WHEN DOES REWARD SELECT INTERACTION? IDENTIFYING AND STEERING INFORMATION PATHWAYS IN MULTI-AGENT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent policies are judged by team return, which does not show which observation channel a policy relies on. On a flocking-alignment task, independent PPO (IPPO) observes a shared reference and its neighbors and reaches polarization 0.977, yet falls to 0.339, near the 0.316 of random headings, when the reference is removed. We prove that return maximization cannot guarantee selection of the peer channel when a peer-omitting policy is optimal, and, for idealized decoupled dynamics, that a training law breaking the correspondence of the shared cue caps every peer-omitting policy at the $N^{-1/2}$ random-heading scale (a *shortcut-incompatibility gap*); a simulator test confirms both sides of this gap. We introduce the Shortcut-Free Causal Information Audit (SCIA), which reads the pathway of a frozen policy through paired, marginal-preserving interventions and is calibrated on controls with known answers; on learned policies, zero-ablation, correspondence, and action influence disagree as these controls predict. With the reward and PPO update fixed, training with independent references (shortcut-incompatible policy optimization, SI-PO) or zeroed references raises peer correspondence from 0.0036 to 0.476 and 0.600 and keeps order at 0.90 and 0.93 after reference loss, whereas a matched mixture that keeps the shared cue in two thirds of worlds fails. The shortcut replicates with 16 agents, on a speed-cue variant, and on a goal-reaching task, and zero-reference training removes it in all three.

1 INTRODUCTION

Multi-agent learners are usually judged by a scalar: team return, success rate, or a group-order statistic such as polarization. These scores are silent about *which information the deployed policy uses*. Two observation channels can support the same score: agents may track a shared reference—a goal, leader direction, or broadcast field—or couple through local peer state. Both mechanisms produce near-perfect order (Vicsek et al., 1995; Olfati-Saber, 2006). The difference (Figure 1a) is invisible during training and decisive at deployment: a policy that relies on the broadcast fails with it, whereas one that relies on its neighbors does not. We call the channel on which a policy's behavior depends its *pathway*.

Figure 1b shows this failure in a standard learner. On a flocking-alignment task, native IPPO (de Witt et al., 2020; Yu et al., 2022) observes both a shared reference and the mean state of its neighbors, and reaches polarization $P = 0.977$. When the reference is zeroed at test time, P drops to 0.339 ($[0.329, 0.348]$) and the evaluation return from 74.6 to -21.1; order falls close to the 0.316 expected for independent uniform headings, although the peer channel is still observed. The learner had the information needed for coordination and did not use it.

This is a multi-agent form of shortcut learning (Geirhos et al., 2020; Shah et al., 2020) and causal confusion (de Haan et al., 2019): high return is predictable from a cue that is cheaper than the intended mechanism. Social learners likewise ignore informative peers when individual cues suffice (Ndousse et al., 2021). We ask two questions. When can return maximization select a peer pathway at all? And can one identify and change the pathway a policy occupies without changing its reward?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

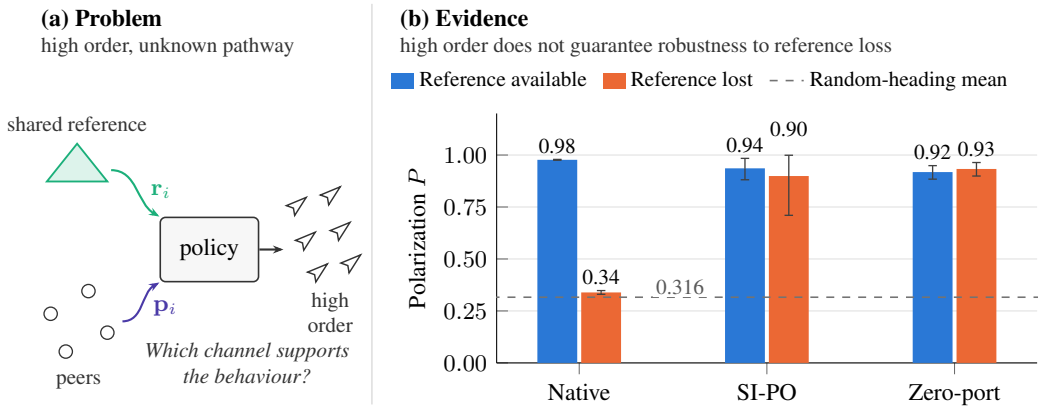

Figure 1: Motivation. (a) High collective order does not reveal whether a policy relies on a shared reference or its peers. (b) Native IPPO loses order when the reference is removed, despite continued access to peers; SI-PO and zero-port retain high mean order. Error bars show 95% cluster-bootstrap confidence intervals over training seeds (ten per recipe).

Contributions. Our contributions answer these two questions in turn. For the first, return alone cannot guarantee peer use: when a channel-omitting policy is optimal, return maximization need not select the peer pathway (Proposition 2). Conversely, for idealized decoupled dynamics, once training breaks the correspondence of the shared cue, every peer-omitting policy is capped at the $N^{-1/2}$ random-heading RMS scale (Proposition 3). A simulator test that satisfies its assumptions confirms both sides of the gap: the pooled second moment of 35 policies with the peer input zeroed lies within 3% of $1/N$, and all 20 trained peer-using policies exceed the cap.

For the second, we introduce SCIA, an audit that reads the pathway of a frozen policy through paired interventions on each channel, with frame and memory controls; it applies unchanged to NEAT and IPPO. Zero-ablation, outcome correspondence, and action influence answer different questions, and on learned policies they disagree in the ways that controls with known answers predict (Figure 3): zero-ablation underestimates channel use by factors of 2.3 and 3.4 in neuroevolved (NEAT) policies that use both channels as substitutes and cannot detect SI-PO's reference use, whereas action influence flags peer responses of native IPPO that do not affect order.

We then steer the pathway by changing only the training distribution, using two instances of the condition in Proposition 3: at an equal world-frame budget, shortcut-incompatible policy optimization with independent references (SI-PO) and zero-reference training raise IPPO's peer correspondence from 0.0036 to 0.476 and 0.600 and retain order after reference loss, while native training and both mixtures do not; SI-PO also raises the mean deployment return over native. The shortcut replicates with 16 agents, on a speed-cue variant, and on a goal-reaching task, and zero-reference training removes it in all three (Sections 4.3–4.5); keeping the reference port brings no advantage in the last two.

Related work. *Shortcut learning.* Shortcut learning and simplicity bias (Geirhos et al., 2020; Shah et al., 2020), causal confusion in imitation learning (de Haan et al., 2019), and social learners that ignore informative peers when individual cues suffice (Ndousse et al., 2021) show that a cheaper cue can displace the intended one. Invariant risk minimization (Arjovsky et al., 2019) changes the objective to learn predictors that remain optimal across environments; we keep the reward, break the correspondence of one named observation channel, and audit the resulting dependence. *Interventional influence.* The causal influence of communication (CIC) measures a message's effect on listener actions (Lowe et al., 2019); Peters et al. (2025) survey emergent communication. Agent-specific effects quantify how an action affects outcomes through other agents (Triantafyllou et al., 2024), and later work decomposes counterfactual effects through actions and state transitions (Triantafyllou et al., 2025). Permutation importance and model reliance (Breiman, 2001; Fisher et al., 2019) permute one input to measure dependence, and interchange interventions (Geiger et al., 2021) swap internal representations. SCIA permutes observation channels, not messages, in a closed multi-agent

loop with paired worlds and frame and memory controls, and separates correspondence from availability; it does not estimate message effects or action-effect decompositions. Its outcome contrasts complement action influence, which can persist with little change in order (Figure 3c). *Training interventions.* Social-influence rewards encourage agents to affect others' actions (Jaques et al., 2019), and Ndousse et al. (2021) induce social learning with environment constraints and an auxiliary loss; both change the objective, whereas we keep the reward fixed. Domain randomization varies simulation conditions for transfer (Tobin et al., 2017); SI-PO instead breaks one named channel's correspondence, motivated by Proposition 3, while zero-port training removes that channel. *Flocking and MARL.* Vicsek alignment obtains order from local interaction (Vicsek et al., 1995); flocking and consensus rules with navigational feedback or a leader also obtain it from a shared direction (Olfati-Saber, 2006; Jadbabaie et al., 2003). A learned policy may occupy either pathway, and high polarization alone does not reveal which. IPPO (de Witt et al., 2020; Yu et al., 2022) optimizes return without distinguishing channels.

2 When does return select a pathway?

Each of N agents executes a shared policy $(\mathbf{a}_i^t, h_i^t) = f(\mathbf{o}_i^t, h_i^{t-1})$ with private recurrent state h_i when the architecture is recurrent. The observation is partitioned as $\mathbf{o}_i = (\mathbf{s}_i, \mathbf{r}_i, \mathbf{p}_i)$: self kinematics $\mathbf{s}_i$, a reference channel $\mathbf{r}_i$, and a local peer statistic $\mathbf{p}_i$ (mean relative neighbor position and velocity). A team return U and a training distribution $\mathcal{D}$ over initial states and reference laws induce $J_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[U]$. Any closed-loop scalar S of the physical trajectory—polarization, remaining distance, or U itself—defines a pathway profile (Section 3).

