target correlation on legacy and frozen outcomes

The signed design removes the U-shape. Unless qualified otherwise, "opposition versus the uncorrelated-target baseline" means the endpoint contrast $\rho = -.33$ versus 0 in the four-agent equicorrelation design; it is not perfect opposition. On the legacy late-training outcome, the normalized-gap slope is -0.755 for IQL and -0.838 for VDN. More positive target correlation lowers the gap; the opposed endpoint does not produce the advantage suggested by the sign-blind profile. Pooling reward-scale levels, however, is not a stake-invariant estimand. In the late-training data, the within-stake opposition-minus-baseline differences change sign and none survives the four-test Holm family; the $\rho \times \sigma$ interactions are not detected ($p = .129$ IQL, .097 VDN). The pooled non-detection is therefore only the equal-weight average over $\sigma \in \{.6, 1\}$, not a general opposition result and not evidence of equivalence.

The conditional values make the cancellation visible. At $\sigma = .6$, opposition lies .644 IQL and .523 VDN *below* the uncorrelated-target baseline (raw $p = .080/.098$); at $\sigma = 1$, it lies $.337/.205$ above it ($p = .466/.609$). The endpoint difference-in-differences is likewise not detected ($p = .099/.157$), and none of the four within-stake contrasts survives its Holm family. The archived pooled differences, $-.154$ IQL and $-.159$ VDN, are therefore estimands for a specified equal mixture of two opposite-signed cells. They cannot support either a conditional zero-effect claim or equivalence. Cooperation-versus-baseline is more consistent in the same outcome: $-.707$ IQL and $-.924$ VDN pooled over stakes, although only VDN survives the four-contrast family ($p_{\text{Holm}} = .0071$; IQL .144).

The archived CSV-only legacy runs cannot be frozen retrospectively. A later deterministic rerun on the same target-support design—not the same trained policies—records its own late-training slopes of -1.082 IQL and -1.078 VDN and clean frozen slopes of -1.156 (95% CI $[-1.695, -.618]$, $p = 2.6 \times 10^{-5}$) and -1.162 ($[-1.515, -.808]$, $p = 1.2 \times 10^{-10}$). All four frozen within-stake point estimates in that rerun place opposition above the baseline, but only IQL at $\sigma = 1$ survives Holm (difference $+1.016$, adjusted $p = .034$). This is a mixed legacy result, not equivalence and not a replicated opposition penalty. It is a cross-run sensitivity, not a within-run protocol contrast.

In that deterministic rerun, the pooled frozen opposition estimates are $+.705$ IQL and $+.328$ VDN;

IQL is raw-significant but not Holm-significant ($p = .0217$, adjusted .0652), while VDN is not detected. Within stakes the four differences are $+.394$, $+1.016$, $+.198$ and $+.459$ for IQL .6, IQL 1, VDN .6 and VDN 1, respectively. This uniform sign but single surviving cell contrasts with the late-training sign reversal. We therefore do not present one protocol as recovering a true legacy opposition effect. Because the archived and frozen values come from different trained runs, their disagreement cannot isolate updates, reset streams or residual exploration. It shows only that the endpoint contrast is sensitive across runs and evaluation constructions while the full-grid target-correlation slope keeps its sign.

Stronger-opposition diagnostics. The four-agent equicorrelation arm cannot pass $-1/3$; its sampled endpoint is $-.33$. Separate two-team and $n = 2$ diagnostics do not establish an opposition advantage, but each changes the estimand: the first adds within-team cooperation and the second changes agent count while delivering feasible-optimum correlation $-.468$. Neither is an equivalence test or evidence about divergence as $\rho \to -1$. Their full tables and qualifications are retained at `TS-OPP`.

Second learner/objective bundle. The signed curve under VDN correlates .976 with the IQL curve and shifts the shared-pool level downward. This is shape replication only. VDN simultaneously replaces independent utilities with additive value decomposition and individual rewards with the summed team reward, converting the training objective to full cooperation. Its level advantage cannot be assigned to centralization, and the separable control shows it is not universal.

Across the 480 matched legacy signed runs, that level shift averages .690 in normalized-gap units (95% CI $[.652, .729]$) and ranges from .627 to .878 across ρ. The near-constant shift is useful precisely because the claim is about shape: the bundle neither creates a second opposition trough nor selectively rescues the adversarial endpoint. Once state is separable, VDN's mean normalized gap is instead slightly higher than IQL's. The reader should therefore carry forward the shared curve ordering, not a general claim that centralized training lowers friction.

