THE AMBIGUITY ZONE: WHEN RANDOMIZED SOCIAL PREFERENCES INDUCE SOCIAL INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized uncertain social preferences (RUSP) can produce reciprocity and reputation-like behavior in multi-agent reinforcement learning by randomizing social rewards and introducing structured uncertainty about others' preferences, but the source of this pressure to understand other agents remains unclear. We propose that there exists a decision-critical region of preference space, the *ambiguity zone*, in which the best action depends on the partner's latent type. The location of this zone can be computed before training from the value of social information. Our results support a two-gate account of social inference. The first gate is *coverage*: across payoff games and training distributions, partner sensitivity increases with the training mass M_{crit} placed on the ambiguity zone (condition-level Spearman $\rho = 0.79$, cluster permutation $p = 0.009$), and the relation extends to a preregistered fourth game. The second gate is *identifiability*: a 2×4 factorial experiment shows zero response outside the ambiguity zone at every information level, while partner-conditioned behavior appears inside the zone only when partner type becomes highly identifiable. A preregistered matched-marginal control reproduces this pattern while holding population cooperation fixed at 0.5. When both gates are open, agents use third-party observations to judge a partner they have never directly interacted with ($S_{\text{rep}} = +0.90 \pm 0.30$); resetting the recurrent state removes the effect, localizing this cross-interaction judgment to memory. Finally, a compact RUSP-form bridge that restores randomized reward transformations, asymmetric preference uncertainty, and multi-agent shared-policy learning preserves the same gate ordering at smaller magnitude. Together, these results distinguish the roles of randomization and uncertainty in RUSP: randomization determines which regions of preference space receive training exposure, while identifiable social evidence determines whether that exposure creates pressure to understand other agents.

1 INTRODUCTION

Agents trained with multi-agent reinforcement learning can develop social behaviors that were not explicitly programmed. Randomized uncertain social preferences (RUSP) provide a striking example (Baker, 2020): agents are trained under randomly transformed social rewards and, in some settings, develop reciprocity, partner discrimination, and reputation-like responses. The empirical phenomenon is established, but the training mechanism remains less clear. In particular, randomizing preferences changes many episodes without explaining which of those episodes create an incentive to infer anything about another agent.

The key distinction is whether partner identity changes the decision. At some social-preference weights, the same action is optimal against every partner type. In those states, partner behavior may be observable, but using it has no decision value. At other weights, different partner types require different actions. We call this region the *ambiguity zone*. Inside the zone, knowing the partner improves expected return; outside it, the task already determines the action. This turns the emergence question into a testable mechanism: social inference should receive training pressure when partner information is both *valuable for the decision* and *recoverable from behavior*.

We formalize these requirements as two gates. Gate 1 is *coverage*. For social-preference level α, we compute the value of social information $V_{\text{social}}(\alpha)$ and define the ambiguity zone $\mathcal{C}$ where that

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.

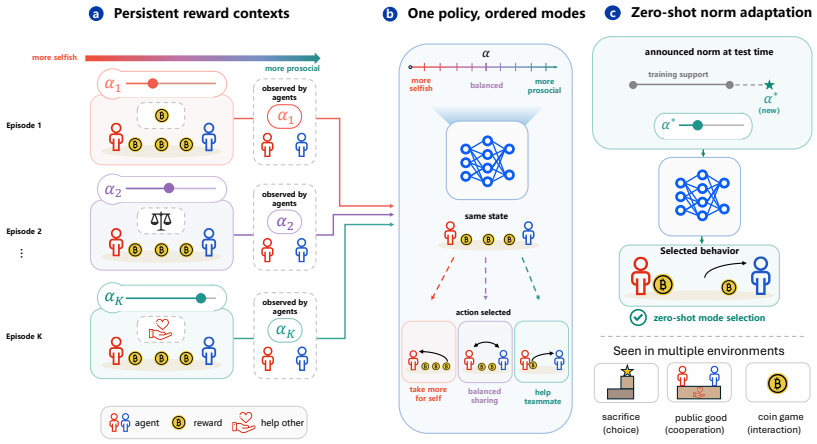


Figure 1: **Persistent social-preference contexts support ordered behavioral modes and zero-shot norm adaptation. (a)** A persistent social-preference context α is drawn per episode and observed noisily. **(b,c)** Preference variation can produce context-conditioned behavioral modes and zero-shot adaptation to announced norms. Appendix F reports these experiments.

value is positive. A training distribution $q(\alpha)$ then supplies a dose $M_{\text{crit}}(q)$, the probability mass it places on $\mathcal{C}$. Gate 2 is *identifiability*. The partner's latent type must be inferable from the interaction history; we quantify this by Bayes-optimal accuracy $\text{Acc}^*(Z \mid H)$. The two quantities separate incentive from information: M_{crit} measures how often understanding the partner matters, while Acc^* measures whether the evidence can support that understanding.

RUSP combines these two ingredients in different roles. Preference randomization allocates training experience across the social-preference axis and therefore controls exposure to decision-critical regimes. Structured uncertainty shapes the social evidence available for identifying persistent partner variation. The two-gate view links them: coverage determines whether partner identity matters during training, while identifiability determines whether behavior reveals the relevant type. Neither is sufficient alone: informative behavior has little effect where action does not depend on type, while critical-zone exposure cannot produce type-conditioned behavior when history does not reveal type.

The experiments follow this two-gate logic. We first vary coverage while holding the partner population fixed and show that M_{crit} organizes social sensitivity across dose ladders, shifted payoff games, and a preregistered fourth game. We then cross zone coverage with partner identifiability in a factorial design. The response remains zero when either gate is closed, and a matched-marginal control confirms the identifiability effect without changing the population cooperation marginal. With both gates open, we test what capability is learned: third-party observation changes first-contact behavior toward a never-met partner, and recurrent-state ablation removes the effect. Finally, we restore the main structural ingredients of RUSP—random reward transformations, asymmetric preference uncertainty, and multi-agent shared-policy learning—and test whether the same gate ordering survives.