Definition 1 (Omitting and admitting classes)**.** Let $\mathcal{F}$ be a class of shared policies for which both channels appear in $\mathbf{o}_i$. The *omitting subclass* $\mathcal{F}_{\setminus c}$ consists of those $f \in \mathcal{F}$ whose action and memory maps are invariant to channel c on the closed-loop support, including every intervened cell used in the audit. The *admitting class* is $\mathcal{F}$ itself: the port for c is present, but incoming weights may be zero.

Proposition 1 (Omission invariance)**.** *If $f \in \mathcal{F}_{\setminus c}$, every paired contrast that changes only the correspondence of c is zero for every trajectory functional S; in particular, the correspondence effect of Section 3 on that channel vanishes: $\Delta_P^{\mathrm{corr}}(f) = 0$ if $c = p$, and $\Delta_R^{\mathrm{corr}}(f) = 0$ if $c = r$.*

Opening a port therefore only enlarges the representable class. Whether the port is *selected* depends on the triple $(\mathcal{F}, U, \mathcal{D})$.

Proposition 2 (Return need not select an admitted channel)**.** *If some $f^\star \in \mathcal{F}_{\setminus p}$ satisfies $J_{\mathcal{D}}(f^\star) = \sup_{f \in \mathcal{F}} J_{\mathcal{D}}(f)$, then $\arg\max J_{\mathcal{D}}$ intersects $\mathcal{F}_{\setminus p}$, every omitting maximizer has $\Delta_P^{\mathrm{corr}} = 0$, and $J_{\mathcal{D}}$ ranks no peer-using policy above $f^\star$.*

The hypothesis is a *sufficient-cue condition*: some map of $(\mathbf{s}_i, \mathbf{r}_i)$ already attains the optimum. It holds for alignment: if every agent follows a shared unit direction $\widehat{\mathbf{g}}$ encoded in $\mathbf{r}_i$, then $P = 1$ while every peer intervention has zero effect. It also holds beyond consensus scores: in shared-target navigation with a remaining-distance return, a greedy target tracker that reads only $\mathbf{r}_i$ is optimal at every step; Section 4.5 trains policies on a simulator version of this task.

Removing the shortcut from training. For a fixed return, the obstruction can be removed through $\mathcal{D}$. Proposition 3 is stated for an idealized alignment model under the following assumptions; Remark 1 states which of them the simulator satisfies.

- (A1) *Local observation and action interface.* Agent i has a frame angle θ_i; observations are expressed in that frame, the self channel is the local velocity $\mathbf{s}_i^t = R(\theta_i)^\top \mathbf{v}_i^t$ only, the local action is applied in the world as $R(\theta_i)\mathbf{a}_i^t$, and the initial memory h_i^0 is a fixed constant.
- (A2) *Decoupled, rotation-equivariant dynamics.* Given the world action $\mathbf{b}_i^t = R(\theta_i)\mathbf{a}_i^t$, each velocity is drawn as $\mathbf{v}_i^{t+1} \sim K(\cdot \mid \mathbf{v}_i^t, \mathbf{b}_i^t)$, independently across agents and of positions (no contact forces), with $K(R\cdot \mid R\mathbf{v}, R\mathbf{b}) = K(\cdot \mid \mathbf{v}, \mathbf{b})$ for every rotation R. Position constraints that leave velocities unchanged, such as a clamp at the arena boundary, are allowed.
- (A3) *Independent initialization and randomization.* The frames $\theta_i \sim \mathcal{U}[-\pi, \pi)$, the initial velocities $\mathbf{v}_i^0$ (rotation-invariant in law), and the per-agent policy noise $\xi_i = (\xi_i^t)_t$ are mutually independent and independent across agents.

(A4) *Polarization-only objective.* $U = \bar{P} = (T - T_w)^{-1} \sum_{t=T_w+1}^{T} P^t$ with $P^t = \|N^{-1} \sum_i \widehat{\mathbf{v}}_i^t\|$ and $\widehat{\mathbf{0}} = \mathbf{0}$.

(A1)–(A3) rule out world-frame cues, dynamical coupling, and common randomness as sources of coordination. Under the *native* law the reference is one shared direction $\mathbf{u}$, observed as $\mathbf{r}_i = R(\theta_i)^\top \mathbf{u}$. A *shortcut-incompatible* law replaces it by directions $\mathbf{u}_i$ drawn once per episode, i.i.d. uniform and independent of the variables in (A3) (independent reference), or by $\mathbf{0}$ (zero reference). Omission (Definition 1) is taken on the support of the law in question.

Proposition 3 (Shortcut-incompatibility gap)**.** *Assume (A1)–(A4) and a shortcut-incompatible law $\mathcal{D}_I$. For every $f \in \mathcal{F}_{\setminus p}$ and every t, the world headings $\widehat{\mathbf{v}}_1^t, \ldots, \widehat{\mathbf{v}}_N^t$ are independent, have mean zero, and are uniform on the circle conditional on $\mathbf{v}_i^t \neq \mathbf{0}$. Hence*

$$\mathbb{E}\big[(P^t)^2\big] \leq \frac{1}{N}, \qquad \sup_{f \in \mathcal{F}_{\setminus p}} J_{\mathcal{D}_I}(f) \leq N^{-1/2}.$$

If $\mathcal{F}$ contains a policy f_c with $J_{\mathcal{D}_I}(f_c) \geq 1 - \varepsilon$, then $\sup_{\mathcal{F}} J_{\mathcal{D}_I} - \sup_{\mathcal{F}_{\setminus p}} J_{\mathcal{D}_I} \geq 1 - \varepsilon - N^{-1/2}$.

The proof shows that an omitting agent's local trajectory is independent of its frame and of every other agent, so its world heading is a uniformly rotated, independent local heading. For the hypothesis on f_c: in the collision-free simulator setting of Section 4.3, which satisfies (A1)–(A3), all 20 trained zero-port and SI-PO policies reach $\mathbb{E}[P] \geq 0.596$, with recipe means up to 0.904 under the independent and 0.972 under the zero law. Under the native law no such gap exists: the omitting reference tracker above already reaches perfect order.

Proposition 3 identifies which training laws remove the shortcut (Table 1) and predicts a value the simulator can check. It concerns the objective: under $\mathcal{D}_I$, high order requires the peer pathway, and Section 4 tests whether PPO finds it. Two consequences follow. (i) Independent and zero references both satisfy the hypothesis, so port preservation is not required by the objective; it matters when the reference also carries information the task needs. (ii) For a mixture that uses the zero-reference law in a fraction α of worlds and a shared reference otherwise, the gap is at least $\alpha\,(J_0(f_c) - N^{-1/2})$, where J_0 is the return under the zero law, provided $\mathcal{F}$ is closed under switching on whether the reference is zero. An independent reference admits no such switch, because an agent cannot distinguish it locally from a shared one.

Remark 1 (Simulator and metric). The simulator satisfies (A1), (A3), and the equivariance in (A2): VMAS integrates holonomic forces with scalar drag and a norm-based speed limit, the local action is rotated before any world-frame operation, and the arena boundary clamps positions without altering velocities. It departs from (A2) through agent–agent collisions, and from (A4) through the composite reward and a polarization averaged over the M agents faster than 0.05; under (A1)–(A3), for $f \in \mathcal{F}_{\setminus p}$ and $\mathcal{D}_I$, this metric satisfies $\mathbb{E}[P_M^2 \mid M = m] = 1/m$, so the $N^{-1/2}$ scale applies when all agents move. Disabling collisions removes the only departure from (A1)–(A3), which Section 4.3 uses to test the bound directly.

3 Auditing and steering pathways

The method has two parts: an audit that reads which channel a frozen policy uses (Section 3.1), and a training intervention that removes a known shortcut without changing the reward (Section 3.2).

3.1 Auditing which channel a frozen policy uses

SCIA takes a frozen policy, the channels to test (here $\mathbf{r}_i$ and $\mathbf{p}_i$), and a closed-loop score S such as polarization, and returns a channel profile in three steps.

Step 1: frame and memory controls. Each agent draws an episode-wise frame $\theta_i \sim \mathcal{U}[-\pi, \pi)$; observations are rotated into it and actions back, so a constant local command has only the random-heading RMS polarization $N^{-1/2}$ and no absolute direction can leak through the action parameterization. Recurrent state is indexed by (policy, world, agent), so evaluation order cannot copy memory between agents.

Step 2: paired interventions. Each cell of a 3×3 grid (Figure 2) intervenes on the observation channels of the frozen policy (Pearl, 2009) and reuses the same initial states, frames, and candidate references. The peer mode is observed, shuffled, or zero ($p \in \{o, h, 0\}$); the reference mode is shared, independent, or zero ($r \in \{s, i, 0\}$). Shuffling cyclically permutes peer inputs across a policy's simultaneous worlds, leaving their empirical distribution over the batch unchanged at each step; independent directions are drawn per agent from the uniform law of the shared one, leaving each agent's reference marginal unchanged. Both break correspondence; zeroing tests availability. Neither fixes the closed-loop trajectory distribution, so contrasts include subsequent closed-loop effects; fidelity checks confirm matching input moments (Sections 4.2–4.3).

