THE PARTNERS YOU KEEP DECIDE WHAT YOU SEE: WHEN REPUTATION-DRIVEN PARTNER CHOICE RESHAPES EXPOSURE IN LLM SOCIETIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) agents can evaluate peers and select future partners, thereby choosing whose behavior remains observable. This makes behavioral trajectories ambiguous because they can reflect both learning from others and changes in exposure. We formalize this *exposure-selection confound* in a reputation-based partner market with finite partner budgets and observer-specific trust. Cross-role exposure is the primary structural variable. Matched-random retention isolates selection, and forced exploration intervenes on exposure. In a one-shot model with role-balanced reputation supply and role-aligned compatibility, Top-K retention yields no more cross-role exposure than matched-random retention. Lower exposure then preserves a larger stationary behavioral gap. A broadcast cross-role subsidy reverses this ordering once its strength exceeds the compatibility bonus. It reaches the integration oracle when cross-role candidates also have uniformly high reputation. Across 1,749 reference-process runs, activation follows partner scarcity and selection precision rather than population size. The result extends to four groups and a continuous trait. Across ten paired seeds per family, DeepSeek shows a modest but detectable exposure reduction, whereas MiniMax and GPT remain near matched-random retention. DeepSeek also follows its reported trust ranking more closely. Instruction-conditioned agents exhibit strong exposure lock-in in all three families. Perturbation-removal experiments separate persistent individual memory from network-mediated persistence, but do not establish a self-sustaining exposure–behavior loop. These results provide an identification framework and measurable activation conditions for determining when reputation-driven selection changes what agents see.

1 INTRODUCTION

Multi-agent systems built on large language models (LLMs) increasingly operate through selective relationships. Agents cooperate, compete, persuade, exchange reputational information, and choose future partners. Consider a persistent software-engineering swarm in which agents retain reviewers from a reputation-ranked pool. Dropping a reviewer also removes its future critiques, so early scores shape later evidence. A retained relationship is therefore more than a communication edge. It determines whose behavior an agent observes, which evidence it accumulates, and whom it can later influence. Once agents choose these relationships, the interaction network becomes an endogenous evidence channel.

This endogeneity creates a measurement problem. After a negative interaction, an observer may lower its trust and replace the partner, consistent with evidence that dynamic partner choice can sustain cooperation (Santos & Pacheco, 2006; Rand et al., 2011). The replacement changes not only the relationship but also the observer's future sample of social behavior. If many agents make similar choices, aggregate behavior can separate because agents learn differently, because they observe different partners, or because both processes reinforce one another. A behavior trajectory alone cannot distinguish these explanations.

We call this ambiguity the *exposure-selection confound*. Observed behavior mixes what agents learned with whom they continued to observe. Existing adaptive-network models establish 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.

strategy and topology can co-evolve (Pacheco et al., 2006; Gross & Blasius, 2008), while social-network methods distinguish selection from influence. Transferring that distinction to language-agent societies is nontrivial because the agents themselves produce trust reports, partner rankings, and sometimes invalid or inconsistent decisions. A useful test must therefore control topology while preserving the agent decision process. It must also manipulate exposure without imputing behavior and treat structured-output validity as part of measurement.

We develop such a test in a finite-budget partner market. Each agent maintains observer-specific trust and retains K partners. Cross-role exposure is the fraction of retained edges crossing two experimental role treatments. It measures evidence composition while degree is held fixed. Matched-random retention supplies the topology control, and forced exploration intervenes directly on exposure. The primary estimand is the exposure difference between endogenous and matched-random retention. Behavioral separation is a downstream outcome rather than a substitute for the structural test. When a designer values heterogeneous evidence, the same contrast also identifies when an exposure subsidy can correct the externality created by local retention.

The theory identifies when this contrast should appear. Under role-balanced reputation supply and role-aligned compatibility, endogenous Top-K retention produces no more cross-role exposure than matched-random retention. A two-group anchored-averaging reduction maps lower exposure to a larger stationary role gap (DeGroot, 1974; Friedkin & Johnsen, 1990; Raineri et al., 2025). The ordering depends on three observable conditions. Behavior differs by role, trust encodes that difference, and partner choice follows the resulting ranking with sufficiently low noise. The same model also yields a control result. A broadcast cross-role subsidy reverses the ranking after its strength exceeds the same-role compatibility bonus. It reaches the integration oracle when cross-role candidates also have uniformly high reputation.