6.3 Shared-state contention and paired partial sharing

The fully separable endpoint removes the only common state. In the archived legacy late-training outcome, its slope is $-.029$ ($p = .87$) IQL and $-.053$ ($p = .76$) VDN; one-way tests and both endpoint contrasts are also not detected. The IQL expectation is structurally ρ-invariant because every marginal problem is invariant to cross-agent target correlation. VDN retains additive team-reward coupling, so its finite-training slope is an empirical non-detection, though the optimal control problem remains separable. In the deterministic feasible-centred rerun, the clean frozen held-out slopes are small and not detected: $+.004$ (95% CI $[-.018, .026]$, $p = .718$) and $+.020$ ($[-.019, .058]$, $p = .318$), and neither paired support change is detected. The endpoint bounds the claim to the tested state coupling; it does not identify a unique mechanism among every possible reward, observation or communication coupling.

Absolute levels reinforce why the endpoint is a scope check rather than a learner comparison. The legacy separable IQL gap averages about 3.90, close to the shared-pool $\rho = 0$ baseline, while VDN's separable gap averages about 4.05 and sits above its shared-pool cells. Thus the VDN bundle's lower

shared-pool level does not transport to factorized dynamics. More fundamentally, IQL separability makes the expected ρ curve flat: changing dependence between other agents' targets cannot alter any one agent's marginal target distribution, private transition or reward. The non-detected VDN slope adds a finite-training check despite team-reward coupling, but neither arm rules out interaction through rewards, observations, communication or opponent modelling.

The paired legacy-Gaussian-support partial-sharing series supplies the interaction-preserving test. Frozen clean slope point estimates steepen monotonically across overlap $w = 0, 1/3, 2/3, 1$: IQL $-.219, -.566, -.807, -1.156$ and VDN $-.125, -.594, -.895, -1.162$. Every partially or fully shared slope is detected, whereas neither separable slope is. The problem-block-clustered interaction is $\rho \times w = -.916$ (95% CI $[-1.304, -.528]$, $p = 3.7 \times 10^{-6}$) for IQL and -1.023 ($[-1.260, -.787]$, $p = 2.0 \times 10^{-17}$) for VDN. Categorical modulation is also detected. Departure from a linear ramp is not ($p = .890/.191$), so four overlap levels support graded modulation but not an exact linear law or majority-sharing threshold.

A baseline-difficulty explanation does not fit additively: at $\rho = 0$ the clustered frozen baseline declines rather than rises with overlap, by $-.390$ IQL ($p = .0045$) and $-.849$ VDN ($p = 3.7 \times 10^{-8}$). The unpaired HC3 sensitivity is less precise for IQL ($p = .119$) but not VDN ($p = .0018$). Overlap could still interact with credit-assignment difficulty in a way not separated here. Under the fixed first-k coordinate ordering, the result identifies modulation with the shared-coordinate fraction, not causal mediation by a uniquely isolated mechanism.

The interpolation is paired at the problem level. One seed identifies each $(\rho, \sigma, \text{replication})$ block and is reused over w, preserving targets, weights, initial networks and evaluation starts while changing only which coordinates receive summed actions. A deterministic short-run smoke gate now covers both learner modes and compares the reported training, frozen-evaluation, convergence and agent-reward summaries. Those summaries agree to 10^{-6} between $k = 0$ and the standalone separable runner and between $k = 3$ and the standalone shared runner; the gate does not separately instrument internal transitions or reset banks. Only $k = 1, 2$ require new training. This matters because independently redrawn endpoints could imitate a dose response through problem difficulty; the paired construction turns overlap into a within-block intervention.

The word "dose" refers to slope modulation, not to a claim that sharing makes the task uniformly harder. At $\rho = 0$, the mean frozen baseline decreases with w, so a simple additive credit-assignment penalty has the wrong sign. The interaction can still combine physical contention with ρ-dependent credit assignment, since both change when a coordinate becomes shared. Four attainable overlap levels also give little power against smooth curvature. The supported statement is therefore that the signed-gradient point estimates strengthen monotonically over the four tested fixed first-k shared-coordinate configurations; the data neither identify an exact linear law nor a unique mediating mechanism.