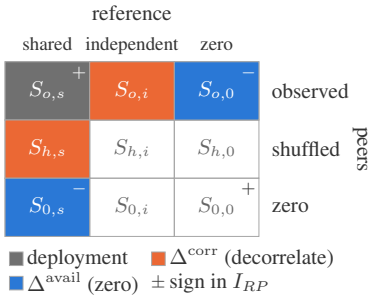

Figure 2: The SCIA grid; cells hold scores $S_{p,r}$. An orange or blue cell's effect is $S_{o,s}$ minus its score (top row: R; left column: P). Signs give I_{RP}.

Step 3: channel profile. With $S_{p,r}$ the score in cell (p, r), the profile is

$$\Delta_R^{\text{corr}} = S_{o,s} - S_{o,i}, \quad \Delta_R^{\text{avail}} = S_{o,s} - S_{o,0},$$
$$\Delta_P^{\text{corr}} = S_{o,s} - S_{h,s}, \quad \Delta_P^{\text{avail}} = S_{o,s} - S_{0,s}, \tag{1}$$
$$I_{RP} = S_{o,s} - S_{o,0} - S_{0,s} + S_{0,0}. \tag{2}$$

I_{RP} is the interaction of the 2×2 zeroing design; the three remaining cells are evaluated but enter no profile component. A positive correspondence effect excludes membership in the omitting class (Proposition 1); a small one does not establish it, so we also report the channel-loss cell $S_{o,0}$. A negative I_{RP} with high scores in both single-channel cells ($S_{o,0}$ and $S_{0,s}$) indicates substitution. To catch use that changes actions but not S, we add a one-step action check,

$$A_c = \mathbb{E}\left\|\pi(\mathbf{o}_i) - \pi(\mathbf{o}_i^{c\leftarrow c'})\right\|, \tag{3}$$

where π is the deterministic action map of the frozen policy f and $\mathbf{o}_i^{c\leftarrow c'}$ replaces channel c by the same channel of another world, on states visited in the deployment cell. Estimates average worlds within a trained policy; 95% intervals resample per-seed means (cluster bootstrap over training seeds).

3.2 Training without the shortcut

Proposition 3 suggests training under a law $\mathcal{D}_I$ that breaks the correspondence of the known sufficient cue while keeping the reward formula and optimizer:

$$J_I(\theta) = \mathbb{E}_{\mathcal{D}_I}[U(f_\theta)]. \tag{4}$$

Definition 2 (SI-PO)**.** Shortcut-incompatible policy optimization trains with PPO under a law $\mathcal{D}_I$ that is (i) *port-preserving*, keeping the reference input and its marginal; (ii) *correspondence-breaking*, replacing the shared direction with independent per-agent directions; (iii) *unmixed*, never restoring the shared reference in training; and (iv) *audit-external*, leaving correspondence scores out of the reward. It adds no new update rule.

Choosing the intervention. Zero-reference training (zero-port) and SI-PO both train under shortcut-incompatible laws, the hypothesis of Proposition 3. They differ in what else they remove: zeroing deletes the whole channel, whereas breaking correspondence keeps the channel's marginal and any content not tied to the shared direction, such as its length. Zeroing suffices when the channel is a pure shortcut. When the channel also carries information the task requires, keeping it is necessary for a policy to use that information, but not sufficient: Section 4.4 finds that SI-PO does not use a retained target-speed code. Table 1 lists the recipes; every frozen policy is evaluated on the same grid, including the shared-reference cell that SI-PO never sees in training.

Table 1: Equal-budget IPPO recipes, all evaluated on the same 3×3 grid. The last column is the lower bound on the gap of Proposition 3 under (A1)–(A4); $J_I(f_c)$ and $J_0(f_c)$ are the returns of a peer-consensus policy under the independent- and zero-reference laws, so the bounds tend to $1-N^{-1/2}$ and, for the two mixtures, $\frac{1}{3}(1-N^{-1/2})$ as these returns tend to one. † Assumes closure under switching on whether the reference is zero. Native: an omitting reference tracker is optimal, so no gap arises.

Recipe	Peer	Reference	Violated constraint	Gap lower bound
Native	observed	shared	(ii), (iii)	none
Balanced	mix	mix	(i)–(iii)	$\frac{1}{3}\big(J_0(f_c)-N^{-1/2}\big)^{\dagger}$
Matched mixture	observed	$\frac{2}{3}$ shared, $\frac{1}{3}$ zero	(i)–(iii)	$\frac{1}{3}\big(J_0(f_c)-N^{-1/2}\big)^{\dagger}$
Zero-port	observed	zero	(i)	$J_0(f_c)-N^{-1/2}$
SI-PO	observed	independent	—	$J_I(f_c)-N^{-1/2}$

4 Experiments

4.1 Setup

The testbed instantiates the alignment structure in VMAS (Bettini et al., 2024): a goal-free $N{=}8$ task with a fixed 22-dimensional observation. Slots $2{:}4$ hold local self velocity, $4{:}6$ the reference vector, and $18{:}22$ the mean relative neighbor position and velocity; absolute-position ($0{:}2$) and lidar ($6{:}18$) slots stay zero. The reward combines polarization, speed, cohesion, collision, and boundary terms.

NEAT. Zero-hidden recurrent NEAT policies (Stanley & Miikkulainen, 2002) are trained under four information regimes (reference-only, peer-only, both, none) with ten seeds, population 50, 30 generations, 32 training worlds, and 200 steps; feed-forward NEAT uses the same per-seed budget with five seeds. *IPPO.* A shared actor and critic (two-layer $\tanh$ MLPs of width 64; diagonal Gaussian action squashed by $\tanh$) are trained with GAE (Schulman et al., 2016) and clipped PPO (Schulman et al., 2017) with a KL stop, using ten training seeds for native, zero-port, and SI-PO and five for balanced. Native IPPO trains on observed/shared worlds; balanced IPPO mixes observed/shared, observed/zero, and zero/shared worlds in equal numbers; zero-port trains on observed/zero; SI-PO on observed/independent; a matched mixture (five seeds) trains on observed/shared and observed/zero worlds in a 2:1 ratio. Each recipe uses 9.6M world frames per seed, 960 parallel worlds, horizon 200, learning rate 3×10^{-4}, clip 0.2, $\lambda = 0.95$, two epochs, and KL stop 0.03. Balanced IPPO additionally normalizes advantages per training mode and selects checkpoints by the minimum mode return; the others, including the matched mixture, use global normalization and mean training return.

Evaluation. Frozen policies are evaluated on all nine cells with four held-out seeds and 64 worlds per seed after a 50-step warm-up, with $10{,}000$ cluster-bootstrap draws (Efron & Tibshirani, 1993) over training seeds. Gates for SI-PO and zero-port were recorded in their protocol files before their numbers were computed. Control gates require $P \geq 0.75$ in informative matched cells, losses of at least 0.20 under both destructive interventions with positive CI lower bounds, and $P \leq 0.55$ (CI upper bound ≤ 0.65) in the no-information cell. The material-effect gate requires $\Delta_P^{\mathrm{corr}} \geq 0.10$ with a positive CI lower bound. For SI-PO and zero-port, additional gates require CI lower bounds of 0.75 on the observed/shared deployment cell and 0.60 on the observed/zero reference-loss cell. These gates were set for an initial five-seed run (seeds 11–71); seeds 89–127 were added afterwards with an identical protocol, and we report the gate decisions for both.

4.2 Validating the audit

Controls with known ground truth. We add a finite diagnostic relay task with a fixed informed sender, two receivers, and seven hand-specified policies. A reference and a peer message both encode a binary target direction. Receivers act for two steps with private local frames and memory; the score S is final-step direction accuracy and ignores speed. Enumerating all 128 equally likely latent worlds in each 3×3 cell gives $8{,}064$ policy-world-cell records.

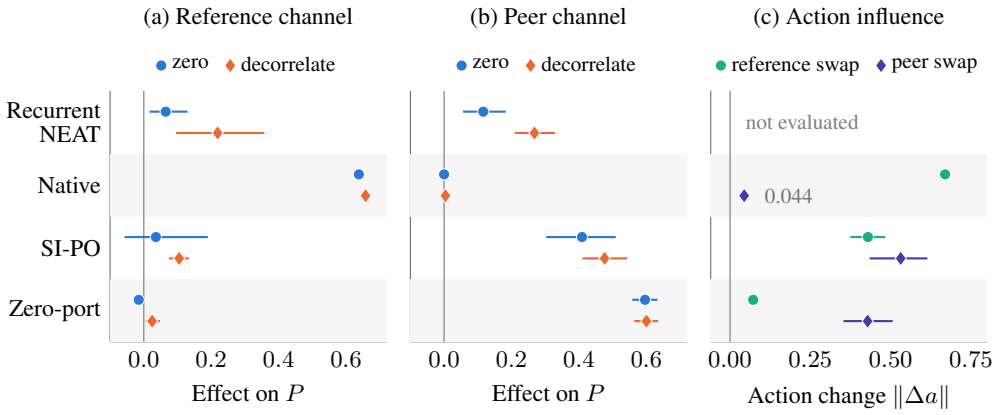

Figure 3: Three readings of frozen policies. (a,b) Closed-loop effects on polarization of zeroing (availability) or decorrelating (correspondence) each channel, for recurrent NEAT and IPPO: zeroing underestimates NEAT's use of both channels, and for SI-PO's reference its interval includes zero while decorrelation's does not. (c) One-step action change when a channel is swapped with another world's (IPPO only; own scale): native IPPO's actions respond to peers (0.044, $[0.037, 0.049]$) although its peer effect on P in (b) is near zero. Lines: 95% cluster-bootstrap intervals.