This account also clarifies the role of randomization. The ambiguity zone depends jointly on the game and the partner population, so its location is generally unknown before solving the interaction structure. Randomization is therefore useful as a coverage mechanism: it distributes training experience across preference space and can place mass on decision-critical social regimes without first locating them. This yields a predictive claim: given a task, a partner population, and a training distribution, the two-gate account specifies where pressure for social inference should appear and where it should vanish.

Contributions. We make three contributions. First, we define the ambiguity zone through the value of social information and derive the critical training mass M_{crit} as a measurable exposure variable. Second, we establish a two-gate empirical account in which zone coverage supplies incentive and partner identifiability supplies usable evidence, including cross-game, preregistered, factorial, and matched-marginal tests. Third, we show that the resulting capability can take the

form of memory-based indirect reputation and that the gate ordering persists in a compact RUSP-form bridge. Together, these results explain preference randomization as a mechanism for exposing learners to social decisions in which modeling another agent is useful.

2 Related Work

Randomized social preferences. RUSP (Baker, 2020) randomizes a reward-sharing transformation with structured asymmetric observation noise and reports emergent direct and indirect reciprocity and team formation. Reward randomization has also been used to diversify strategic behavior (Tang et al., 2021), while heterogeneous social value orientations shape mixed-motive MARL populations (McKee et al., 2020). Our analysis characterizes which parts of a randomized preference distribution make partner identity decision-relevant, how much training exposure those parts receive, and whether interaction history makes the relevant partner variation identifiable.

Opponent modeling and reputation. Inferring other agents' goals, types, or policies is a mature area (Albrecht and Stone, 2018), including recursive reasoning, opponent-learning awareness (Foerster et al., 2018), Self–Other Modeling (Raileanu et al., 2018), and theory-of-mind networks (Rabinowitz et al., 2018). Reputation and indirect reciprocity have likewise been modeled through image scoring, standing strategies, and learned social norms (Nowak and Sigmund, 1998; Anastassacos et al., 2021; Vinitsky et al., 2023). These approaches characterize mechanisms for representing or using information about others. We study the conditions under which ordinary recurrent model-free learning receives enough reward and information pressure for partner-dependent representations to become behaviorally useful.

Incomplete information and value of information. Classical repeated-game theory showed that incomplete information about an opponent's behavior or payoffs can sustain cooperation and reputation in finite games (Kreps et al., 1982; Kreps and Wilson, 1982). More generally, information-structure analyses of multi-agent learning use value of information to quantify the utility of observations for strategic decision making (Li et al., 2022). Our formulation connects these ideas to learned social preferences: V_{social} localizes where partner type changes the optimal action, M_{crit} measures how often training visits those decisions, and Acc^* separately measures whether behavior reveals the relevant type.

Social dilemmas in MARL. Sequential social dilemmas and extensions based on inequity aversion or social influence study how reward design changes cooperation (Leibo et al., 2017; Hughes et al., 2018; Jaques et al., 2019). Our dependent variable is partner-conditional behavior under matched states; an agent that cooperates equally with all partners has zero social-inference sensitivity by construction.

3 Method

3.1 Social-Preference Setting

We study iterated two-action social dilemmas with stage payoffs (R, S, T, P). Agent i receives a socially mixed reward $r_i = \text{own} + \alpha\,\text{other}$, where α is fixed within an episode and varies across training according to $q(\alpha)$. Let Z denote the partner's latent behavioral type, H the interaction history, and $Q(a \mid Z, \alpha)$ the finite-horizon value of action a against a known type. The partner type is never supplied directly to the learner.

3.2 Ambiguity Zone and Critical Mass

The value of social information at preference level α is

$$V_{\text{social}}(\alpha) \;=\; \mathbb{E}_Z\Big[\max_a Q(a \mid Z, \alpha)\Big] \;-\; \max_a \mathbb{E}_Z[Q(a \mid Z, \alpha)]\,. \tag{1}$$

This is the expected return gained by conditioning the decision on partner type. We define the **ambiguity zone** (formally, the critical region) as $\mathcal{C}_\epsilon = \{\alpha : V_{\text{social}}(\alpha) > \epsilon\}$; we use $\epsilon = 0$

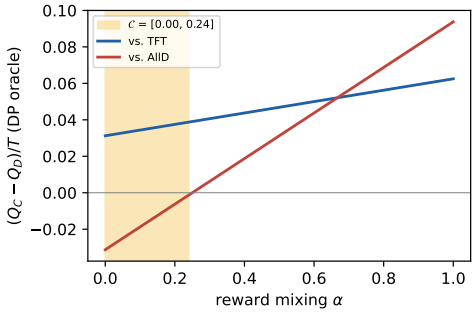


Figure 2: **Ambiguity zones from exact dynamic programming.** The oracle action preference $(Q_C - Q_D)/T$ is evaluated against two partner types before training. The shaded region $\mathcal{C}$ contains preference weights for which partner type changes the optimal action. The zone moves with the game and partner population, making it a property of their joint interaction structure.

throughout. Inside $\mathcal{C}_\epsilon$, different partner types change the optimal action; outside it, one action is optimal across the population. For a training distribution $q(\alpha)$, the associated exposure is the **critical training mass**

$$M_{\text{crit}}(q) \;=\; \int_{\mathcal{C}_\epsilon} q(\alpha)\, d\alpha. \tag{2}$$

Thus M_{crit} measures how much training occurs where partner identity has positive decision value. It measures exposure rather than within-zone value magnitude: two distributions can share M_{crit} while weighting different V_{social} values. Appendix C reports a preregistered within-zone comparison that holds $M_{\text{crit}} = 1$ while varying $\mathbb{E}_q[V_{\text{social}}]$.

Figure 2 illustrates two consequences. With standard payoffs $(3, 0, 5, 1)$, exact dynamic programming over T=32 gives $\mathcal{C} = (0.25, 0.66)$ against an {All-C, All-D} population, with peak V_{social} at $\alpha = 0.43$. Against a {TFT, All-D} population, the zone shifts to $[0, 0.24]$. The same α can therefore be decision-critical in one population and irrelevant in another. This dependence motivates random coverage: a broad training distribution can encounter a critical region even when its location is unknown to the designer.