We test three levels of activation. A deterministic reference process establishes existence across scale and representation changes. Instruction-conditioned agents establish executability. Cross-role exposure falls to 0 in two model families and at most 0.21 in the third. Naive agents reveal a graded activation boundary. One family shows a modest but detectable spontaneous reduction relative to matched-random retention, while the other two remain near that baseline. The activating family also follows its reported trust ranking substantially more faithfully, although all three remain noisier than deterministic Top-K retention. Together, these results locate activation in the joint presence of role-relevant trust and sufficiently faithful selection. Reputation-based retention is therefore neither generically segregating nor generically innocuous.

This paper makes the following principal contributions:

1. **Selection-aware identification.** We formulate the exposure-selection confound and a finite-budget partner-market protocol that distinguishes endogenous selection from topology alone using matched-random retention, then tests the consequence of exposure using forced exploration.
2. **Activation and control theory.** We prove when Top-K retention reduces cross-role exposure and when a broadcast subsidy reverses the ordering above $\gamma^*{=}b$. Systematic scans characterize the boundaries in scarcity and selection noise and extend the mechanism to four groups and a continuous trait.
3. **An empirical activation boundary.** Across three model families, validity-gated experiments show strong activation when role-relevant trust and compliant ranking are supplied, but graded spontaneous activation otherwise. A symmetric audit identifies selection fidelity as the clearest observable cross-family discriminator, while role-specific trust remains weakly identified.

2 Related Work

Adaptive networks, selection, and influence. Adaptive-network research shows that strategy and topology can co-evolve (Pacheco et al., 2006; Gross & Blasius, 2008), dynamic partner choice can sustain cooperation (Santos & Pacheco, 2006; Rand et al., 2011), and reputation can make partner choice depend on interaction history (Ren et al., 2026). Observational network studies show that homophilous selection and peer influence can be generically confounded (Shalizi & Thomas, 2011). DeGroot and Friedkin–Johnsen models instead describe influence on a fixed network (DeG-

root, 1974; Friedkin & Johnsen, 1990; Raineri et al., 2025; Disarò & Valcher, 2024); we use their anchored-averaging logic only to map exposure into a scalar diagnostic. We bring the identification problem into a controlled language-agent market where the retained network is both an outcome and a manipulable channel for future evidence.

Partner selection in LLM societies. Recent work shows that LLM agents form selective ties through personality-based partner choice (Wang et al., 2026), coached relationship formation (Schneider et al., 2025), gossip-mediated indirect reciprocity (Zhu et al., 2026), and homophily-driven community formation (Mehdizadeh & Hilbert, 2025; Ren et al., 2026). Where much of this literature treats selective ties primarily as an outcome, we treat the retained graph as a manipulable evidence channel. Matched-random retention isolates when selection changes exposure, and interventions test how that exposure can be realigned.

Language-agent social evaluation. LLM-agent studies evaluate cooperation, coordination, and strategic behavior in social dilemmas (Park et al., 2023; Piatti et al., 2024; Wu et al., 2024; Akata et al., 2025). SOTOPIA and MultiAgentBench broaden interactive social and coordination evaluation (Zhou et al., 2024; Zhu et al., 2025). These benchmarks address complementary questions. Our focus is a causal control that separates behavioral influence from changes in who remains observable.

Information mediation and control of social learning. Information design studies how a mediator can steer decentralized decisions through signals (Kamenica & Gentzkow, 2011). Recent work extends this idea to sequential social learning by controlling the precision of private evidence (Arghal et al., 2026). Our primary question concerns endogenous information availability. Partner choice changes which behaviors remain observable before any mediator intervenes. The broadcast subsidy is a secondary control result. It adds one public scalar per candidate pair and corrects the resulting exposure externality without forcing edges or changing observed behavior.

Echo chambers and information cocoons. Selective-exposure research documents content homogeneity under platform recommendation and social choice (Bakshy et al., 2015; Flaxman et al., 2016; Barberá et al., 2015; Sunstein, 2017). In an agent society, the retained graph is generated by the same agents whose behavior is being evaluated. We compare endogenous retention with matched-random retention and manipulate cross-role exploration while holding the agent interface fixed. This turns selective exposure from a descriptive network statistic into a testable component of multi-agent evaluation.

3 Formalizing Exposure Selection

3.1 Behavior, trust, relationships, and exposure

Agents $i \in \{1, \ldots, n\}$ act at rounds $t = 0, 1, \ldots$.

Definition 1 (Role treatment)**.** *A role treatment z_i is an externally assigned type or intervention, used to define groups for exposure measurement, not a measured latent personality. In words, it is the experimenter's grouping of who counts as which type.*

Definition 2 (Observer-specific reputation)**.** *Let a_{iqt} be agent i's valid observed response on probe q at time t; invalid responses remain missing, and a behavioral estimate $\widehat{p}_i(t) \in [0, 1]$ aggregates valid responses. For observer i and target j, $r_{ij}(t)$, called reputation in the model and reported trust in the language-agent interface, is a score derived only from information available to i (observed actions, prior interactions). A simple specification is*

$$r_{ij}(t+1) = (1 - \eta_i) r_{ij}(t) + \eta_i g(a_j(t), o_{ij}(t)), \tag{1}$$

where $o_{ij}(t)$ records whether and how i observed j. Public reputation is an ablation in which all observers receive the same score.