These controls expose three inference errors. The reference-only control has one bit of observational peer/action mutual information but no peer influence. The redundant control loses no score when either channel is zeroed, yet peer decorrelation costs 0.25 ($I_{RP} = -0.50$): zero-only ablation misses a backup path. The presence gate has availability effect 0.50 but zero correspondence: availability is not evidence that peer content supports the score. The same controls expose the audit's limits: the score-blind policy uses peer sign only for speed, which every score contrast misses but action-level influence detects, and the delayed policy has zero first-action but positive two-step influence, at thresholds 0.05, 0.10, and 0.20.

In the simulator, SCIA recovers four scripted controls encoding reference-only, peer-only, redundant-OR, and dual-channel-AND use: irrelevant-channel effects are zero and $I_{RP} \in \{-1.000, 1.000\}$ matches OR versus AND, with matched-cell collision CIs below 0.05.

Learned NEAT policies. In the single-channel families, with peers held at zero, reference-only policies reach $P = 0.991$ ($[0.989, 0.993]$); independent directions reduce P by 0.673 ($[0.671, 0.675]$) and zeroing by 0.734 ($[0.656, 0.811]$), so order depends on a *shared* direction, not a nonzero input. Peer-only policies reach $P = 0.955$ ($[0.953, 0.958]$); cyclic peer shuffling lowers P by 0.637 ($[0.632, 0.642]$) and zeroing by 0.631 ($[0.625, 0.637]$). Peer-input means and second moments and reference norms differ across cells by less than 6×10^{-8}, so the contrast is correspondence, not magnitude. The no-information cell stays at $P = 0.324$ ($[0.318, 0.333]$), close to the random-heading mean 0.316 and below the RMS scale $1/\sqrt{8} \approx 0.354$.

With both channels available, recurrent NEAT reaches $P = 0.959$ ($[0.945, 0.972]$) with $\Delta_R^{\text{corr}} = 0.219$ ($[0.096, 0.358]$) and $\Delta_P^{\text{corr}} = 0.268$ ($[0.211, 0.329]$); availability effects are smaller (0.065, 0.116). The interaction $I_{RP} = -0.454$ ($[-0.513, -0.390]$) shows that either channel alone nearly preserves order: evolution retains both pathways as substitutes. A zero-only audit would read this policy as weakly dependent on either channel; decorrelation reveals effects 3.4 and 2.3 times larger (Figure 3a,b), the pattern of the hand-built redundant control. Zeroing recurrent paths changes P by under 0.10.

4.3 Moving IPPO from the reference to its peers

Native and mixture training. All families (Table 2) pass the intervention-fidelity checks. Native IPPO reaches the highest deployment order ($P = 0.977$) with peer correspondence 0.0036 and reference correspondence 0.658: shuffling peers leaves $P = 0.973$, and losing the reference drops it to 0.339 (Figure 1b) in every one of ten seeds (per-seed range 0.311–0.360). This is the situation

Table 2: Pathway profiles of frozen policy families. $S_{o,0}$ is polarization after reference loss. Intervals are 95% cluster bootstrap over training seeds; recurrent NEAT and native, zero-port, and SI-PO IPPO use ten training seeds; feed-forward NEAT, balanced, and matched-mixture IPPO use five.

Family	Δ_P^{corr}	Δ_R^{corr}	I_{RP}	$S_{o,0}$
Recurrent NEAT (both)	$0.268\ [0.211, 0.329]$	$0.219\ [0.096, 0.358]$	-0.454	$0.893\ [0.830, 0.941]$
Feed-forward NEAT	$0.350\ [0.290, 0.411]$	0.104	-0.438	$0.968\ [0.959, 0.978]$
IPPO, native	$0.0036\ [0.0029, 0.0043]$	0.658	-0.027	$0.339\ [0.329, 0.348]$
IPPO, balanced	$0.0058\ [0.0017, 0.0118]$	$0.535\ [0.267, 0.672]$	-0.073	$0.141\ [0.000, 0.424]$
IPPO, matched mixture	$0.020\ [0.018, 0.021]$	0.652	-0.046	$0.369\ [0.354, 0.385]$
IPPO, zero-port	$0.600\ [0.563, 0.634]$	$0.025\ [0.006, 0.048]$	-0.025	$0.933\ [0.899, 0.964]$
IPPO, SI-PO	$0.476\ [0.410, 0.542]$	$0.105\ [0.075, 0.135]$	-0.206	$0.899\ [0.711, 0.999]$

of Proposition 2: the shared cue suffices, and PPO takes it. Its actions still respond to peers ($A_P = 0.044$, $[0.037, 0.049]$, against $A_R = 0.669$; Figure 3c), a use that leaves order unchanged. The balanced mixture, which zeroes the reference in one third of training worlds, also stays below the material-effect gate (0.0058) and reaches only 0.141 after reference loss. That recipe also changes peer availability, normalization, and checkpoint selection; a matched mixture that differs from SI-PO and zero-port only in its reference law fails as well: peer correspondence 0.020 ($[0.018, 0.021]$) and order 0.37 after reference loss. During training its return in zero-reference worlds fell from about 15 to about -55 while the shared-reference return rose, so PPO did not pursue the weighted gap of Table 1.

Zero-port and SI-PO training. Zero-port IPPO yields $\Delta_P^{\mathrm{corr}} = 0.600$ ($[0.563, 0.634]$) and $\Delta_R^{\mathrm{corr}} = 0.025$. On the held-out deployment cell it reaches $P = 0.918$ ($[0.884, 0.949]$), which shuffling peers drops to 0.318; the policy does not revert to the reference when that port returns, and it keeps $P = 0.933$ after reference loss. SI-PO yields $\Delta_P^{\mathrm{corr}} = 0.476$ ($[0.410, 0.542]$) with a small residual reference effect ($\Delta_R^{\mathrm{corr}} = 0.105$) and deployment order $P = 0.936$ ($[0.881, 0.984]$), passing the material-effect and deployment gates. Both recipes thus move the same reward and optimizer from the reference to the peer pathway, as Proposition 3 predicts. On the deployment cell the two recipes differ in return. SI-PO's mean evaluation return exceeds native's (111.6 versus 74.6; difference $+37.0$, 95% CI $[4.4, 65.0]$), although two of its ten seeds fall below every native seed; zero-port's is lower in point estimate (65.7; difference -8.8, $[-21.6, 3.9]$), so a return cost cannot be excluded. After reference loss both far exceed native (96.9 and 74.4 versus -21.1). At $N = 16$, under gates fixed before training (five seeds each), native again takes the shortcut (peer correspondence 0.018; $P = 0.246$ after reference loss) and zero-port again uses peers (0.707; $P = 0.951$ after reference loss, 0.246 with peers shuffled).

Direct test of Proposition 3. Zeroing the peer input places each frozen policy in $\mathcal{F}_{\setminus p}$, and disabling collisions makes the simulator satisfy (A1)–(A3) (Remark 1). Across the 35 IPPO policies, $\mathbb{E}[P^2]$ is 0.1259 ($[0.1240, 0.1278]$) under independent and 0.1260 ($[0.1241, 0.1279]$) under zero references; both intervals contain the predicted $1/N = 0.125$ and lie within 3% of it, and no policy exceeds $\mathbb{E}[P] = 0.334 < N^{-1/2}$. Re-enabling collisions changes $\mathbb{E}[P^2]$ by at most 0.003. With peers observed in the same setting, all 20 trained zero-port and SI-PO policies exceed this cap (mean $\mathbb{E}[P]$ 0.81–0.97).

Seed-level variation. In the initial five seeds, seed 23 fell to $P = 0.111$ after reference loss, so the reference-loss CI ($[0.466, 1.000]$) missed the pre-set 0.60 bound; five seeds added afterwards all exceed 0.88 and the ten-seed CI ($[0.711, 0.999]$) clears it, a post hoc result. Seed 23 is the only SI-PO policy that collapses without the reference (availability effect 0.677, at most 0.001 elsewhere), yet all ten use it: reference correspondence is positive in every seed ($\Delta_R^{\mathrm{corr}} = 0.105$, $[0.075, 0.135]$) while zeroing barely changes order ($\Delta_R^{\mathrm{avail}} = 0.036$, $[-0.057, 0.190]$; Figure 3a), a use a zero-only audit would miss. With $I_{RP} = -0.206$ ($[-0.346, -0.055]$), most SI-PO policies use the reference when present but do not need it; both effects are much weaker in zero-port ($\Delta_R^{\mathrm{corr}} = 0.025$, $I_{RP} = -0.025$). The same split appears in actions: SI-PO responds to both channels ($A_R = 0.429$, $A_P = 0.531$), zero-port mainly to peers ($A_R = 0.072$, $A_P = 0.428$). Zero-port is the cleaner recipe for peer use here; SI-PO attains the higher mean return.