3.3 Partner Identifiability

Coverage creates a reason to distinguish partner types, but the history must contain enough information to do so. We quantify the second gate by Bayes-optimal identifiability $\text{Acc}^*(Z \mid H_L)$, the accuracy of the optimal classifier of latent type Z from an L-step interaction history. In the controlled experiments, Acc^* is obtained by exact enumeration or Monte-Carlo evaluation of the known partner process. Thus M_{crit} measures exposure to decisions where type matters, whereas Acc^* measures whether behavior reveals the type. The gate is defined by information remaining in interaction history rather than by a particular noise parameter; whether learning converts that information into partner-conditioned behavior is tested below.

3.4 Training and Evaluation

Agents are GRU policies (Cho et al., 2014) with hidden size 32 trained by PPO (Schulman et al., 2017) on horizons of 16–32. Observations contain both players' last actions, the last reward, the time index, and the agent's noisy preference view, but no partner identity. The main measure is trust sensitivity S_{trust}, the difference in cooperation toward different latent partner types under matched states at a fixed probe α. We use nonzero partner-conditioned behavior under matched states as the behavioral criterion for social inference. For cross-game comparisons we use $S_{\text{directed}} = d_G S_G$, with the sign d_G fixed by the dynamic-programming oracle before the corresponding confirmatory runs. Indirect reputation uses S_{rep}, the difference in the focal agent's first action toward a never-met partner after observing that partner interact with a third party. Statistics use exact paired seed-level sign-flip tests unless stated otherwise. Confirmatory predictions are stored in timestamped JSON files before the corresponding runs.

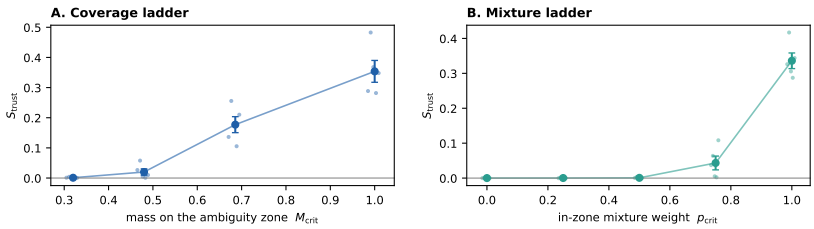


Figure 3: **Critical-mass dose response. (A)** Widening $U[0, x]$ lowers the fraction of training inside the ambiguity zone and reduces partner sensitivity ($+0.354 \to +0.177 \to +0.020 \to \approx 0$). **(B)** With the off-zone components fixed, increasing only the in-zone mixture weight restores the response ($\approx 0/\approx 0/\approx 0/+0.043/+0.336$). Dots are seeds ($n$=5); bars are mean $\pm$ SEM.

4 Results

4.1 Critical-Mass Dose Response

Dose ladders. With scripted {TFT, All-D} populations, whose zone is $\mathcal{C} = [0, 0.24]$, training distributions with greater mass on $\mathcal{C}$ produce stronger partner sensitivity. For $\alpha \sim U[0, x]$, S_{trust} decreases from $+0.354$ at $x = 0.24$ to $+0.177$ at $x = 0.35$, $+0.020$ at $x = 0.50$, and approximately zero at $x = 0.75$. A complementary mixture ladder keeps the off-zone component fixed and varies only the in-zone probability p_{crit}, yielding $0.000/0.000/0.001/0.043/0.336$ for $p_{\text{crit}} = 0, 0.25, 0.5, 0.75, 1$. Matched-width off-band randomization and IID observation-noise controls are zero in every seed. In self-play, the fixed-α profile of $|S_{\text{trust}}|$ is bell-shaped and peaks at the dynamic-programming band center (0.085 at $\alpha = 0.45$, $n = 10$); band randomization $U[0.25, 0.65]$ recovers 85% of that peak ($|S| = 0.072$, CI $[0.058, 0.087]$; off-band and IID-noise controls 0.000; $10/10$ seeds, exact $p = 0.002$, $d_z = 2.9$). Together, the two ladders organize the response by zone exposure, while the matched off-band and IID-noise controls remain at zero.

Cross-game alignment. The dose ladders vary exposure within one interaction structure. We next move the decision-critical region by changing the payoff matrix: if critical mass captures the relevant training pressure, targeted exposure should track the shifted zone. Three games have dynamic-programming bands $(0.05, 0.34)$, $(0.25, 0.66)$, and $(0.55, 0.89)$. A fixed $\alpha = 0.45$ produces directed sensitivity only when that value falls inside the game's band ($S = -0.502$, $5/5$; zero in the shifted games). Broad $U(0, 1)$ training gives -0.001 and -0.014 in the shifted games, where its critical mass is only 0.29–0.34, while targeted bands restore the predicted directional effects ($+0.171$ and -0.265, each $5/5$). Across the ten game–distribution conditions, M_{crit} organizes S_{directed} with condition-level Spearman $\rho = 0.794$ and cluster-respecting permutation $p = 0.0087$; within-game correlations are 0.80–1.00, and leave-one-game-out values are 0.77–0.83.

Preregistered fourth game. For payoffs $(3, 0, 7, 2)$, we froze the zone $[0, 0.37]$, the oracle direction, three equal-width windows with $M_{\text{crit}} = 0.10/0.55/1.00$, their predicted ordering, and the statistical test before any G4 training. The observed S_{directed} values are $-0.078 \to +0.150 \to +0.311$ ($n = 10$ each), with a top-vs-bottom gap of 0.389 and exact $p = 0.002$. Four of five preregistered criteria pass. The fifth predicted the lowest-mass window to be indistinguishable from zero; its measured $|S| = 0.078$ has the opposite directed sign. The out-of-sample ordering follows the same critical-mass axis.