The directed partner network has adjacency $A(t)$, with $A_{ij}(t) = 1$ meaning i retains j as a partner. Each agent has a finite partner budget K; a candidate can enter only by replacing a retained partner, preventing unbounded proposals from masquerading as endogenous structure.

Definition 3 (Cross-role exposure)**.** *Suppose agents receive one of two role treatments. For a retained directed network $A(t)$, cross-role exposure is*

$$\omega(t) = \frac{\sum_{ij} A_{ij}(t)\mathbf{1}\{z_i \neq z_j\}}{\sum_{ij} A_{ij}(t)}. \tag{2}$$

In words, $\omega(t)$ is the share of retained partnerships connecting the two roles, measuring evidence composition rather than density.

3.2 Reference process

Let $b_i \in [0,1]$ be a fixed behavioral anchor and $\alpha_i \in (0,1]$ an effective resistance-to-influence parameter. Conditional on a row-stochastic matrix $W(t)$, define

$$\mathbf{p}(t+1) = \mathbf{D}_\alpha \mathbf{b} + (I - \mathbf{D}_\alpha)W(t)\mathbf{p}(t), \qquad \mathbf{D}_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_n). \tag{3}$$

An empty neighborhood uses a self-loop, so every row remains stochastic. In a future residual diagnostic, $W(t)$ may be estimated from partner exposure and reputation. The diagnostic predictor is defined by equation 3; no residual-based result is claimed here.

For a mechanistic simulator, a lagged public attractiveness proxy may be defined by $s_j(t) = \sigma(\beta p_j(t-1))$ and

$$w_{ij}(t) = \frac{A_{ij}(t)s_j(t)}{\sum_k A_{ik}(t)s_k(t)}. \tag{4}$$

We call s_j attractiveness rather than observer-specific reputation. The one-step lag makes the update conditional; at $t = 0$ scores use $\mathbf{p}(0)$.

3.3 Counterfactual network controls

Table 1 summarizes five identification networks (fixed, matched-random, endogenous, forced-exploration, and public-score) plus an exposure-signal condition for alignment. Matched-random-private is the primary topology-only control. It matches population, role balance, degree capacity, edge count, and rounds while randomizing retention. Forced-exploration-private restores cross-role exposure, and endogenous-public tests whether private observer histories are required.

3.4 Global objective, oracle, and social signals

The identification design treats low cross-role exposure as the structural phenomenon. To state what a desirable configuration would be, define a global integration objective

$$\mathcal{U}(X) = \sum_{ij} x_{ij}\left[\bar{r}_{ij} + b\,\mathbf{1}\{z_i = z_j\}\right] + \lambda \sum_{ij} x_{ij}\mathbf{1}\{z_i \neq z_j\}, \tag{5}$$

where $\bar{r}_{ij} \in [r_{\mathrm{lo}}, r_{\mathrm{hi}}] \subseteq [0,1]$ is the planner-visible reputation value and $\lambda \geq 0$ weights cross-role integration. The oracle network maximizes $\mathcal{U}$ subject to each agent retaining exactly K partners. Let m_i be the number of cross-role candidates available to agent i. The objective is separable across agents, so for $\lambda > r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$ every agent fills cross-role capacity and

$$\omega_{\mathrm{oracle}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\min(K, m_i)}{K}, \tag{6}$$

which reduces to $\min(K,m)/K$ when the balanced two-role model gives every agent $m_i = m$. Local agents maximize their own scores and ignore the λ term; the resulting natural–oracle gap is the exposure externality. A social control signal is a broadcast quantity $s_{ij}(t)$ added to local retention utility,

$$\widetilde{u}_{ij}(t) = u_{ij}(t) + \gamma\, s_{ij}(t), \tag{7}$$

where $\gamma \geq 0$ is the signal strength. This paper studies the exposure signal $s_{ij}(t) = \mathbf{1}\{z_i \neq z_j\}$: a single scalar per candidate pair, identical for every agent, derived only from public role labels. It is a Pigouvian subsidy for the externality in equation 5: it does not force edges, does not modify reputation or observed behavior, and reveals no private history.