4.4 A SPEED-CUE VARIANT: RETAINING INFORMATION VERSUS USING IT

We test whether retaining the reference port helps when it carries useful information as well as a directional shortcut. The reference length now encodes a per-world target speed $v^\star \sim \mathcal{U}[0.2, 0.5]$ as $v^\star/0.5$. The speed reward becomes the mean of $\max\{0, 1 - |\|\mathbf{v}_i\| - v^\star|/v^\star\}$ across agents; other reward terms are unchanged. SI-PO randomizes reference directions while retaining this speed code; zero-port removes both. Each recipe uses five seeds; hypotheses and gates for shortcut replication, peer use, and a speed-matching benefit were fixed before training.

The shortcut and peer-pathway findings replicate. Native IPPO reaches $P = 0.940$, yet its peer correspondence is only 0.013 ($[0.012, 0.014]$), and order falls to 0.335 ($[0.328, 0.344]$) after reference loss. SI-PO and zero-port instead have peer correspondence 0.420 ($[0.353, 0.512]$) and 0.384 ($[0.310, 0.506]$), retaining order 0.76 and 0.69 after reference loss.

The predicted benefit of retaining the port is not established. On the observed/shared deployment cell, SI-PO's speed-match advantage over zero-port is $+0.165$ ($[-0.020, 0.317]$); the interval crosses zero and fails the pre-set positive-lower-bound criterion. Best checkpoints occur at updates 34–49 of 50, motivating a post hoc check at twice the training budget. Both recipes then attain high order (deployment 0.90 and 0.89, after reference loss 0.92 and 0.88) and peer correspondence near 0.58, but the estimated speed-match difference is only $+0.005$ ($[-0.040, 0.053]$). Neither budget establishes a port-preservation advantage; the interval also does not establish equivalence.

Across worlds, speed correlates with $v^\star$ at 0.85 for native IPPO and 0.02 for SI-PO at 19.2M frames: SI-PO retains the speed code, but its speed barely tracks it. Its speed score nonetheless matches native's (0.646 and 0.641 for SI-PO and zero-port, 0.636 for native).

Across the twelve alignment profiles of Sections 4.3 and 4.4, the five with peer correspondence below 0.02 fall to at most 0.37 after reference loss, and the seven above 0.38 keep at least 0.69: the correspondence effect measured with the reference present separates the policies that remain ordered without it.

4.5 A GOAL-REACHING TASK

The last task simulates the shared-target navigation of Section 2 (target at distance 3; a goal-proximity term replaces alignment and speed); only informed agents see the target direction. Native IPPO trains with all agents informed, the peer recipes with two of eight informed and the rest given zero (zero-port analogue) or random directions (SI-PO analogue), 19.2M frames per seed, with gates fixed before training; at deployment two agents are informed. Native's goal score falls from 0.48 to 0.11 (drop 0.37, $[0.34, 0.39]$; peer correspondence 0.030). With all agents informed, the zero-port analogue scores 0.80 ($[0.79, 0.81]$), above native. The zero-port analogue has peer correspondence 0.26 ($[0.23, 0.30]$) and a goal score 0.15 ($[0.13, 0.17]$) above native; its uninformed agents score 0.18, against 0.02 for native and -0.17 with peers shuffled, so peers are partly followed, and its informed agents score 0.40, against 0.35. The SI-PO analogue fails the gate (-0.016, $[-0.034, 0.001]$ against native): its uninformed agents score 0.07, but its informed agents only 0.13 (native 0.35). A scripted greedy tracker reaches 0.93 with all agents informed (return 143, against 69 for native); the comparisons above are between recipes at equal budget.

5 DISCUSSION AND CONCLUSION

Reward performance and pathway use are different claims, and three lessons transfer. First, zero-ablation misreads pathways predictably, missing substitutes and residual uses that decorrelation reveals. Second, for a fixed reward, whether a pathway can be selected depends on the training distribution (Propositions 2 and 3), across three tasks. Third, interaction claims become checkable: name the intended pathway, break the correspondence of cues that make omitting it optimal, and audit the frozen policy, including the channel-loss cell.

Limitations. Proposition 3 assumes decoupled dynamics, a polarization-only objective, and an available peer-consensus policy, and does not guarantee optimizer convergence. The audit and both recipes require the tested channel to be named in advance.

REPRODUCIBILITY STATEMENT

Experimental settings are specified in Section 4.1, and proofs are provided in Appendix A. Protocols were fixed before their corresponding runs but were not externally preregistered.

AI USE STATEMENT

Generative AI tools were used to search for literature and to polish the abstract and conclusion. The related-work section was written by the authors, who take responsibility for the final content of the text.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. VMAS: A vectorized multi-agent simulator for collective robot learning. In *Distributed Autonomous Robotic Systems*, volume 28 of *Springer Proceedings in Advanced Robotics*, pp. 42–56. Springer, 2024. doi: 10.1007/978-3-031-51497-5_4. URL https://link.springer.com/chapter/10.1007/978-3-031-51497-5_4.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/947018640bf36a2bb609d3557a285329-Abstract.html.

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. URL https://arxiv.org/abs/2011.09533.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. URL https://www.nature.com/articles/s42256-020-00257-z.

Ali Jadbabaie, Jie Lin, and A. Stephen Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003. URL https://web.mit.edu/~jadbabai/www/papers/JadLiMorseTAC_June03.pdf.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 3040–3049, 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 693–701, 2019. URL https://ifaamas.org/Proceedings/aamas2019/pdfs/p693.pdf.

Kamal Ndousse, Douglas Eck, Sergey Levine, and Natasha Jaques. Emergent social learning via multi-agent reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 7991–8004, 2021. URL https://proceedings.mlr.press/v139/ndousse21a.html.

Reza Olfati-Saber. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on Automatic Control*, 51(3):401–420, 2006. URL https://hal.elte.hu/~vicsek/downloads/papers/flocking_tac06-engineering.pdf.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. URL https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B.

Jannik Peters, Constantin Waubert de Puiseau, Hasan Tercan, Arya Gopikrishnan, Gustavo Adolpho Lucas De Carvalho, Christian Bitter, and Tobias Meisen. Emergent language: A survey and taxonomy. *Autonomous Agents and Multi-Agent Systems*, 39:18, 2025. doi: 10.1007/s10458-025-09691-y. URL https://link.springer.com/article/10.1007/s10458-025-09691-y.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Kenneth O. Stanley and Risto Miikkulainen. Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2):99–127, 2002. URL https://nn.cs.utexas.edu/downloads/papers/stanley.ec02.pdf.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. URL https://arxiv.org/abs/1703.06907.

Stelios Triantafyllou, Aleksa Sukovic, Debmalya Mandal, and Goran Radanovic. Agent-specific effects: A causal effect propagation analysis in multi-agent MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48578–48607, 2024. URL https://proceedings.mlr.press/v235/triantafyllou24a.html.

Stelios Triantafyllou, Aleksa Sukovic, Yasaman Zolfimoselo, and Goran Radanovic. Counterfactual effect decomposition in multi-agent sequential decision making. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60072–60098, 2025. URL https://proceedings.mlr.press/v267/triantafyllou25a.html.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229, 1995. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.75.1226.

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

A PROOFS

Proof of Proposition 1. The paired cells share initial physical states, frames, memories, and exogenous randomness. By Definition 1, changing channel c changes neither action nor memory on their joint reached support. Equal actions produce equal next physical states and memories. Induction therefore gives identical trajectories and identical values of every trajectory functional S, so the paired contrast is zero. □

Proof of Proposition 2. The hypothesized $f^\star$ is both an omitting policy and a maximizer, so $\arg\max J$ intersects $\mathcal{F}_{\setminus p}$ and cannot be contained in its complement. Proposition 1 gives zero correspondence for every omitting maximizer. No peer-using policy has a larger value than $J(f^\star)$; return maximization therefore cannot exclude this omitting solution. □

Lemma 1 (Order does not identify peer use)**.** *If every agent ignores peer state and maintains velocity along a shared unit direction $\widehat{\mathbf{g}}$ encoded in the reference, then $P = 1$ while every peer intervention has zero effect. For any return that is a nondecreasing function of polarization and dynamics that admit constant aligned velocities, this tracker is an omitting maximizer, so Proposition 2 applies under the native law.*

Proof of Lemma 1. Let every agent apply $\mathbf{a}_i = v_0\widehat{\mathbf{g}}$ in world coordinates (locally $R(\theta_i)^\top v_0\widehat{\mathbf{g}}$, a function of $\mathbf{r}_i$) and ignore $\mathbf{p}_i$. Then $P = \|N^{-1}\sum_i \widehat{\mathbf{g}}\| = 1$. The policy lies in $\mathcal{F}_{\setminus p}$, so Proposition 1 gives a zero peer contrast. No policy exceeds $P = 1$, hence the tracker attains $\sup J$ for every nondecreasing function of P. □

Proof of Proposition 3. Fix $\mathcal{D}_I$ and $f \in \mathcal{F}_{\setminus p}$. Under the zero law set $\mathbf{u}_i = \mathbf{0}$ below.