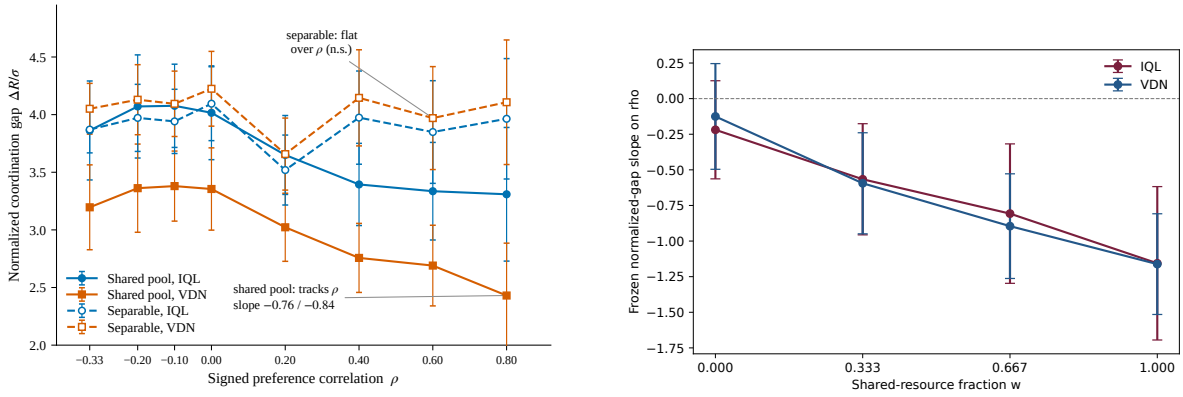


Figure 2: Interaction-structure scope and overlap response on legacy-Gaussian-at-zero target support. Left: the signed gradient appears with shared state; the separable IQL expectation is ρ-invariant and the VDN slope is not detected. Right: on clean frozen held-out evaluation, the slope point estimates steepen over the four tested fixed first-k shared-coordinate configurations. Absolute levels are not compared across learners; error bars on the right are HC3 95% CIs.

6.4 Control: dynamic benchmarking separates feasibility from frozen-policy performance

The static quantity R^{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. For the clean greedy outcome R_T^π, an exact $T = 100$ dynamic oracle yields

$$R^* - \mathbb{E}R_T^\pi = \underbrace{R^* - R^{\text{cont}}}_{\Phi\text{ continuous feasibility}} + \underbrace{R^{\text{cont}} - \mathbb{E}R^{\text{lat}}}_{Q\text{ reachable-lattice cost}} + \underbrace{\mathbb{E}(R^{\text{lat}} - R_T^{\text{dyn}})}_{T_c\text{ travel transient}} + \underbrace{\mathbb{E}(R_T^{\text{dyn}} - R_T^\pi)}_{P\text{ frozen-policy residual}} . \tag{9}$$

R^* and R^{cont} are deterministic conditional on a target-weight instance; R^{cont} uses its clipped reward-scale-weighted target mean, while R^{lat} is the best state on the exact reset-specific 21-point lattice per resource; and R_T^{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 3.98×10^{-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; the 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 oracle is exact here because the joint problem factorizes by resource even though the agents' policies do not. For one coordinate, averaging the four quadratic rewards gives $-c(x-\bar{\tau}_w)^2 - v$, where c is the mean total weight, $\bar{\tau}_w$ is the weighted target mean and v is the irreducible dispersion term. The continuous optimum clips $\bar{\tau}_w$ to the box. Four integer agent actions give an aggregate move in $\{-4,\dots,+4\}$; from a reset with fractional part f, every later state lies on the 21-point interleaving $\{0, f, 1, f+1, \dots, 9, f+9, 10\}$, with duplicates harmless when $f = 0$. The transition table on that lattice is independent of the nonzero offset. Backward induction over 100 stages therefore enumerates every feasible joint aggregate action without approximating a continuous controller. Summing the three independent coordinate values is exact because both stage reward and the aggregate action set factor as Cartesian products.

Three gates guard the implementation rather than merely its algebra. First, the full three-resource dynamic program and the sum of separate one-resource programs agree to 3.98×10^{-13}. Second, a self-contained verifier exhausts all two-step aggregate-action sequences for 54 starts—the two boundaries, two integer interior starts and 50 random starts—and matches the oracle in all 54 cases. Third, the analysis checks the four-term identity for every replication with maximum residual below 2×10^{-5} and rejects negative quantization, transient or policy components beyond numerical tolerance. These checks validate the controller for this deterministic transition family; they do not prove that P isolates a general learning construct.

The target-support choice changes the answer. Table 2 reports HC3 slopes of the reward-scale-normalized components on ρ ($n = 480$ 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 does not: $+0.136$ ($p = .463$) IQL and $+0.130$ ($p = .051$) VDN. The corresponding 95% CIs are $[-.228, .500]$ and $[-.001, .261]$; 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% 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 bootstrapped component and total slopes, not constrained mixture proportions; 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.

The support problem is visible before any learner is trained. With targets marginally $\mathcal{N}(0,1)$ and state restricted to $[0,10]$, about half of coordinates prefer a negative state and are clipped to the same lower boundary. The resulting feasible-optimum correlation is about $-.217$ at nominal $\rho = -.33$ and $+.743$ at nominal $+.8$, retaining roughly 66% of the opposed treatment but 93% of the cooperative treatment. At the intermediate requested values $-.20, +.20, +.40, +.60$, feasible correlations are approximately $-.138, +.155, +.335, +.530$. The legacy axis is monotone but asymmetric and compressed toward zero, especially on the opposed side.