4.2 Partner Identifiability as the Second Gate

We cross coverage (critical band versus matched-width off-band) with partner informativeness in one environment. A scripted partner follows its type-consistent action independently at each step with probability q, where $q = 0.5$ removes type information and $q = 1$ makes behavior deterministic. The corresponding Bayes-optimal identifiability values are $\text{Acc}^*(Z \mid H_{16}) = 0.50/0.81/0.97/1.00$ for $q = 0.5/0.65/0.8/1.0$. All eight cells use the same training protocol with $n = 10$ seeds.

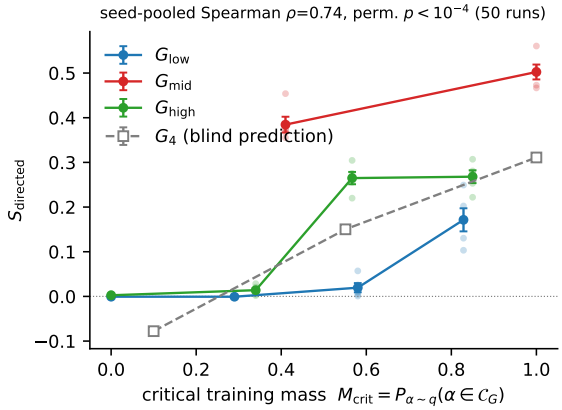


Figure 4: **Cross-game alignment with critical training mass.** Each filled point is one game–distribution condition averaged over five seeds. The primary condition-level analysis gives Spearman $\rho = 0.794$ with cluster permutation $p = 0.0087$; the in-panel annotation reports the seed-pooled diagnostic across 50 runs ($\rho = 0.74$, $p < 10^{-4}$). Open squares show the preregistered fourth game, whose zone, direction, windows, and ordering were fixed before training.

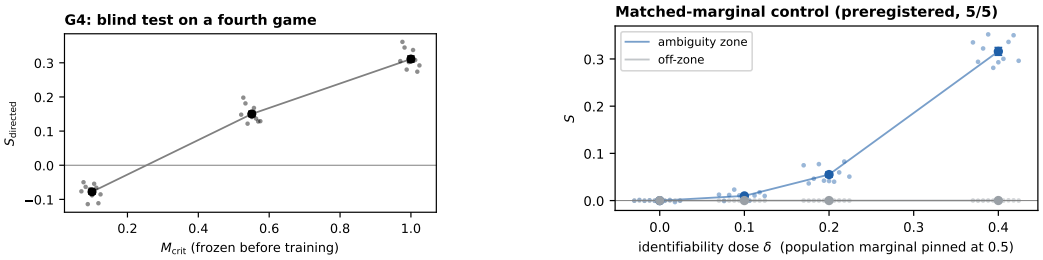


Figure 5: **Cross-game and matched-marginal tests of the two-gate account.** *Left:* the fourth-game test uses a zone, direction, training windows, and ordering fixed before training; S_{directed} rises $-0.078 \rightarrow +0.150 \rightarrow +0.311$ as M_{crit} increases. *Right:* a matched-marginal identifiability experiment pins population cooperation at 0.5 for every dose; the in-zone response rises $-0.000/+0.010/+0.055/+0.316$, while the off-zone row remains at zero ($n = 10$ per cell).

The two gates interact sharply (Table 1). Every off-critical cell remains at zero across all four identifiability levels. Inside the critical band, sensitivity is also zero at $\text{Acc}^* \leq 0.81$, then rises to $+0.282$ (CI $[0.25, 0.31]$, $10/10$) at $\text{Acc}^* = 0.97$ and $+0.343$ (CI $[0.31, 0.38]$) at $\text{Acc}^* = 1.0$; the $q = 1.0$ versus $q = 0.5$ contrast is positive in $10/10$ seeds with exact $p = 0.002$. The preregistered strict-monotonicity prediction is not met because the two low-information cells are tied. At $\text{Acc}^* = 0.81$ the type is recoverable above chance, while the learned response remains zero within this training budget. The gap distinguishes information present in the history from information that training turns into policy behavior.

Matched-marginal identifiability control. Increasing q also changes the marginal cooperation rate of the partner population. We therefore use a preregistered control in which partner types are non-reactive Bernoulli cooperators with $P(C \mid \text{hi}) = 0.5 + \delta$ and $P(C \mid \text{lo}) = 0.5 - \delta$, so the population marginal is exactly 0.5 for every δ. The critical row is evaluated at the analytic indifference point $\alpha^* = 3/7$ and the off-zone row at $\alpha = 0.9$, with the identical partner population in both. All five frozen predictions are met: in-zone sensitivity rises $-0.000/+0.010/+0.055/+0.316$ for $\delta = 0/0.1/0.2/0.4$ ($10/10$ seeds positive at $\delta \geq 0.2$, one-sided exact $p = 0.001$ in the preregistered oracle direction), the off-zone row stays within $|S| < 10^{-3}$ at every dose, and realized training marginals are 0.500 ± 0.003. The right panel of Figure 5 shows that partner information changes the response even when marginal cooperation is held fixed. Together, the in-zone and off-

Table 1: **Two-gate factorial.** Mean S_{trust} at the in-band probe (n=10 seeds per cell). Zone coverage and high partner identifiability are jointly associated with the nonzero response.

	q=0.5 Acc^* =0.50	q=0.65 Acc^* =0.81	q=0.8 Acc^* =0.97	q=1.0 Acc^* =1.00
critical band	0.000	0.000	+0.282	+0.343
off-critical	0.000	0.000	0.000	0.000

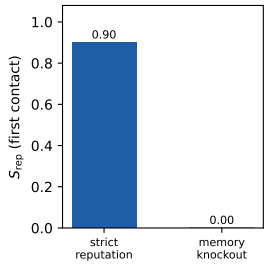


Figure 6: **Indirect reputation and memory localization.** Under critical-band training, third-party observation changes first-contact behavior toward a never-met partner. Zeroing the recurrent state between observation and encounter removes the effect. Bars are seed means ($n = 10$).

zone rows show that social conditioning appears when partner information is both identifiable and decision-relevant.