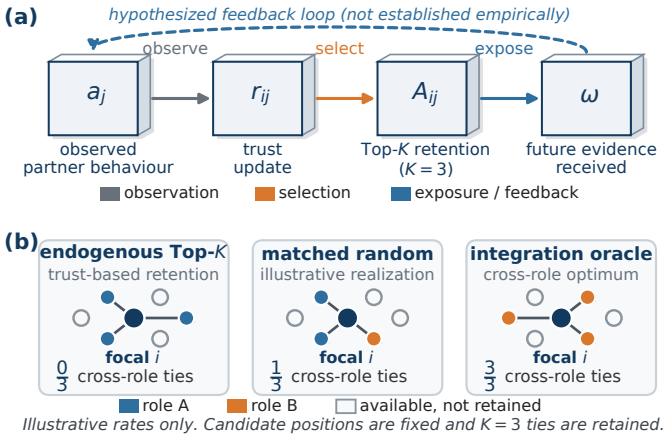


Figure 1: Selection-exposure loop. (a) Behavior shapes trust, retention, and future evidence. The dashed return path is a hypothesized feedback loop not established empirically. (b) At fixed focal agent, candidate pool, and partner budget, the three rules select subsets with different cross-role composition. Open nodes are available candidates not retained. Rates are schematic.

4 The Mechanism and Its Limits

The retention theorems require role-aligned compatibility, role-balanced reputation supply, and strict lexicographic Top-K ranking; later language-agent experiments test these premises.

4.1 From Retention to Future Evidence

A relationship tie changes whom an agent subsequently observes, so selective retention produces *endogenous exposure bias*:

$$\text{interaction outcome} \longrightarrow r_{ij}(t) \longrightarrow A_{ij}(t+1) \longrightarrow \mathcal{N}_i(t+1), o_{ij}(t+1) \longrightarrow a_i(t+1). \quad (8)$$

Trust filtering can improve immediate relational safety while reducing counterevidence, preserving apparent role differences even when no response rule changed. The experiments test successive links in equation 8, but the longitudinal evidence does not establish the full self-sustaining loop. Figure 1 visualizes the loop and compares endogenous Top-K, matched-random, and oracle retention at the same capacity.

4.2 A Minimal Retention-to-Exposure Result

The theorem asks when selective retention changes future exposure relative to the matched-random baseline, independent of outcome semantics.

The general feedback loop is difficult to analyze. We therefore isolate one retention step. The model has two role groups, each with m agents. A focal agent i retains exactly K of the other $2m-1$ agents. Let $a_i \in [0,1]$ denote behavior. Candidate j has reputation r_{ij} and compatibility indicator

$$c_{ij} = \mathbf{1}\{|a_i - a_j| \le \epsilon\}. \quad (9)$$

Agent i ranks candidates by $u_{ij} = r_{ij} + bc_{ij}$ and retains the top K. A matched-random rule retains K distinct candidates uniformly at random.

Candidates have high or low reputation. Role-balanced reputation supply means that the r_{ij} are iid, equal to r_{hi} with probability φ and r_{lo} otherwise, independently of z_j, with uniform random tie-breaking. Thus reputation supply is exchangeable across roles. We impose $r_{\text{hi}} - r_{\text{lo}} > b > 0$ and $1 \le K \le 2m-2$, so compatibility does not reverse the high–low ordering.

Theorem 1 (Top-K retention and cross-role exposure)**.** *Suppose compatibility is role-aligned, so $c_{ij} = 1$ exactly when $z_j = z_i$. Under the conditions above, endogenous Top-K retention never creates more cross-role exposure than matched-random retention:*

$$\mathbb{E}[\omega_{\text{endo}}] \le \mathbb{E}[\omega_{\text{random}}]. \quad (10)$$

The inequality is strict for every admissible K.

Corollary 1 (Exposure and the reduced role gap)**.** *Let $\alpha \in (0,1)$ and $\Delta_b > 0$. If the stationary exposure reduction applies at two rates satisfying $\omega_{\mathrm{endo}} < \omega_{\mathrm{random}}$, then*

$$\Delta^*(\omega_{\mathrm{endo}}) > \Delta^*(\omega_{\mathrm{random}}), \qquad \Delta^*(\omega) = \frac{\alpha\Delta_b}{\alpha + 2(1-\alpha)\omega}. \tag{11}$$

Thus the retention rule can preserve a larger role gap by changing partner composition, even when both rules have the same edge count.

Thus, under role-balanced reputation supply, the compatibility-adjusted Top-K rule yields less cross-role exposure than chance.

The one-step result further assumes the ranking is actually computed and that reported trust predicts rankings; asymmetric reputation supply, unaligned compatibility, or full capacity break it.