Step 1 (decoupling). By omission, on the support of $\mathcal{D}_I$ agent i's local action and memory update are functions of $(\mathbf{s}_i^t, \mathbf{r}_i, h_i^{t-1})$ and its own policy noise, where $\mathbf{r}_i = R(\theta_i)^\top \mathbf{u}_i$ is constant within the episode. By (A2), $\mathbf{v}_i^{t+1}$ is drawn from a kernel that depends only on $\mathbf{v}_i^t$ and agent i's action, independently across agents. Since the initial variables $(\theta_i, \mathbf{v}_i^0, \mathbf{u}_i)$ and the policy noise are independent across agents by (A3) and the i.i.d. reference law, and all transition kernels factorize over agents, the agents' trajectories are independent.

Step 2 (frame independence). Let $\mathbf{z}_i^t = R(\theta_i)^\top \mathbf{v}_i^t$ and $\tilde{\mathbf{u}}_i = R(\theta_i)^\top \mathbf{u}_i$. By (A1) the local action and memory are drawn from kernels $\pi(\cdot \mid \mathbf{z}_i^t, \tilde{\mathbf{u}}_i, h_i^{t-1})$ that do not involve θ_i. Conditional on $\theta_i = \vartheta$ and on the local history up to t, (A2) gives, for every Borel set B,

$$\Pr\big(\mathbf{z}_i^{t+1} \in B\big) = K\big(R(\vartheta)B \mid R(\vartheta)\mathbf{z}_i^t, R(\vartheta)\mathbf{a}_i^t\big) = K\big(B \mid \mathbf{z}_i^t, \mathbf{a}_i^t\big),$$

which does not involve ϑ. Conditional on $\theta_i = \vartheta$, the initial pair $(R(\vartheta)^\top \mathbf{v}_i^0, R(\vartheta)^\top \mathbf{u}_i)$ has the law of $(\mathbf{v}_i^0, \mathbf{u}_i)$, by independence from θ_i and rotation invariance of their laws, and h_i^0 is fixed. The conditional law of the local process $(\mathbf{z}_i^t, h_i^t)_t$ given $\theta_i = \vartheta$ is determined by this initial law and these kernels (Ionescu-Tulcea), none of which depends on ϑ; hence $(\mathbf{z}_i^t)_t$ is independent of θ_i.

Step 3 (headings). $\widehat{\mathbf{v}}_i^t = R(\theta_i)\widehat{\mathbf{z}}_i^t$ with θ_i uniform and independent of $\widehat{\mathbf{z}}_i^t$. Therefore $\mathbb{E}[\widehat{\mathbf{v}}_i^t] = \mathbb{E}[R(\theta_i)]\,\mathbb{E}[\widehat{\mathbf{z}}_i^t] = \mathbf{0}$, since $\mathbb{E}[R(\theta_i)]$ is the zero matrix, and conditional on $\widehat{\mathbf{z}}_i^t = \mathbf{w}$ with $\|\mathbf{w}\| = 1$ the heading $R(\theta_i)\mathbf{w}$ is uniform. Independence across agents follows from Step 1.

Step 4 (bound and gap). Expanding the square,

$$\mathbb{E}[(P^t)^2] = \frac{1}{N^2}\Big(\sum_i \mathbb{E}\|\widehat{\mathbf{v}}_i^t\|^2 + \sum_{i\neq j} \mathbb{E}[\widehat{\mathbf{v}}_i^t]\cdot\mathbb{E}[\widehat{\mathbf{v}}_j^t]\Big) \le \frac{1}{N},$$

because $\|\widehat{\mathbf{v}}_i^t\| \le 1$ and the cross terms vanish. Jensen's inequality gives $\mathbb{E}[P^t] \le N^{-1/2}$ for each t, and averaging over $t \in \{T_w+1, \dots, T\}$ gives $J_{\mathcal{D}_I}(f) \le N^{-1/2}$. As f is arbitrary, $\sup_{\mathcal{F}_{\setminus p}} J_{\mathcal{D}_I} \le N^{-1/2}$, while $\sup_{\mathcal{F}} J_{\mathcal{D}_I} \ge J_{\mathcal{D}_I}(f_c) \ge 1-\varepsilon$. □

For the metric of Remark 1, fix $f \in \mathcal{F}_{\setminus p}$ under $\mathcal{D}_I$, let $S^t = \{i : \|\mathbf{v}_i^t\| > c\}$ ($c = 0.05$ in the simulator), and $P_M^t = \| |S^t|^{-1} \sum_{i \in S^t} \widehat{\mathbf{v}}_i^t \|$. Since $\|\mathbf{v}_i^t\| = \|\mathbf{z}_i^t\|$, the set S^t is a function of $(\mathbf{z}_i^t)_i$, which is independent of $(\theta_i)_i$ by Steps 1–2. Conditional on $(\mathbf{z}_i^t)_i$, the headings $R(\theta_i)\widehat{\mathbf{z}}_i^t$, $i \in S^t$, are independent unit vectors with mean zero, so $\mathbb{E}[(P_M^t)^2 \mid |S^t| = m] = 1/m$ for $m \ge 1$. With polarization as the only reward, an omitting policy could therefore exceed the $N^{-1/2}$ scale in this metric by stopping most agents; the speed term of the experimental reward penalizes this.

For consequence (ii) of Proposition 3, let a law use the zero-reference law with probability α and a law with a nonzero reference otherwise, and write $J = \alpha J_0 + (1-\alpha) J_{\neq 0}$. Assume $\mathcal{F}$ is closed under the switch that, for $f, g \in \mathcal{F}$, acts as f when the reference is nonzero and as g when it is zero (the reference is constant within an episode). For any $f \in \mathcal{F}_{\setminus p}$, the switch g_f of f and f_c has the same trajectories as f in nonzero-reference worlds, and $J_0(f) \le N^{-1/2}$ by the proposition applied to the zero component. Hence $J(g_f) - J(f) = \alpha\,(J_0(f_c) - J_0(f)) \ge \alpha\,(J_0(f_c) - N^{-1/2})$, and taking the supremum over f bounds the gap from below by the same quantity. With an independent reference in place of the zero law, the switch would have to detect whether its unit reference is shared, which is not a function of local observations.

For the navigation example of Section 2, agents follow $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \eta \mathbf{a}_i^t$ with $\|\mathbf{a}_i^t\| \le 1$ and no inter-agent forces; the shared target $\mathbf{g}$ is encoded only in $\mathbf{r}_i^t = R(\theta_i)^\top(\mathbf{g} - \mathbf{x}_i^t)$, and the return is $U = -\sum_{t=0}^{T-1} \sum_{i=1}^{N} \|\mathbf{x}_i^{t+1} - \mathbf{g}\|$.

Lemma 2 (Greedy displacement)**.** *Let* $\mathbf{d} = \mathbf{g} - \mathbf{x}$ *and* $\eta > 0$*. For* $\|\mathbf{a}\| \le 1$*,* $\|\mathbf{x} + \eta\mathbf{a} - \mathbf{g}\|$ *is minimized by* $\mathbf{a}^\star = \mathbf{0}$ *if* $\mathbf{d} = \mathbf{0}$*, and uniquely by* $\mathbf{a}^\star = (\mathbf{d}/\|\mathbf{d}\|)\min(1, \|\mathbf{d}\|/\eta)$ *otherwise.*

Proof of Lemma 2. Minimizing $\|\mathbf{d} - \eta\mathbf{a}\|$ over the unit ball projects $\mathbf{d}/\eta$ onto that ball. Its unique projection is zero for $\mathbf{d} = 0$, and otherwise $(\mathbf{d}/\|\mathbf{d}\|)\min(1, \|\mathbf{d}\|/\eta)$. From initial distance d_i^0, the triangle inequality bounds any trajectory's distance after $t+1$ steps below by $\max(d_i^0 - (t+1)\eta, 0)$. The greedy map attains all these bounds simultaneously, hence minimizes cumulative remaining distance. It depends on the reference alone and is an omitting maximizer whenever included in the class, proving the navigation application of Proposition 2. □

B Experimental settings

Algorithm 1 states the audit of Section 3.1.