4.3 Indirect Reputation and Memory

The two-gate measurements establish when partner-conditioned behavior appears. We next test whether this conditioning supports a stronger social capability across interaction boundaries. In the strict reputation protocol, focal agent A observes partner B interact with a third party for 16 steps and then meets B for the first time. B is Grim or All-D. At the start of phase 2, the observation contains no phase-1 trace, so any first-contact discrimination must be carried by A's recurrent state. The metric S_{rep} is the type-conditional difference in A's first action toward B.

Critical-band training yields $S_{\text{rep}} = +0.900 \pm 0.300$ on the first move ($9/10$ seeds, exact $p = 0.0039$, $d_z = 2.85$) and $+0.981 \pm 0.040$ over the whole phase ($10/10$, $p = 0.002$). Every off-band cell is 0.000. Resetting the GRU state between phases reduces the first-move effect from $+0.900$ to 0.000 ($9/10$, $p = 0.0039$), locating the judgment in memory. Earlier stochastic-partner experiments show that when the partner's last action already serves as a sufficient statistic, sensitivity is knockout-insensitive; when that shortcut is removed in strict reputation, multi-step memory is recruited. Appendix D reports these diagnostics. The result refines the second gate: available information can be used through the cheapest sufficient statistic, while longer memory appears when the decision requires it.

4.4 RUSP-Form Bridge

The preceding experiments isolate the two gates under controlled partner processes. We finally restore a compact RUSP-form setting with three agents, per-episode random reward transformations, asymmetric uncertainty about own versus others' preferences ($\sigma_{\text{own}} = 0.1$, $\sigma_{\text{other}} = 0.3$), all three agents controlled by one shared learning policy during training, and hold-out evaluation at the identity transformation in which the observed agent is replaced by a scripted type, as in the original protocol. Four conditions vary critical mass and stable partner heterogeneity.

The first bridge design sampled pairwise preferences independently and produced values near zero in every condition. Under that design, behavior toward one agent does not reveal disposition toward another, so third-party history carries little information for the later interaction. The revised preregistered design uses row-constant generosity types, restoring cross-relationship consistency

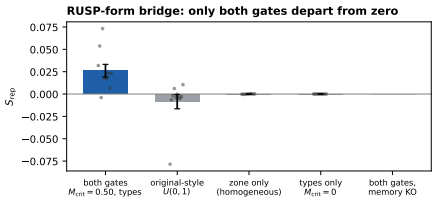


Figure 7: **RUSP-form bridge.** With random reward transformations, asymmetric uncertainty, and three interacting agents trained with one shared policy, the condition combining critical-zone mass $M_{\text{crit}} = 0.50$ with consistent type heterogeneity produces the nonzero reputation signal ($+0.026$, $9/10$ seeds positive, exact seed-level permutation $p = 1.2 \times 10^{-4}$). Memory knockout returns the response to zero. Dots are seeds ($n = 10$ per condition).

while preserving the asymmetric uncertainty structure. In this version, the both-gates condition ($M_{\text{crit}} = 0.50$ with consistent heterogeneity) reaches $S_{\text{rep}} = +0.026 \pm 0.022$, with $9/10$ seeds positive and an exact label-permutation $p = 1.2 \times 10^{-4}$ against the enriched condition. Memory knockout again gives zero. The original-style broad $U(0, 1)$ condition has $M_{\text{crit}} = 0.24$ and yields $-0.009 \approx 0$; pure critical-band training with homogeneous types ($M_{\text{crit}} = 1.0$) is zero to three decimals because cooperation does not bootstrap; heterogeneity without critical mass is also zero in every seed.

The bridge magnitudes follow the earlier dose relation: $M_{\text{crit}} = 0.24$ lies near the region where the scripted curve is approximately zero, while $M_{\text{crit}} = 0.50$ produces $+0.026$ compared with $+0.020$ near mass 0.48 in the scripted experiments. The preregistered magnitude targets ($0.10/0.05$) are not met. In this bridge, reward randomization controls ambiguity-zone exposure, consistent preference heterogeneity under asymmetric uncertainty provides identifiable cross-relationship variation, and recurrent memory carries that variation into first-contact behavior. Appendix E gives both bridge versions and the full-matrix boundary.

5 Discussion

The experiments support a simple organization of the RUSP mechanism: social inference requires both a reason to distinguish partners and evidence that can support the distinction. The ambiguity zone supplies the first condition. It identifies preference weights where the optimal action changes with partner type, and M_{crit} measures how often training visits those decisions. Partner identifiability supplies the second. The factorial and matched-marginal experiments show that informative behavior does not generate social conditioning outside the zone, while zone coverage does not generate conditioning when partner type is insufficiently recoverable. The two quantities therefore separate *decision value* from *behavioral evidence*.

Randomization functions as a coverage mechanism over preference space. Because the ambiguity zone changes with the game and population, broad randomization can encounter a critical region without locating it in advance, but it may under-allocate mass when that region is narrow. Structured uncertainty shapes the evidence from which partner variation can be recovered. Its empirical role is therefore captured by identifiability rather than noise magnitude itself, consistent with the zero IID-noise controls and matched-marginal manipulation.

The identifiability gate has an additional learning consequence. Bayes-optimal identifiability reaches 0.81 before any learned response appears in the factorial, showing that statistical availability is weaker than behavioral learnability at the tested budget. The memory experiments sharpen the point: agents use a one-step statistic when it is sufficient and recruit recurrent memory only when the decision cannot be supported by the shortcut. Social inference therefore need not correspond to a single internal architecture; the training pressure selects the cheapest representation that can exploit the available social information.

Appendix F provides the own-context comparison: one policy can condition on its *own* persistent context across the preference axis, whereas inference about another agent is localized by decision relevance and identifiability.

Together, these results turn a qualitative observation into a predictive workflow. For a new social-learning setting, one can locate where partner identity changes the best response, measure how much training mass covers that region, and test whether behavioral history identifies the relevant variation. These steps separate decision relevance, exposure, and usable evidence, which are easily conflated in randomized social training. The factorial further shows that statistical identifiability is not a universal behavioral threshold: $\mathrm{Acc}^* = 0.81$ is already above chance while the learned response remains zero at the tested budget. Training and representation can therefore affect when the second gate becomes behaviorally expressed.