4.3 From lock-in to alignment: the exposure subsidy signal

The minimal result diagnoses a gap; the question this paper adds is whether a social signal can close it. Under the objective in equation 5, a planner with $\lambda > r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$ fills cross-role capacity and attains the oracle exposure $\omega_{\mathrm{oracle}} = \min(K, m)/K$ defined in equation 6. The natural network is then strictly below the global optimum, and the difference is the exposure externality ignored by local scores. The broadcast cross-role signal $s_{ij} = \mathbf{1}\{z_j \neq z_i\}$ with strength γ is defined in equation 7.

Theorem 2 (Alignment threshold)**.** *Under the one-shot model with role-balanced reputation supply:*

1. *if $b < \gamma < r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$, then $\mathbb{E}[\omega_{\mathrm{sig}}(\gamma)] > \mathbb{E}[\omega_{\mathrm{endo}}]$ for every admissible K;*
2. *under the same interval, $\mathbb{E}[\omega_{\mathrm{sig}}(\gamma)] > m/(2m-1) = \mathbb{E}[\omega_{\mathrm{random}}]$ for every admissible K;*
3. *for $0 \leq \gamma < b$ the retained composition equals the natural one, whereas at $\gamma = b$ role compatibility cancels and $\mathbb{E}[\omega_{\mathrm{sig}}(b)] = m/(2m-1)$; hence the first reordering threshold is $\gamma^* = b$;*
4. *if $b < \gamma < r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$ and every cross-role candidate has reputation r_{hi}, then $\omega_{\mathrm{sig}}(\gamma) = \omega_{\mathrm{oracle}}$ for every admissible K; otherwise the nonnegative residual $\omega_{\mathrm{oracle}} - \mathbb{E}[\omega_{\mathrm{sig}}(\gamma)]$ is an information component.*

In other words, the signal works only once it outweighs the same-role compatibility bonus.

Corollary 2 (Intervention efficacy)**.** *For every admissible K and $b < \gamma < r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$, the oracle gap shrinks:*

$$\mathrm{Gap}(0) - \mathrm{Gap}(\gamma) = \mathbb{E}[\omega_{\mathrm{sig}}(\gamma) - \omega_{\mathrm{endo}}] > 0, \qquad \mathrm{Gap}(s) = \omega_{\mathrm{oracle}} - \mathbb{E}[\omega(s)]. \tag{12}$$

Appendix B.3 gives complete proofs of the retention, role-gap, alignment, and efficacy results.

The signal is deliberately minimal: one public scalar per candidate pair, whose sign relative to the threshold matters; the threshold $\gamma^*{=}b$ is a property of the finite lexicographic ranking.

Under soft selection (logit choice, temperature τ) the threshold degrades continuously rather than vanishing, from 0.000 at $\tau \leq 0.05$ to 0.43 at $\tau{=}3$; the natural$<$random ordering is preserved at low temperature and melts as choice noise grows, so the theorems describe the zero-noise limit of a choice-model family.

4.4 Causal predictions and falsification

Theorem 2 and Corollary 2 yield five causal predictions: (I) endogenous-private below matched-random-private exposure, (II) lower exposure with greater behavioral separation, (III) forced exploration restoring exposure and attenuating separation, (IV) endogenous-public weakening the mechanism under private histories, and (V) an above-threshold signal raising exposure toward the oracle. This paper directly reports (I), (II), and (V); (III)–(IV) remain protocol extensions rather than completed confirmatory analyses. If endogenous and matched-random retention match, selection is not operative. The inference unit is a complete seeded run, with paired contrasts and bootstrap.

Table 1: Partner-market conditions.

Condition	Info	Retention	Role
Fixed	Private	None	History
Matched-random	Private	Uniform	Control
Endogenous	Private	Ranked Top-K	Mechanism
Forced explore	Private	Top-K + cross-role	Intervention
Public score	Public	Ranked Top-K	Ablation
Exposure signal	Public	Top-K + subsidy	Alignment

5 Evidence and Causal Identification

Evidence comes in three tiers: deterministic reference-process runs, mean-field diagnostics, and the prespecified partner-market study. Every run has a unique identifier; invalid outputs remain missing, and a 95% structured-output gate is required before behavioral interpretation. The LLM validity pilot (four agents, two rounds, three seeds per model) checks pipeline validity.

The first five conditions in Table 1 identify lock-in; the sixth tests alignment. Let E denote endogenous-private retention and R denote matched-random-private retention. The primary estimand is

$$\tau_\omega = \mathbb{E}[\omega_E - \omega_R]\,. \tag{13}$$

Exposure lock-in requires $\tau_\omega < 0$. With Δ_s the absolute between-role difference in mean behavior, the downstream contrast is

$$\tau_\Delta = \mathbb{E}[\Delta_E - \Delta_R]\,. \tag{14}$$

Exposure selection is identified by $\tau_\omega < 0$. The distinct downstream implication is tested by τ_Δ: a full exposure-mediated behavioral result requires $\tau_\omega < 0$ and $\tau_\Delta > 0$, but the structural result does not. Forced-exploration-private and endogenous-public are protocol-level diagnostics not included in the completed ten-seed core contrast.