The centred control adds five to the same latent Gaussian rather than clipping or rescaling it. Translation preserves its covariance and requested ρ while placing essentially all mass in the box; the runner rejects any rare out-of-box draw. A 20,000-draw pre-run gate confirms all checked targets are feasible and correlations match the grid within .001. Paired hashes verify that a feasible row and its legacy counterpart share the underlying standardized draw, weights, initialization and evaluation starts. The slope difference is therefore attributed to target location relative to the box, subject to ordinary training variation, rather than to a newly drawn collection of tasks. It is nevertheless a support intervention, not a pure feasibility-only manipulation: translation also changes reset-to-target distance and boundary geometry.

The feasible-centred endpoint contrasts make that consistency explicit. At the four-agent equicorrelation endpoint, $\rho = -.33$ minus $\rho = 0$ is $+2.216$ for IQL at $\sigma = .6$ ($p_{\text{Holm}} = .0346$), $+2.113$ for IQL at $\sigma = 1$ ($p_{\text{Holm}} = .00280$), $+1.528$ for VDN at .6 ($p_{\text{Holm}} = .0153$), and $+2.467$ for VDN at 1 ($p_{\text{Holm}} = 4.19 \times 10^{-5}$). The corresponding pooled contrasts are also penalty-signed, but the within-stake values are the relevant answer to the cancellation criticism. Their full-grid reward-scale-normalized $\rho \times \sigma$ interactions are not detected ($p = .527/.515$), so the data support a penalty at both sampled reward scales without supporting moderation of the normalized slope by σ.

At the already-sampled interior values $\rho = -.20, -.10$, all eight feasible-centred point estimates remain above the 0 baseline, although only two survive Holm correction: directional continuity, not eight detected effects.

Table 2: Exact finite-horizon decomposition of the clean frozen-policy gap. Entries are HC3 slopes of each reward-scale-normalized component on ρ; 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	Φ	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

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 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-stake 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.

6.5 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 3 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 ρ rises. Thus the noise penalty is positive on average, but the multiplicative target-correlation–noise claim does not.

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 $\sigma = 1$; neutral and cooperative cells carry most of the mean penalty. Within-σ 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 contrasts keep learning, measurement noise and exploration distinct.

The conditional interaction estimates show where the average comes from. At $\sigma = .6$, the $\rho \times \varepsilon$ slope is $+2.246$ IQL ($p_{\text{Holm}} = .0308$) and $+.766$ VDN (also .0308); at $\sigma = 1$ it is $+.872$ IQL ($p_{\text{Holm}} = .312$) and $+1.158$ VDN ($p_{\text{Holm}} = 4.20 \times 10^{-5}$). The difference between the two reward-scale-specific slopes is not detected for either learner. Collapsing reward scale instead, the clean noise penalty grows from 1.223 at $\rho = -.33$ to 2.340 at $\rho = 0$ and 3.132 at $\rho = .8$ for IQL; the VDN sequence is $.215, .607, 1.320$. After correction the pooled opposition values remain non-detections ($p_{\text{Holm}} = .097/.306$). The evidence therefore supports a positive average penalty and a positive signed interaction in this declared mixture, not the statement that noise hurts in every sampled cell.

Table 3: Functional falsification. Panel A compares the pre-stated ratio with reward-scale-only, independent-effects, degree-one homogeneous and flexible-index alternatives on the signed target-correlation $\varepsilon = 0$ grid. Panel B crosses signed target correlation with observation noise under feasible-centred support.

*Panel A: signed-grid model comparison (*16 *condition means per learner)*

Learner	Model	R^2	AICc	ΔAICc vs M5
IQL	M1: $\sigma/(1+\rho)$	.666	29.840	33.447
	M4: $\rho+\sigma$	.935	7.381	10.988
	M5: $\sigma(\beta_0+\beta_\rho\rho)$	**.959**	$\mathbf{-3.607}$	0
VDN	M1: $\sigma/(1+\rho)$	.717	22.408	30.512
	M4: $\rho+\sigma$	.925	4.704	12.808
	M5: $\sigma(\beta_0+\beta_\rho\rho)$	**.958**	$\mathbf{-8.104}$	0

Panel B: clean frozen signed-noise control ($n = 180$ pairs per learner)

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

AICc uses the strict small-sample parameter count, including residual variance. M5 is fit without an intercept to enforce degree-one reward-scale homogeneity. M0, M2 and the curved-index M7/M8 alternatives are reported in the supplement; M6 adds a free intercept, with none improving AICc. Panel B averages equally over the declared three-ρ, two-reward-scale mixture; interaction uncertainty is HC3 and Holm adjustment covers the two learner interactions. M3 is not estimable here because $\sigma\varepsilon(1-\rho) = 0$ throughout the $\varepsilon = 0$ grid; the complete estimator robustness table is at `TS-COND`.