Algorithm 1 SCIA audit of a policy family

Require: frozen policies $\{\pi_k\}_{k=1}^K$ (one per training seed); score S; evaluation seeds $\mathcal{E}$; W worlds per seed; warm-up T_0; bootstrap draws B

Ensure: 95% intervals for $\Delta_R^{\text{corr}}, \Delta_R^{\text{avail}}, \Delta_P^{\text{corr}}, \Delta_P^{\text{avail}}$, I_{RP}, $S_{o,0}$, A_R, and A_P

for each policy π_k and evaluation seed $e \in \mathcal{E}$ **do**
  draw W worlds (states, frames $\theta_i \sim \mathcal{U}[-\pi, \pi]$, directions), reused in every cell ▷ Step 1
  for each cell $(p, r) \in \{o, h, 0\} \times \{s, i, 0\}$ **do**
    peers: o observed, h from another simultaneous world, 0 zeroed ▷ Step 2
    reference: s shared, i the agent's own candidate, 0 zeroed
    roll out π_k in closed loop in frame θ_i, memory per (policy, world, agent)
    $S_{p,r}^{k,e} \leftarrow$ mean of S over worlds, $t > T_0$
  end for
  for each channel $c \in \{R, P\}$ **do**
    $A_c^{k,e} \leftarrow$ mean $\|\pi_k(\mathbf{o}_i) - \pi_k(\mathbf{o}_i^{c \leftarrow c'})\|$ on states of cell (o, s) ▷ Eq. (3)
  end for
end for
$S_{p,r}^k \leftarrow$ mean over e; form each contrast of Equations (1)–(2) per policy ▷ Step 3
for $b = 1, \dots, B$ **do**
  resample K policies with replacement; record each contrast's mean ▷ cluster bootstrap
end for
report estimates, 2.5%/97.5% bootstrap quantiles, and pre-set gate decisions

For the random-frame baseline, if every agent outputs the same local command, independent uniform frames make the world headings independent and uniform, so $\mathbb{E}\|N^{-1}\sum_i \hat{\mathbf{a}}_i\|^2 = 1/N$ (RMS scale $N^{-1/2}$). The corresponding mean is 0.316 for $N = 8$.

In the hand-built controls of Section 4.2, seven independent uniform bits in $\{-1, +1\}$ define 128 equally likely worlds: a target z, receiver frames f_1, f_2, peer donor signs d_1, d_2, and reference signs e_1, e_2. A sender relays z. Receiver i sees the peer input $p_i = f_i z$ (observed), $f_i d_i$ (shuffled), or 0, and the reference $r_i = f_i z$ (shared), $f_i e_i$ (independent), or 0. Shuffling thus keeps the peer marginal exactly and removes its link to z. Each receiver acts for two steps with position update $x_i^{t+1} = x_i^t + f_i a_i^t v_i^t$ from $x_i^0 = 0$, and the score is $S = \frac{1}{2}\sum_i \mathbf{1}[f_i a_i^1 = z]$. All cells reuse the world and reset memory.

The seven policies are, with $b(u) = +1$ for $u \geq 0$ and -1 otherwise, and unit speed unless stated: reference-only $b(r_i)$; peer-only $b(p_i)$; redundant $b(r_i + p_i)$; presence gate $b(r_i)$ if $p_i \neq 0$, else $+1$; score-blind $b(r_i)$ with speed $1 + \mathbf{1}[p_i < 0]$; delayed-peer $b(h_i^t)$ with $h_i^{t+1} = p_i$ and $h_i^0 = 0$; null $+1$.

For the action diagnostic, with the reference fixed at its shared value, the peer input is set to $m \in \{-1, +1\}$ with $q(m) = 1/2$, and C is the mutual information between m and the action A (first action, C_1; two-step trace, $C_{1:2}$):

$$C = \mathbb{E}_{r,f} \sum_{m,a} q(m)\pi(a \mid r, f, \mathrm{do}(m)) \log_2 \frac{\pi(a \mid r, f, \mathrm{do}(m))}{\sum_{m'} q(m')\pi(a \mid r, f, \mathrm{do}(m'))}. \tag{5}$$

This is the CIC form of Lowe et al. (2019) with a uniform intervention law.

For the direct test of Proposition 3, with the peer input set to zero, a frozen policy belongs to $\mathcal{F}_{\setminus p}$; with collisions disabled, the simulator satisfies (A1)–(A3) (Remark 1). Under an independent or zero reference, Proposition 3 then gives $\mathbb{E}[(P^t)^2] = \bar{\nu}_t/N$, where $\bar{\nu}_t$ is the fraction of agents with nonzero velocity. Every agent moved in every world and step ($\bar{\nu} = 1$), so the prediction is $\mathbb{E}[P^2] = 1/N = 0.125$ and $\mathbb{E}[P] \leq N^{-1/2} = 0.354$.

The test uses all 35 trained IPPO policies (native, zero-port, and SI-PO with ten seeds each; balanced with five), 256 fresh worlds per policy and cell, 200 steps, and a 50-step warm-up. P is normalized by all N agents as in (A4); intervals are 95% bootstrap over policies.

In the speed-cue task of Section 4.4, each world draws a target speed $v^\star \sim \mathcal{U}[0.2, 0.5]$, encoded as the reference length $v^\star/0.5$. The speed term of the reward becomes $\mathrm{mean}_i \max\{0, 1 - |\|\mathbf{v}_i\| - v^\star|/v^\star\}$; other terms are unchanged. Native, zero-port, and SI-PO use seeds 11–71 and 9.6M world frames. Zero-port and SI-PO were extended to 19.2M frames post hoc; native had converged (best checkpoints at updates 13–14).

In the goal-reaching task of Section 4.5, each world places one target at distance 3 from the spawn centre in a uniform direction. The reward is $\mathrm{mean}_i \,\mathrm{clamp}(1 - \|\mathbf{x}_i - \mathbf{g}\|/3, -1, 1)$ minus the original cohesion, collision, and boundary terms (no alignment or speed term). Informed agents observe the unit direction to the target. Native trains with all agents informed; the subset recipes train with two of eight informed and give the others zero (zero-port analogue) or a random direction (SI-PO analogue). Seeds 11–71, 19.2M world frames. Evaluation: the few-informed (two informed, others zero) and all-informed deployments, four evaluation seeds with 64 worlds, goal score averaged after a 50-step warm-up.

For the $N = 16$ runs, native and zero-port IPPO are retrained with $N = 16$; all other settings match the $N = 8$ runs (seeds 11–71, 9.6M world frames, full grid). The random-heading RMS scale is 0.25.

C Additional results

Table 3 lists the hand-built control results. The ground truth is whether a nonzero peer value can change the two-step action; it holds for peer-only, redundant, score-blind, and delayed-peer. At threshold 0.10:

- peer zeroing gives one false positive (presence gate) and two false negatives (redundant, score-blind);

- score correspondence gives one false negative (score-blind);
- first-action influence gives one false negative (delayed-peer);
- two-step influence is correct for all seven policies.

Thresholds 0.05 and 0.20 give the same result.

Table 3: Hand-built controls (exact enumeration, no sampling error). S: direction accuracy; C_1, $C_{1:2}$: uniform-intervention mutual information in bits for the first action and the two-step trace.

Control	$S_{o,s}$	Δ_P^{avail}	Δ_P^{corr}	I_{RP}	C_1	$C_{1:2}$
Reference only	1.00	0.00	0.00	0.00	0.00	0.00
Peer only	1.00	0.50	0.50	0.00	1.00	1.00
Redundant	1.00	0.00	0.25	-0.50	0.50	0.50
Presence gate	1.00	0.50	0.00	0.50	0.00	0.00
Score blind	1.00	0.00	0.00	0.00	1.00	1.00
Delayed peer	1.00	0.50	0.50	0.00	0.00	1.00
No information	0.50	0.00	0.00	0.00	0.00	0.00

Table 4 lists the direct test of Proposition 3. With peers zeroed, per-policy $\mathbb{E}[P^2]$ ranges from 0.111 to 0.138, and no policy exceeds $\mathbb{E}[P] = 0.334$. Seven of the eight family intervals contain 0.125; the exception (SI-PO, zero reference, 0.1287) gives 0.1236 ($[0.1210, 0.1263]$) in a replication with 1,024 worlds per policy. With peers observed, all 20 zero-port and SI-PO policies exceed 0.354 (minimum 0.596). Mean $\mathbb{E}[P]$ under the independent and zero references is 0.904 ($[0.860, 0.945]$) and 0.957 ($[0.915, 0.986]$) for zero-port, and 0.812 ($[0.755, 0.853]$) and 0.972 ($[0.916, 1.000]$) for SI-PO. The gap is therefore at least 0.55 (independent) and 0.62 (zero). Native policies stay at 0.316 and 0.344.

Table 4: Direct test of Proposition 3: frozen IPPO policies with peers zeroed; n is the number of policies. Prediction under independent and zero references: $\mathbb{E}[P^2] = 0.125$ and $\mathbb{E}[P] \leq 0.354$. Shared-reference rows are shown for contrast.