6 LIMITATIONS

The evidence comes from matrix-game social dilemmas with scripted or learned partners, and the largest bridge contains three agents controlled by one shared learning policy, with the observed agent's type scripted only in the hold-out evaluation probe. The two gates are therefore established as organizing variables in these settings, while larger populations and richer sequential environments remain to be tested. M_{crit} is a coarse exposure statistic. A preregistered comparison holds $M_{\mathrm{crit}} = 1$ while changing $\mathbb{E}_q[V_{\mathrm{social}}]$ by $4.5\times$ and finds no detectable difference ($+0.305$ versus $+0.297$, paired exact $p = 0.24$), but both distributions may lie near a response ceiling. The observed identifiability onset near $\mathrm{Acc}^* \approx 0.97$ is specific to the tested task and training budget; Acc^* characterizes information in the interaction history rather than a universal threshold on raw observation noise. Finally, the compact RUSP-form bridge reproduces the gate ordering at small magnitude, while the broad $U(0, 1)$ recipe has no measurable indirect-reputation signal at this scale. A faithful large-scale replication is needed to determine how much of the original RUSP magnitude is explained by the same mechanism.

7 CONCLUSION

This work asks when randomized social preferences create pressure for an agent to understand other agents. Across the tested settings, the answer is organized by two gates. Social information must change the decision, captured by training exposure to the ambiguity zone, and the relevant partner variation must be identifiable from behavior. Critical-mass dose responses, cross-game transfer, a preregistered fourth game, a two-factor identifiability experiment, a matched-marginal control, and a memory knockout all support this account. The RUSP-form bridge further shows that the same ordering survives when random reward transformations, asymmetric uncertainty, and multi-agent shared-policy learning are restored.

The broader significance is that preference randomization can be analyzed as a training-distribution design problem. Useful episodes place agents where another agent's latent type has decision value, while identifiability determines whether that value can be exploited. Randomization therefore allocates exposure across preference space, and structured uncertainty shapes the usable social evidence. In richer settings, the same logic suggests locating decision-critical social regimes, then measuring their exposure and the identifiability of the partner variation that changes the decision. This perspective provides a way to predict, diagnose, and eventually design for social inference within ordinary recurrent model-free learning.

REPRODUCIBILITY STATEMENT

Every preregistered prediction is stored in timestamped JSON before the corresponding runs. The registry includes oracle bands, band-shift designs, the two-gate factorial, the G4 test, the two bridge versions, the matched-marginal control, and the M_{crit}-versus-$\mathbb{E}[V_{\mathrm{social}}]$ comparison. The corresponding failures and partial confirmations remain in the manuscript. All statistics use training seeds as the unit of inference, and the condition-level reanalysis scripts and raw per-seed JSON files accompany the code.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, provide feedback on experimental design, support result interpretation, assist with code implementation and debugging, and support manuscript drafting, translation, revision, and grammar editing. The research hypotheses and core conceptual framing were developed and refined by the authors. All AI-assisted outputs incorporated into the work were reviewed and verified by the authors, who take responsibility for the final content.

REFERENCES

Bowen Baker. Emergent reciprocity and team formation from randomized uncertain social preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhenggang Tang, Chao Yu, Boyuan Chen, Huazhe Xu, Xiaolong Wang, Fei Fang, Simon Du, Yu Wang, and Yi Wu. Discovering diverse multi-agent strategic behavior via reward randomization. In *International Conference on Learning Representations (ICLR)*, 2021.

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International Conference on Machine Learning (ICML)*, 2018.

Martin A. Nowak and Karl Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393:573–577, 1998.

Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. Cooperation and reputation dynamics with reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

David M. Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252, 1982.

David M. Kreps and Robert Wilson. Reputation and imperfect information. *Journal of Economic Theory*, 27(2):253–279, 1982.

Tao Li, Yuhan Zhao, and Quanyan Zhu. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control*, 53:296–314, 2022.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

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 *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020.

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

Eugene Vinitsky, Raphael Köster, John P. Agapiou, Edgar A. Duéñez-Guzmán, Alexander S. Vezhnevets, and Joel Z. Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2), 2023.

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

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

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

A Oracle Construction and Game-Specific Zones

Dynamic-programming oracle. For a repeated two-action game with stage payoffs (R, S, T, P) and the same reward mixing used in the main experiments, $r_i^{\text{eff}} = r_i + \alpha r_j$, the oracle solves the finite-horizon control problem of a focal agent facing a fixed reactive partner by exact backward induction over horizon $T = 32$, undiscounted. The state contains the partner's one-step memory. On the grid $\alpha \in \{0, 0.01, \ldots, 1\}$, the decision-critical region is $\mathcal{C}_G = \{\alpha : a^*(z_1, \alpha) \neq a^*(z_2, \alpha)\}$ for a specified partner pair (z_1, z_2). The semantic trust band additionally requires $a^*(\text{TFT}) = C$ and $a^*(\text{AllD}) = D$.

Game-specific bands. For the base prisoner's dilemma $(3, 0, 5, 1)$, the anti-coordination band against AllC versus AllD is $[0.25, 0.66]$ and the trust band against TFT or Grim versus AllD is $[0.00, 0.24]$. The shifted games were frozen in `results_shift/design.json` before training: $G_{\text{low}} = (4.0, 0.0, 5.4, 0.25)$ with trust band $[0.05, 0.34]$; $G_{\text{mid}} = (3.0, 0.0, 5.0, 1.0)$ with band $[0.25, 0.66]$; and $G_{\text{high}} = (3.0, 0.0, 5.7, 2.02)$ with band $[0.55, 0.89]$. The fixed-α control point is 0.45, which lies inside only the G_{mid} band. G4 is $(3.0, 0.0, 7.0, 2.0)$ with band $[0.00, 0.37]$, frozen before G4 training.