The three frozen protocols answer different questions. Clean greedy evaluation asks whether training under noisy observations changed the learned policy when that policy is scored without measurement noise. Matched-noise evaluation adds the policy's observation-time error. Restoring the 1% random-action floor adds execution noise while keeping parameters frozen. The last two means are close, so residual action exploration contributes little to their difference; both are larger than the clean penalty, showing that evaluation-time observation noise is not negligible. None of these comparisons recovers the original continued-training estimand, which also includes updates and its own reset stream. Their agreement on the interaction sign is stronger evidence than treating one training window as policy quality.

6.6 Control-battery verdict

Four apparent findings entered the battery and none retained its original interpretation. The nominal-α U was a sign-blind magnitude profile. Reward-scale dominance was degree-one reward scaling. On legacy-Gaussian support, the estimated target-correlation gradient was bounded to shared-state coupling and graded with the shared fraction. The final error was attribution: a continued-training gap against an unconstrained static ideal did not identify policy quality. Frozen evaluation found a same-signed gradient, feasible support estimated a steeper one, and the exact oracle split it into achievability and a support-dependent policy residual.

The feasible-centred endpoint yields a different, support-specific result without eliminating the

pooled legacy ambiguity. The four-agent $\rho = -.33$ endpoint is worse than $\rho = 0$ within both reward-scale levels and learners after Holm correction, whereas legacy frozen estimates are mixed and legacy late-training contrasts change sign. The discrepancy is itself the construct-validity result: conclusions about opposition depend on whether the environment actually delivers the named treatment inside its state space.

7 Discussion

7.1 What the functional retains and loses

The case study yields neither wholesale vindication nor a reason to discard every coordinate. More positive target-coordinate correlation lowers the gap in the shared-state family, consistent with the sign of the proposed alignment denominator only for this narrow instantiation. The effect is not a context-free law of objective or reward-function alignment. At separation the IQL expectation is invariant and the VDN slope is not detected; across the four tested overlap levels on paired legacy-Gaussian-at-zero support, the point estimates grade monotonically. Its policy-residual channel also changes with target support. The proposed divergence toward perfect opposition is not observed, and the equicorrelated many-agent geometry cannot deliver $\rho = -1$ in the first place.

Stakes remain directionally related to raw loss because the reward algebra imposes degree-one scaling. That is a property of the instrument, not independent evidence that stakes dominate learning. Observation noise has a positive mean cost under the repaired feasible-centred mixture, but its target-correlation interaction runs opposite to Equation (1). A canonical ratio retains degree-one scaling in σ, but its one-dimensional form imposes a specific coupling among reward scale, target correlation and observation noise that the separately estimated effects do not follow. The independent-effects model's advantage and the reversed signed-noise interaction therefore weigh against the maintained single-index composition in this environment, not every qualitative intuition motivating it.

The appropriate feedback to the companion theory is specific: retain objective alignment as a candidate theoretical direction, but do not treat this target-correlation experiment as its direct measurement; impose an interaction-structure scope condition; drop the symmetric quadratic refinement; weaken the adversarial singularity; distinguish imposed stake homogeneity from empirical support; and treat information loss as a separate channel whose interaction must be estimated rather than inherited from a convenient numerator.

7.2 Why the controls work together

No single repair would have located the fourth result. Freezing policies removes late updating and separates the action-exploration floor. Held-out resets prevent the training reset stream from being mistaken for evaluation. Moving the same signed Gaussian into the box delivers the nominal treatment. The exact controller then replaces a static state optimum with the best attainable finite-horizon trajectory. Only their combination distinguishes continuous feasibility, reachable-lattice quantization, travel cost and the residual to a matched controller.

The separable endpoint and partial-sharing series play a similar complementary role. Separation is a clean negative endpoint but is structurally flat for IQL, so it cannot alone identify a mechanism. Partial sharing retains multi-agent interaction while changing how many coordinates receive summed actions. That manipulation also changes transition gain and, on the affected coordinates, reachable-lattice, clipping and travel geometry. Pairing latent problems and network initialization across overlap makes the slope change interpretable as modulation within this shared-coordinate transition bundle. It does not isolate physical contention from credit assignment or those dynamical changes, so "shared-coordinate modulation" is the supported claim; a unique psychological, algorithmic or dynamical mechanism is not.

7.3 An estimand hierarchy, not one headline number

The battery leaves an ordered set of evidence rather than a preferred curve selected after the fact. The late-training gap documents what the original experiment measured, including updates, exploration and its reset stream. The clean frozen gap is the primary learned-policy outcome: it holds parameters fixed and uses common unseen starts. The feasible-centred frozen gap asks whether that result persists when the signed target treatment is deliverable inside the state box. The oracle residual finally conditions on the best trajectory the same transitions and starts permit. Moving down this hierarchy removes a specific nuisance, but also changes the estimand. The correct question is not which number is "the real gap"; it is which claim each number can support.