Family	n	Reference	$\mathbb{E}[P^2]$ (no collisions)	$\max \mathbb{E}[P]$	$\mathbb{E}[P^2]$ (collisions)
Native	10	independent	0.1261 $[0.1224, 0.1303]$	0.334	0.1262
		zero	0.1231 $[0.1194, 0.1266]$	0.324	0.1238
		shared	0.9538 $[0.9485, 0.9587]$	0.982	0.9542
Balanced	5	independent	0.1245 $[0.1230, 0.1262]$	0.318	0.1254
		zero	0.1276 $[0.1224, 0.1332]$	0.331	0.1299
		shared	0.9751 $[0.9701, 0.9794]$	0.990	0.9756
Zero-port	10	independent	0.1260 $[0.1230, 0.1288]$	0.328	0.1261
		zero	0.1255 $[0.1226, 0.1283]$	0.328	0.1257
		shared	0.1345 $[0.1294, 0.1404]$	0.350	0.1347
SI-PO	10	independent	0.1262 $[0.1219, 0.1304]$	0.333	0.1263
		zero	0.1287 $[0.1262, 0.1312]$	0.332	0.1291
		shared	0.3396 $[0.2151, 0.4898]$	0.903	0.3398
All (pooled)	35	independent	0.1259 $[0.1240, 0.1278]$	0.334	0.1261
		zero	0.1260 $[0.1241, 0.1279]$	0.332	0.1267
		shared	0.5473 $[0.4228, 0.6780]$	0.990	0.5476

Table 5 gives the SI-PO grid of Section 4.3. Figure 4 gives the NEAT profiles of Section 4.2, and Table 6 a guide to reading profiles.

Table 5: SI-PO polarization on the evaluation grid (95% cluster bootstrap over ten training seeds).

Cell	P	95% CI
observed / shared (deployment)	0.936	$[0.881, 0.984]$
observed / independent (training)	0.830	$[0.786, 0.867]$
observed / zero (reference loss)	0.899	$[0.711, 0.999]$
shuffled / shared	0.460	$[0.389, 0.539]$
zero / shared	0.527	$[0.423, 0.654]$
zero / zero	0.284	$[0.228, 0.316]$

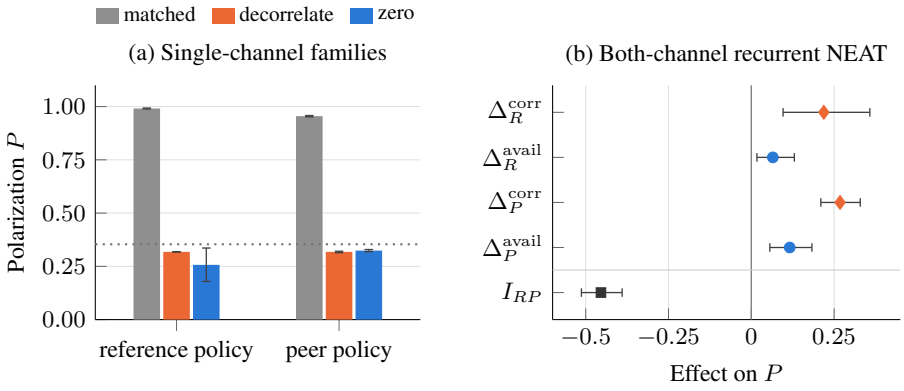

Figure 4: SCIA on frozen NEAT policies. (a) Single-channel families lose order under marginal-preserving decorrelation or zeroing; the dotted line is the $1/\sqrt{8}$ RMS scale of random headings. (b) Both-channel recurrent NEAT has positive correspondence on each channel and a negative interaction I_{RP}: substitution. Estimates with 95% cluster-bootstrap intervals over ten trained policies.

For the action-level check, on states visited in the deployment cell, each policy's deterministic action is recomputed with one channel replaced by the same channel of another world (256 worlds per policy, 50-step warm-up, one step). Action changes (mean action norm 0.600 for native):

- native: reference 0.669 ($[0.660, 0.679]$), peers 0.044 ($[0.037, 0.049]$);
- zero-port: reference 0.072 ($[0.055, 0.088]$), peers 0.428;
- SI-PO: reference 0.429 ($[0.374, 0.483]$), peers 0.531.

Table 6: Reading a channel profile. "Zero-only" is the conclusion of an audit that only zeroes channels; examples are from this paper.

Pattern	Signature	Reading	Zero-only reads	Example
Reference pathway	$\Delta_R^{\text{corr}} \gg 0$, $\Delta_P^{\text{corr}} \approx 0$, low $S_{o,0}$	order from the shared cue	correct	native IPPO
Peer pathway	$\Delta_P^{\text{corr}} \gg 0$, $\Delta_R^{\text{corr}} \approx 0$, high $S_{o,0}$	order from peers	correct	zero-port IPPO
Substitutes	both $\Delta^{\text{corr}} > 0$, small Δ^{avail}, $I_{RP} < 0$	both used, each sufficient	neither used	recurrent NEAT; redundant control
Partial substitution	$\Delta_R^{\text{corr}} > 0$, $\Delta_R^{\text{avail}} \approx 0$, high $S_{o,0}$	reference used when present, not needed	reference ignored	SI-PO
Presence gating	$\Delta^{\text{avail}} > 0$, $\Delta^{\text{corr}} \approx 0$	presence, not content, matters	content used	presence-gate control
Score-invisible use	all score contrasts ≈ 0, action influence > 0	use that leaves the score unchanged	(missed by every score contrast)	score-blind control; weakly, native IPPO's peers

The speed-cue gates, fixed before training, and their outcomes are:

- H1 (port value): SI-PO minus zero-port speed match, CI lower bound > 0. Fails: $+0.165$ ($[-0.020, 0.317]$) at 9.6M and $+0.005$ ($[-0.040, 0.053]$) at 19.2M.
- H2 (peer use): peer correspondence ≥ 0.10, CI lower bound > 0, for SI-PO and zero-port. Passes at both budgets.
- H3 (shortcut): native peer correspondence < 0.10 and reference-loss polarization CI upper bound ≤ 0.55. Passes.

Table 7 lists the results.

Table 7: Speed-cue task. $S_{o,s}$ and $S_{o,0}$ are polarization in the deployment and reference-loss cells; speed match is on the observed/shared cell; $r(v, v^\star)$ is the across-world correlation of mean speed with $v^\star$. Intervals are 95% cluster bootstrap over five training seeds.

Recipe	Frames	Δ_P^{corr}	$S_{o,s}$	$S_{o,0}$	Speed match	$r(v, v^\star)$
Native	9.6M	$0.013\ [0.012, 0.014]$	0.940	0.335	0.636	0.85
Zero-port	9.6M	$0.384\ [0.310, 0.506]$	0.701	0.694	0.459	-0.22
SI-PO	9.6M	$0.420\ [0.353, 0.512]$	0.757	0.762	0.623	0.05
Zero-port	19.2M	$0.574\ [0.507, 0.622]$	0.889	0.883	0.641	-0.07
SI-PO	19.2M	$0.584\ [0.562, 0.605]$	0.901	0.916	0.646	0.02

The goal-reaching gates, fixed before training, and their outcomes are:

- H1 (shortcut): native few-informed peer correspondence < 0.10 and an all-to-few drop ≥ 0.20 with CI lower bound > 0. Passes: drop 0.369 ($[0.343, 0.390]$).
- H2 (peer use): for each subset recipe, few-informed peer correspondence ≥ 0.10 and a goal score at least 0.10 above native, both with CI lower bound > 0. Passes for the zero-port analogue ($+0.146$, $[0.125, 0.167]$); fails for the SI-PO analogue (-0.016, $[-0.034, 0.001]$).

With peers shuffled, the zero-port analogue's uninformed agents score -0.166. The scripted greedy tracker moves every agent at full speed along its reference; its return under the training reward is 143.1, against 69.1 for native. Table 8 lists the results.

Table 8: Goal-reaching task: goal scores and few-informed peer correspondence. Intervals are 95% cluster bootstrap over five training seeds.

Policy	Few-informed	Δ_P^{corr} (few)	All-informed	Inf./uninf. (few)
Scripted greedy	—	—	0.925	—
Native	$0.108\ [0.103, 0.113]$	$0.030\ [0.016, 0.047]$	$0.477\ [0.449, 0.496]$	0.352 / 0.020
Zero-port analogue	$0.254\ [0.234, 0.274]$	$0.261\ [0.229, 0.295]$	$0.801\ [0.789, 0.812]$	0.402 / 0.184
SI-PO analogue	$0.092\ [0.075, 0.108]$	$0.131\ [0.107, 0.155]$	$0.370\ [0.340, 0.401]$	0.129 / 0.068

The $N = 16$ gates, fixed before training, and their outcomes are:

- H1 (shortcut): native peer correspondence < 0.10 and reference-loss polarization CI upper bound ≤ 0.55. Passes.
- H2 (peer use): zero-port peer correspondence ≥ 0.10 with CI lower bound > 0, and CI lower bounds of 0.75 (deployment) and 0.60 (reference loss) on polarization. Passes.

Table 9 lists the results.

Table 9: $N = 16$ runs. Intervals are 95% cluster bootstrap over five training seeds.

Recipe	Δ_P^{corr}	Δ_R^{corr}	$S_{o,s}$	$S_{h,s}$	$S_{o,0}$
Native	$0.018\ [0.016, 0.019]$	$0.750\ [0.744, 0.758]$	0.972	0.954	$0.246\ [0.240, 0.252]$
Zero-port	$0.707\ [0.694, 0.723]$	$0.002\ [0.000, 0.005]$	0.953	0.246	$0.951\ [0.923, 0.978]$