Directed social sensitivity. Cross-game results use $S_{\text{directed}} = d_G S_G$, where S_G is the measured sensitivity in game G. The direction d_G is determined by a two-level rule: first by the probe-pair dynamic-programming sign, $d = \mathbf{1}[Q_C > Q_D \mid \text{AllC}] - \mathbf{1}[Q_C > Q_D \mid \text{AllD}]$ at the probe α; if that comparison is indifferent ($d = 0$), the training-pair band direction $+1$ is used. G_{mid} and G_{high} resolve at the first level with $d = -1$; G_{low} and G4 use the second level with $d = +1$. The same rule was stored in the G4 prediction file before training.

B Experimental Conditions and Dose Designs

Scripted-partner factorization. The primary partner population is `tft_alld`, with in-band center 0.12 and off-band center 0.60; `allc_alld` uses 0.45 and 0.85. Conditions include `fixed_in`/`fixed_off`, `band_in`/`band_off`, exploratory `sweep_in`, IID-noise `iidn_off`, `full_range` $U(0, 1)$, the coverage ladder `cover_x` with $x \in \{0.24, 0.35, 0.50, 0.75\}$, and the mixture ladder `mixp_p` with $p \in \{0, 0.25, 0.5, 0.75, 1\}$. Coverage results ($n = 5$) are $+0.354/+0.177/+0.020/+0.001$ for $x = 0.24/0.35/0.50/0.75$; mixture results are $\approx 0/\approx 0/\approx 0/+0.043/+0.336$ for $p = 0/0.25/0.5/0.75/1$.

Band-shift design. The initial design, frozen before training, used a fixed condition ($\alpha = 0.45$ every episode) and `urange` ($\alpha \sim U(0, 1)$), five seeds each. The `urange` prediction failed on the shifted games, with means -0.001 and -0.014 versus G_{mid}'s -0.385. A dose addendum was frozen before the additional runs. Targeted windows $U(0, 0.5)$/$U(0, 0.35)$ for G_{low} and $U(0.4, 1)$/$U(0.5, 0.9)$ for G_{high} yield $+0.171$ and -0.265 at the highest targeted masses. Across all 50 training runs, the seed-pooled correlation is $\rho = 0.74$ with $p < 10^{-4}$; the condition-level analysis in the main text remains primary because seeds within a game–distribution cell share the same design point.

Two-gate factorial. The design in `twogate_frozen.json` was frozen before the corresponding runs. Criticality is `band_in` or `band_off` and consistency is $q \in \{0.5, 0.65, 0.8, 1.0\}$, with $n = 10$ seeds per cell. Partner type is resampled each episode, and q is the probability that the partner follows its type-consistent action at each step.

C Identifiability and Marginal Controls

Bayes-optimal identifiability. $\mathrm{Acc}^*(Z \mid H_L)$ is the MAP accuracy of recovering partner type Z from $L = 16$ action-history steps. For the consistency manipulation, it is estimated with 2×10^5 Monte-Carlo episodes per reference subject. The random-subject values are $0.499/0.812/0.971/1.000$ for $q = 0.5/0.65/0.8/1.0$; the AllC reference gives $0.500/0.887/0.996/1.000$. In the original factorial, population cooperation decreases from 0.500 at $q = 0.5$ to 0.266 at $q = 1.0$. The off-band row experiences the same population change while staying within $|S| < 0.005$ for every seed.

Matched-marginal control. Partners are non-reactive Bernoulli cooperators drawn with equal prior: $P(C \mid \mathrm{hi}) = 0.5 + \delta$ and $P(C \mid \mathrm{lo}) = 0.5 - \delta$, for $\delta \in \{0, 0.1, 0.2, 0.4\}$. The population cooperation marginal is therefore 0.5 at every dose. With reward $r = \mathrm{own} + \alpha \cdot \mathrm{partner}$ and standard PD payoffs, the one-shot gap is $\Delta(p, \alpha) = 4\alpha - 1 - p(\alpha + 1)$. Against $p = 0.5$, the indifference point is $\alpha^* = 3/7$; at $\alpha = 0.9$, cooperation dominates against both types. Closed-form $\mathrm{Acc}^*(Z \mid H_{16})$ is $0.50/0.79/0.95/1.00$ for $\delta = 0/0.1/0.2/0.4$. The design, direction, and five predictions were frozen before training in `results_mm/frozen_predictions_mm.json`. Training uses 600 updates, batch 128, and $n = 10$ seeds per cell. Realized training marginals deviate from 0.5 by at most 0.003. In this control, memory knockout does not remove the response because a single last action identifies type with accuracy $0.5 + \delta$.

Critical mass versus expected social-information value. A preregistered discrimination test fixes $\delta = 0.4$ and compares two width-0.10 training distributions lying entirely inside the critical band $\alpha \in (0.282, 0.613)$, so $M_{\mathrm{crit}} = 1$ in both. `evsi_center` is $U(0.40, 0.50)$ and `evsi_edge` is uniform over two 0.05-wide slivers at the band edges. Their $\mathbb{E}[V_{\mathrm{social}}]$ values differ by $4.5\times$, 0.238 versus 0.053 per step. The observed sensitivities are $+0.305 \pm 0.038$ and $+0.297 \pm 0.040$ ($10/10$ seeds positive in both), with paired exact sign-flip $p = 0.24$. At this scale, the test does not separate these two within-zone doses.

D Reputation and Memory Diagnostics

Strict reputation protocol. Each 32-step episode has two phases. During the first 16 steps, focal learner A observes partner B interact with a TFT bystander; A takes no payoff-relevant action toward B. During the second 16 steps, A plays B for the first time. B's latent type is Grim or AllD with equal probability. $S_{\mathrm{rep}} = P(C_A \text{ at } t = 16 \mid \mathrm{good}) - P(C_A \text{ at } t = 16 \mid \mathrm{bad})$ measures first-contact discrimination before direct experience. Carry mode includes B's last phase-1 action in the phase-2 observation, while strict mode removes it and leaves only A's recurrent state. Training uses 1500 updates, batch 48, GRU hidden size 32. The preference conditions are `band_in` $U(0, 0.24)$ and `band_off` $U(0.36, 0.60)$. For $n = 10$ at probe $\alpha = 0.10$, strict `band_in` gives $S_{\mathrm{rep}} = +0.900$ (9/10 seeds at 1.0, exact $p = 0.0039$); hidden-state knockout at first contact gives 0.000, carry mode gives $+0.700$, and `band_off` is 0.000 in every mode. Five additional confirmation seeds (10–14) all reach 1.0.