That hierarchy explains apparent contradictions without averaging them away. Legacy late-training opposition contrasts reverse sign across stakes. Legacy frozen contrasts are all penalty-signed but mostly uncertain. Feasible-centred frozen contrasts are penalty-signed and detected in all four stake-by-learner cells. At the same time, the full-grid target-correlation slope is negative under both supports. These statements can coexist because support changes the feasible preference geometry, while freezing changes the object scored. The disagreement is evidence about construct sensitivity, not a licence to retain only the most favourable protocol.

The decomposition introduces a second hierarchy. The total gap answers how far the outcome is from the collection of individual ideals. Φ answers how much of those ideals no common continuous state can satisfy. Q and T_c add lattice and travel restrictions. P answers how far the frozen policy lies from a joint controller after those restrictions. Only the last component is even a candidate policy-performance measure, and it remains bundled with decentralization and objective mismatch. The reader-facing claim therefore stops at a "frozen-policy residual"; terms such as pure learning friction, coordination intelligence or bargaining failure would exceed the design.

This ordering provides a practical reporting rule for mixed-motive sweeps. Lead with the estimand closest to the substantive claim, but place its adjacent protocol and reference-point sensitivities nearby. If a training window and frozen policy disagree, report both. If a static optimum and dynamic oracle disagree, decompose the difference. If target support changes the delivered treatment, do not transport endpoint language across supports. That practice is more informative than announcing a single preferred

result after silently changing what was measured.

7.4 Implications for Cooperative AI

Alignment interventions are often discussed as though greater goal agreement should generally improve cooperation. Here only ideal-point correlation is manipulated, and its dividend appears precisely where agents must reconcile control of a common state. When each agent owns an independent pool, target correlation has no detected effect. That distinction matters for design: increasing target-coordinate agreement can help under physical or state contention, but the same intervention may be irrelevant where tasks factorize. It also matters for evaluation. A target sweep should verify signed coupling, and a benchmark should separate whether target agreement makes a good joint outcome attainable from whether agents learn a better policy conditional on attainability.

The result is not a claim about commons governance. The state is rival in control but is not appropriated, depleted or replenished. Nor is it architecture-general: IQL and VDN are both value-based, and the VDN arm also changes the reward objective. Policy-gradient methods, monotonic-mixing methods and standard mixed-motive benchmarks remain external-validity tests.

The feasible component is not a nuisance to be subtracted and forgotten. In systems with one jointly controlled state, making objectives more compatible can create value even if the learning algorithm is already optimal, because a better common outcome becomes possible. The policy component asks a different design question: conditional on what is possible, do decentralized agents exploit it? Under feasible-centred targets both channels improve as target correlation becomes positive, whereas under legacy support only the first is detected. Cooperative-AI interventions should state which channel they target. Preference shaping may improve achievability without improving learning; communication or planning may improve the residual without changing achievability.

The partial-sharing result offers one route to richer benchmarks. Rather than compare only a fully joint task with a factorized twin, a benchmark can parametrize how much of the state, reward or information is coupled and test whether an algorithm's target-correlation sensitivity changes with that dose. Here state overlap is coarse and also affects credit assignment. A future factorial crossing state, reward and observation coupling would separate those axes, while crossing architecture with reward information would separate the two changes currently bundled in VDN. Until then, the finding is about state-overlap modulation in this construction.

7.5 The methodological caution

The sampler failure is seductive because its code resembles a natural signed knob: $\tau_i = \alpha b + (1 - |\alpha|)\xi_i$. Yet the shared anchor's sign cancels in cross-agent correlation. Computing realized treatment before training would have caught the problem for almost no cost. The transferable check has two parts: compute the coupling actually realized, then compare it with the estimand named in the design. A sampler that is always positive or always negative is not defective unless the paper interprets it as sweeping the sign.

The two-part wording is not pedantic. An early audit of the two-team control computed its realized

mean pairwise correlation as $-\text{opp}^2/3$ and briefly flagged the design because the mapping is even in the nominal opposition-strength label. That was a false positive: the design never claimed that the nonnegative strength parameter was itself signed, and its delivered between-team correlation was exactly the treatment it named. The flag was retracted and preserved in the forensic history. A realized correlation should be compared with the declared estimand, not judged against a generic expectation that every knob must be signed.