The control question adds a sixth condition, endogenous-private with an exposure signal at strength γ above threshold. Using the same-capacity oracle exposure from equation 6, we evaluate the intervention through the oracle gap and its reduction,

$$\mathrm{Gap}(s) = \omega_{\mathrm{oracle}} - \mathbb{E}[\omega_s], \qquad \mathrm{Eff}(\gamma) = \mathrm{Gap}(0) - \mathrm{Gap}(\gamma) = \mathbb{E}[\omega_{\mathrm{sig}}(\gamma) - \omega_E], \tag{15}$$

for $s \in \{E, \mathrm{sig}(\gamma)\}$. Theorem 2 predicts $\mathrm{Eff}(\gamma) > 0$ above threshold; a null result would mean the local wedge is not the binding constraint. Section 6.6 reports estimates from live runs that pass the strict partner-market inclusion gate.

6 Results

6.1 Measurement validity and calibration

Table 3 shows that all nine runs in the validity pilot passed the prespecified output gate. The pilot included three model families and three runs per family, with all 72 decision slots valid. Separately, all 60 calibration measurements across the three families were valid, and extreme persona interventions produced complete choice separation with $\widehat{p}$ at 0 or 1, revealing prompt-conditioned policy control rather than a continuous personality scale. Invalid outputs are recorded as missing and excluded from behavioral conclusions.

6.2 Deterministic boundary evidence

The original adaptive simulator fills capacity exactly, so sparse topology is never informative and no tested condition shows a cooperation transition; every mean-field scan has one stable root at each β, and a 20,000-distribution random search found no multi-root example.

6.3 DETERMINISTIC SIGNAL EVIDENCE

A 40-agent, two-role, capacity-K=5 pilot tests the alignment layer with the broadcast subsidy γ over ten seeds. Exposure is 0.2470 under natural retention, below the matched-random baseline 0.5065; it stays below 0.47 for $\gamma \leq 0.05$, jumps to 0.8690 at $\gamma = 0.08$, and saturates at 1.0000 for $\gamma \geq 0.10$, the finite-lexicographic step predicted by Theorem 2 (Figure 10a). Applied to the language-agent market over four seeds, no threshold jump appears: exposure stays within 0.54–0.67 for every family and $\gamma \in \{0.1, \ldots, 0.5\}$, never below matched-random retention (Figure 10b).

6.4 SYSTEMATIC EVIDENCE ON BOUNDARIES AND GENERALITY

A generalized reference process sweeps Scale $\times$ Mechanism $\times$ Generality over N=16 to 128 in 1,749 runs.

At the baseline N=32, K=5, w=0.8, retention fully segregates: exposure 0.000 vs 0.505 random and 1.000 oracle, separation 0.360 vs 0.086, and planner-objective shortfall 318 vs 161 vs 0 (endogenous, random, oracle).

Boundaries. Lock-in persists for $K/(N-1) \leq 0.39$ and weakens near 0.52 (exposure 0.062); the ordering appears for any $w > 0$ and reaches full long-horizon segregation by $w \gtrsim 0.3$ (at $w = 0$ exposure is 0.512, near random); reputation noise up to σ=0.4 does not unlock exposure, while increasing softmax temperature melts lock-in monotonically (0.000 at $\tau \leq 0.1$ to 0.449 at $\tau = 3$).

Scale and generality. The boundary follows the relative budget $K/(N-1)$: exposure is 0.000 at 0.3 and 0.17–0.22 at 0.6 for every tested N. Four groups and a continuous trait reproduce the ordering, with exposure-diversity collapse under the trait; Appendix D.1 reports the full diagnostic grid.

6.5 MECHANISM EXECUTABILITY

An instruction-conditioned arm reveals roles and requires same-role trust ranking. Cross-role exposure reaches 0.000 in DeepSeek and GPT and ≤ 0.21 in MiniMax, versus 0.38–0.71 in a roles-only control over seven seeds. MiniMax-M3 is an explicit exception (Appendix D.4), while a planner finds the signal threshold (Appendix D.3). The required decision infrastructure combines role-relevant trust with compliant selection. Instruction supplies both explicitly, whereas naive agents do not consistently exhibit both.