Cheapest-sufficient-statistic controls. With stochastic non-reactive partners defined by cooperation rates $q_{\mathrm{good}}/q_{\mathrm{bad}}$, critical-band training produces reactive sensitivity without multi-step memory. In v1 ($q = 0.25/0.75$, $\alpha = 0.44$, $n = 5$), $S = -0.153$ and the knockout value is -0.154. In v2b with 1500 updates, $S = -0.214$ and the knockout result is identical. In v3, masking the partner's last action and adding reward noise $\sigma = 2$ reduces discrimination to -0.018 instead of recruiting memory. These controls show that the learned policy uses a cheaper one-step statistic when that statistic can support the decision; the strict reputation protocol forces the relevant evidence to persist across the phase boundary.

E RUSP-FORM BRIDGE AND BOUNDARY DIAGNOSTICS

Bridge structure. The bridge contains three agents A, B, and X. A observes the B–X interaction in phase 1 and meets X for the first time in phase 2. Preferences enter through a reward-transformation matrix with per-entry observation noise. During training, all three agents are controlled by one shared learning policy; partner heterogeneity comes from the sampled transformation, not from scripts. The hold-out evaluation uses $T = I$ and $\Sigma = 0$ with X replaced by a scripted AllC or AllD type, so that first-contact discrimination is measured against controlled types. Training uses 1500 updates, batch 48, and seeds 0–9 per condition.

Version 1. The preregistered v1 conditions were `selfish`, `original` $U(0, 1)$, `enriched` $U(0, 0.24)$, and `depleted` $U(0.36, 0.60)$, with IID matrix entries and $\sigma_{\text{obs}} = 0.2$. All four conditions produce values near 0.000. Under independent α_{ij}, X's behavior toward B provides no stable information about X's disposition toward A; low-α self-play also fails to bootstrap the heterogeneity required for discrimination. These observations were recorded before the v2 design.

Version 2. The v2 design, frozen before its runs, uses row-constant generosity $\alpha_{ij} = g_i$ for all j and asymmetric uncertainty $\sigma_{\text{own}} = 0.1$, $\sigma_{\text{other}} = 0.3$. The cells are `corr_mix` (half $U(0, 0.24)$, half $U(0.6, 1)$; $M_{\text{crit}} = 0.50$), `corr_original` ($U(0, 1)$; $M_{\text{crit}} = 0.24$), `corr_enriched` ($U(0, 0.24)$; $M_{\text{crit}} = 1.0$ but homogeneous), and `corr_depmix` (off-band mixture; $M_{\text{crit}} = 0$). Outcomes are $+0.026 \pm 0.022$, -0.009, and zero to three decimals for the last two conditions. A-memory knockout reduces `corr_mix` to zero. The exact label-permutation test of `corr_mix` against `corr_enriched` enumerates all $\binom{20}{10} = 184{,}756$ assignments and gives $p = 1.2 \times 10^{-4}$. Predictions P3/P4/P6 pass; P1/P2, which predicted magnitudes of at least $0.10/0.05$, fail.

Full-matrix boundary. A four-agent self-play variant with a full preference matrix and $\sigma_{\text{obs}} = 0.10$ was run under frozen coverage-ordering predictions. Across v1, v2, and v3-style controls, directed sensitivity remains flat at $|S| \leq 0.05$ without a coverage ordering: v1 is underpowered, v2 remains near the reactivity floor ≈ 0.045, and the v3-style counterfactual-probe, own-preference-observation, $\sigma = 0$, single-population, and narrow-band controls do not restore the ordering. The homogeneous learner population does not provide stable behavioral variation that identifies partner type, placing this setting on the closed-identifiability side of the two-gate account.

F CONTEXT-CONDITIONED MODE FAMILY

Figure 1b,c summarizes a companion experiment suite on persistent observable preference contexts. Across a sacrifice game, a threshold public good, and the Coin Game, per-episode α is observed by the agent and held fixed throughout the episode. One shared policy learns an ordered family of context-conditioned decisions, and announced test-time contexts outside the training support select the corresponding behavior. A seven-condition factorization shows that deterministic persistent schedules preserve this mode family, whereas step-wise resampling and meaning-stripped reward noise of matched magnitude do not produce comparable structure. Together, these runs characterize conditioning on an agent's own persistent context across the preference axis; the main experiments characterize partner-conditioned social inference through coverage and identifiability.

G TRAINING DETAILS AND HYPERPARAMETERS

The policy is a single-layer GRU with hidden size 32, an observation skip connection, and separate policy and value heads. Adam uses learning rate 3×10^{-4}. PPO uses clip 0.2, value coefficient 0.5, entropy coefficient 0.01, four epochs per update, gradient-norm clip 1.0, and GAE with $\gamma = 0.96$, $\lambda = 0.95$. The horizon is $T = 32$; reputation and bridge environments split it into $16 + 16$. Scripted-partner experiments (factorization, band shift, G4) use 600 updates and batch 128. Self-play, reputation, and bridge use 1500 updates and batch 48. Probe evaluation uses 192 episodes per S_{trust} estimate and 256 for reputation. The two-gate factorial, G4, reputation, and bridge use 10 seeds per condition; band shift, dose, and the stochastic-partner memory controls use five. Except for the explicitly reported correlation permutation analyses, all reported p-values use exact seed-level tests (sign flip or label permutation), with no episode-level pseudoreplication. Experiments run on a

single workstation GPU; each training run takes minutes, and the full experiment tree reproduces in under two GPU-days.