The same discipline applies to the other choices. Algebra can reveal scale effects before ANOVA; a factorized or partially coupled twin can test whether the environment supplies the apparent target-correlation effect; and an oracle matched to transitions can prevent a geometric floor from being called learning failure. These checks do not establish how prevalent the traps are. They show, with a documented failure of our own, why experiment construction belongs inside reproducibility rather than being treated as a fixed prelude to statistics.

This paper consequently has two legitimate reading surfaces. The present article carries the argument, formal MDPs, principal estimands, conditional contrasts, oracle validation and release status. The accompanying technical supplement retains the complete factorial tables, secondary diagnostics, defect chronology, alternative outcome protocols and legacy figures under stable navigation identifiers `TS-FIRST`–`TS-RELEASE`. Compression is therefore navigational rather than evidential: material moves out of the main line, but non-detections, corrections and provenance remain citable. A later eight-page conference version would require a further genre-specific reduction and should point to both records; it cannot substitute for the full scientific account.

8 Limitations and Future Work

First, all post-diagnostic controls are exploratory. Protocol timestamps and row-level hashes prevent silent rematching, but there was no preregistration or independent replication. Corrected multiplicity controls false positives within declared families; it does not change that status. The effect sizes should therefore be read as demonstrations in this environment, not population estimates for MARL research.

The chronology is unusually important. The signed DGP, separable arm, partial-sharing dose, feasible support, frozen protocols, oracle and signed-noise crossing were designed after defects or interpretive ambiguities in the earlier sweep became visible. Dated scripts and repository history can show when a control existed and row-level hashes can prevent rematching; neither substitutes for an immutable preregistration. The named Holm families answer specific reviewer-raised questions, but selection across the sequence of possible controls remains outside those corrections. An independent rerun should therefore reproduce directions and conditional patterns before the reported p-values are treated as more than exploratory uncertainty summaries.

Second, the signed axis is geometrically and environmentally bounded. Equicorrelation with four agents stops at $-1/3$, and the experiment samples $-.33$ just inside that bound. The team design confounds between-team opposition with within-team cooperation, while the $n = 2$ arm changes agent count. Under legacy support, clipping attenuates feasible opposition more than cooperation: nominal

$-.33$ becomes about $-.217$, whereas $+.8$ becomes about $+.743$. The feasible-centred control repairs this in the tested box, but it is one translated Gaussian support, not a universal adversarial environment.

The positive-semidefinite restriction becomes tighter as the population grows: $-1/(n-1) \to 0$. Mutual pairwise opposition at -1 is possible only for two agents, while larger groups require block or otherwise non-equicorrelated structures. The team block and $n = 2$ controls therefore answer different diagnostic questions and cannot be pooled with the four-agent continuum. Likewise, centring the Gaussian at five preserves covariance and makes the sampled targets feasible, but it changes their location relative to random resets. It is the clean support repair for this box, not proof that boundary placement is irrelevant in other tasks.

Third, the total gap is not a pure learning statistic. The oracle decomposition removes the continuous, lattice and transient components from the clean held-out outcome, yet the residual still combines policy suboptimality with decentralization and, for IQL, mismatch between individual training rewards and the oracle's team objective. Calling that residual "pure learning friction" would repeat the category error the control was designed to correct.

The oracle is also centralized and optimizes mean team reward with complete knowledge of weights, targets and state. That is deliberately an upper bound, not a policy available to the decentralized learners. For VDN the training objective matches the team aggregation more closely; for IQL it does not. Comparing their residual levels would therefore mix learner competence, information, decentralization and objective alignment. The component slopes are useful because they show how the same benchmark changes with ρ, but they do not identify which of those ingredients produces the support-dependent P gradient.

Fourth, learner scope is narrow. IQL is deliberately simple; VDN is additive and trained on a team reward. Their similar shared-state shapes show that the ordering recurs under one bundled change, not that centralization is causal or that policy-gradient and monotonic-mixing methods will agree. Auxiliary MAPPO results are not headline evidence because advantage normalization is not crossed with an ablation that isolates its scale consequences. QMIX and independent policy-gradient controls remain unrun.

This omission is substantive for the stake result. Per-replication advantage normalization can remove or distort reward-scale sensitivity, so adding a policy-gradient baseline without an ablation would conflate architecture with a different scale treatment. A useful extension would cross value-based versus policy-gradient optimization, centralized versus decentralized training, individual versus team reward, and normalized versus unnormalized advantages. The present two bundles occupy only two corners of that design. Similar curve shape across them is reassuring but cannot support a main effect for centralization or objective sharing.

Reward rescaling is not exactly a learning-rate change under Adam because moment normalization and its ε term intervene. With the optimizer fixed, this is sensitivity of the training bundle, not an optimizer-invariant law or isolated step-size mechanism.