6.6 LANGUAGE-AGENT SPONTANEITY

Across ten paired seeds, endogenous exposure is 0.45, 0.49, and 0.55 for DeepSeek, MiniMax, and GPT, versus matched-random values of 0.54, 0.55, and 0.54. DeepSeek's paired estimate is -0.09, with a 95% interval $[-0.15, -0.04]$ that excludes zero. The MiniMax interval (-0.06 $[-0.14, +0.02]$) and GPT interval ($+0.01$ $[-0.04, +0.07]$) contain zero. Retention therefore produces a detectable exposure reduction in one family but not in the other two (Appendix Table 4). Among valid behavior pairs, τ_Δ is -0.10 $[-0.22, +0.03]$ for DeepSeek (n=9), $+0.15$ $[+0.03, +0.26]$ for MiniMax, and 0.00 for GPT. The DeepSeek result identifies changed exposure, not behavioral amplification.

DeepSeek shows higher Top-K overlap (0.82 vs 0.65–0.68, with disjoint clustered intervals), and trust-only reconstruction matches its exposure (0.442 vs 0.446), unlike MiniMax or GPT. Same-role trust bias is not significant under run-level clustering in any family, so the role signal remains plausible but not separately identified. Appendix Table 5 reports the audit and marks seed-level associations as exploratory.

Figure 2 locates the cross-family boundary. Complete-set Top-3 agreement of 37–65% places preference noise near the reference process's random-exposure regime. Populations of 8, 16, 32, and 64 reproduce the endogenous contrasts, while instructed exposure remains 0.000. With the roles-only arm, this localizes activation to role-relevant trust and compliant selection.

Exploratory intervention scan. In the four-seed cells of Table 2, no intervention activates lock-in in all families: budget stays near random, assortative and Watts–Strogatz ring-rewire initializations

lower exposure only in some families, and role-payoff shows no consistent shift; the Barabási–Albert initialization raises exposure descriptively.

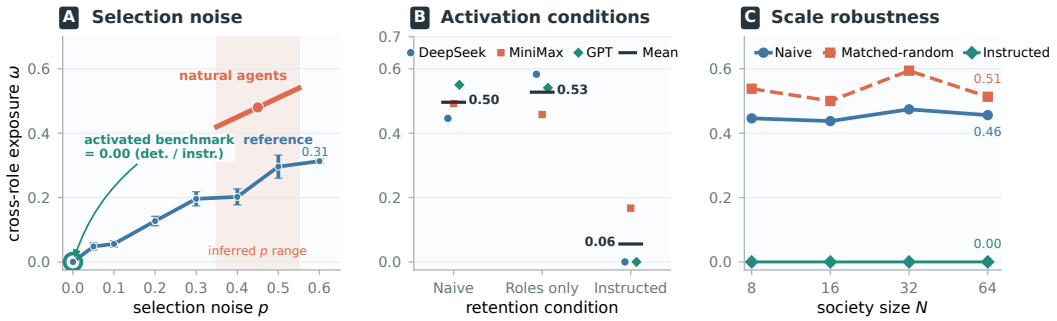


Figure 2: Activation boundary. (a) Increasing selection noise progressively melts lock-in and restores exposure toward the matched-random regime. The shaded band marks the natural-agent fidelity range. (b) Instruction, but not role labels alone, activates lock-in. (c) The regimes remain separated from $N{=}8$ to 64.

6.7 Longitudinal perturbation removal

We tested whether a temporary behavioral difference persists after its source is removed. GPT did not respond to the prompt induction. For DeepSeek and MiniMax, endogenous exposure lies below matched-random retention ($\tau_{\omega}^{\mathrm{E-M}} = -0.072\ [-0.144, -0.010]$), but the difference is driven by the random realization and every behavioral-persistence contrast contains zero. A twenty-paired-seed MiniMax-M3 action shock confirms action-sensitive trust and individual memory, while post-removal exposure matches random retention. Together, these experiments separate evidence for persistent individual memory from evidence for an exposure-mediated loop. The former is supported, whereas the latter is not established. Appendix D.5 gives the protocols and full estimates.

7 Discussion and Limitations

Our results identify three conditions under which reputation-driven retention changes exposure: behavior reveals roles, trust captures them, and selection follows trust. The reference process builds in this alignment, and instruction-conditioned agents execute it strongly. Among naive agents, the highest-fidelity family shows the clearest reduction, while the others remain near matched-random retention. Behavioral gaps do not identify learning when composition changes. Studies should therefore measure exposure, include a topology control, and intervene before attributing social influence.

Retention turns local ranking into information allocation. Myopic choices determine which counterfactual behaviors remain observable, connecting trust to macro-level persistence. Exposure interventions distinguish informational lock-in from behavioral rigidity. Aggregate results cannot resolve path dependence in individual tie replacement, which requires longer runs.