Fifth, the task family is bespoke: four agents, three continuous coordinates, bounded summed ac-

tions, quadratic losses and two stake levels in the signed battery. The partial-sharing grid has only four attainable overlaps and always shares the first k coordinate positions, so its failure to detect curvature does not prove linearity and its overlap gradient is not separated from coordinate identity. Counterbalancing or randomizing equally sized shared subsets is a necessary invariance check. Reward, observation or communication coupling without common state remains untested. Standard benchmarks, larger populations, alternative reward shapes and finer overlap grids are needed before claiming external generality.

The environment is also not a common-pool-resource model in the institutional sense. Agents jointly move a state, but they do not extract from, deplete or replenish a stock. Quadratic rewards are smooth and separable by coordinate, which makes the exact oracle possible and may make target compatibility unusually salient. Discrete bargaining, network congestion, communication games and depletable commons could shift the balance between feasibility and learned-policy residual. The control logic should transport more readily than the numeric slopes.

Sixth, no prospective power or minimum-detectable-effect calculation governed the post-diagnostic controls. The separable VDN slope, departures from a linear partial-sharing ramp, legacy opposition contrasts and $\rho \times \sigma$ interactions are non-detections, not evidence of absence or equivalence, and they were not designed to exclude small effects. Their intervals and achieved sample sizes are the appropriate descriptive sensitivity record; post-hoc power would not repair the absence of prospective design justification.

The public replication boundary is now explicit. Zenodo v3.0.0 record 10.5281/zenodo.22004215 contains the checked August source and raw outputs; its inventories and hashes identify that immutable payload. The mutable GitHub tree and any later correction must still identify their own commit, but availability of the August raw evidence is no longer a blocker. The post-v3 bootstrap delta is preserved in the arXiv ancillary provenance bundle and remains pending repository/Zenodo synchronization.

Future work should first replicate the frozen feasible-centred battery independently. It should then cross learner architecture with reward information, extend partial sharing to a finer grid, test non-state forms of coupling, and connect the exact controller to analytic equilibrium benchmarks. Those additions are more informative than merely increasing the original factorial's seed count.

9 Conclusion

This study began with an apparent confirmation and ends with a control battery. A shared-anchor sampler never delivered negative correlation and manufactured a symmetric U-shape. Degree-one reward scaling made the implemented reward coefficient look structurally dominant. A shared-state environment made a shared-state-bounded target-correlation effect look like a general preference effect. Yet this experiment varies ideal points inside a fixed quadratic reward family; it does not directly measure objective- or reward-function alignment. An out-of-box target distribution, continued-training outcome and static reference then misattributed part of a real target-correlation gradient.

The repaired result is narrower and better identified. Frozen policies retain a negative target- correlation slope on held-out resets. Feasible-centred targets steepen it, the four-agent $\rho = -.33$ endpoint

exceeds $\rho = 0$ within both reward-scale levels and learners, and the exact finite-horizon decomposition assigns roughly 54–63% of the gradient to feasibility (95% bootstrap intervals spanning 51–72%) and 37–46% to a support-dependent policy residual (28–49%). In the feasible-centred separable control, the IQL expectation is correlation-invariant and the VDN slope is not detected; across four tested legacy-Gaussian-support overlap levels the slope point estimates steepen monotonically. Observation noise raises the mean feasible-centred gap but interacts with target correlation in the opposite direction from the proposed single index. Taken together, the exploratory controls reject the implemented composite as a maintained single-predictor description in this environment; this remains distinct from a valid confirmatory H4 verdict. Its coordinates survive only as separately estimated, scope-dependent variables.

These are exploratory results from one environment family, not a prevalence estimate or architecture-general law. The contribution is the discipline they instantiate: verify treatment, remove algebraic scale, vary interaction structure, freeze evaluation and match the benchmark to the dynamics. A result that changes under those controls has not been weakened by the audit; it has finally been identified.

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`. The stable Zenodo concept DOI 10.5281/zenodo.20636467 resolves to the latest version; the exact v3.0.0 payload used here is archived at immutable record DOI 10.5281/zenodo.22004215. It contains the checked August raw outputs, prior release source boundary, inventories and hashes. The 24 August bootstrap and protocol-summary analysis delta postdates that immutable record and is not claimed to be part of v3; repository/Zenodo synchronization is a later release task.

Reproducibility note. The February CSVs predate deterministic PyTorch seeding and cannot be regenerated bit for bit. Deterministic seeding was added before the post-fix runs. GPU and CPU schedules still differ in how they index replication streams; the replication remains the inferential unit.

AI Assistance. Claude (Anthropic) was used for literature synthesis, analysis and simulation-code development, LaTeX preparation and revision. ChatGPT (OpenAI) was used for adversarial review and cross-checking. Automated review platforms supplied additional pre-submission checks. All claims, results and errors remain the author's 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.

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

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.

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.

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.

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.

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

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.

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.

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.

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.