The boundary separates two regimes. Weak role signals and noisy choice keep exposure near random, whereas salient compatibility and precise selection activate lock-in. Forced exploration guarantees mixed evidence. A broadcast subsidy is softer when compatibility is the binding wedge. Cross-role exposure measures evidence availability, not attention, uptake, or welfare. Anchored averaging is a scalar diagnostic, not a model of multidimensional reasoning. Role labels and the planner objective remain design choices. Larger preregistered markets should vary trust relevance and compliance over longer horizons.

These findings suggest a practical diagnostic. Test whether role-linked behavior predicts observer-specific trust and whether trust predicts retained partners. Only if both links hold should exposure interventions change trajectories. This makes echo-chamber concerns falsifiable, locates failure before costly longitudinal experiments, and distinguishes a missing signal from a failure to act on it.

REFERENCES

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Raghu Arghal, Kevin He, Shirin Saeedi Bidokhti, and Saswati Sarkar. Steering the herd: A framework for LLM-based control of social learning. In *International Conference on Learning Representations (ICLR 2026)*, pp. 34584–34621, 2026.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, 2015. doi: 10.1126/science.aaa1160.

Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542, 2015. doi: 10.1177/0956797615594620.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974. doi: 10.1080/01621459.1974.10480137.

Giorgia Disarò and Maria Elena Valcher. Balancing homophily and prejudices in opinion dynamics: An extended Friedkin–Johnsen model. *Automatica*, 166:111711, 2024. doi: 10.1016/j.automatica.2024.111711.

Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016. doi: 10.1093/poq/nfw006.

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990. doi: 10.1080/0022250X.1990.9990069.

Thilo Gross and Bernd Blasius. Adaptive coevolutionary networks: A review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008. doi: 10.1098/rsif.2007.1229.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Aliakbar Mehdizadeh and Martin Hilbert. Homophily-induced emergence of biased structures in LLM-based multi-agent AI systems. *Social Network Analysis and Mining*, 15(1):108, 2025. doi: 10.1007/s13278-025-01535-7.

Jorge M. Pacheco, Arne Traulsen, and Martin A. Nowak. Coevolution of strategy and structure in complex networks with dynamical linking. *Physical Review Letters*, 97(25):258103, 2006. doi: 10.1103/PhysRevLett.97.258103.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22, New York, NY, USA, 2023. ACM. doi: 10.1145/3586183.3606763.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 111715–111759, 2024.

Roberta Raineri, Lorenzo Zino, and Anton Proskurnikov. FJ-MM: The Friedkin–Johnsen opinion dynamics model with memory and higher-order neighbors. *European Journal of Control*, 86: 101306, 2025. doi: 10.1016/j.ejcon.2025.101306.

David G. Rand, Samuel Arbesman, and Nicholas A. Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108 (48):19193–19198, 2011. doi: 10.1073/pnas.1108243108.

Siyue Ren, Wanli Fu, Xinkun Zou, Chen Shen, Yi Cai, Chen Chu, Zhen Wang, and Shuyue Hu. Reputation as a solution to cooperation collapse in LLM-based MASs. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '26)*, pp. 245–253, Richland, SC, 2026. International Foundation for Autonomous Agents and Multiagent Systems. doi: 10.65109/UEHN4980.

Francisco C. Santos and Jorge M. Pacheco. A new route to the evolution of cooperation. *Journal of Evolutionary Biology*, 19(3):726–733, 2006. doi: 10.1111/j.1420-9101.2005.01063.x.

Philipp J. Schneider, Lin Tian, and Marian-Andrei Rizoiu. Learning to make friends: Coaching LLM agents toward emergent social ties. In *NeurIPS 2025 Workshop on Scaling Environments for Agents (SEA)*, 2025. doi: 10.48550/arXiv.2510.19299.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011. doi: 10.1177/0049124111404820.

Cass R. Sunstein. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, Princeton, NJ, 2017. doi: 10.1515/9781400884711.

Tao Wang, Hsiang-Ling Chiu, Chihang Wei, Zhonghao Hou, and Yang Xiu. Not birds of a feather: Personality-based partner selection in LLM agents. *arXiv preprint arXiv:2607.19785*, 2026.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. Shall we team up: Exploring spontaneous cooperation of competing LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5163–5186, 2024. doi: 10.18653/v1/2024.findings-emnlp.297.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 8580–8622, 2025. doi: 10.18653/v1/2025.acl-long.421.

Shuhui Zhu, Yue Lin, Shriya Kaistha, Wenhao Li, Baoxiang Wang, Hongyuan Zha, Gillian K. Hadfield, and Pascal Poupart. Talk, judge, cooperate: Gossip-driven indirect reciprocity in self-interested LLM agents. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, 2026.

AI USE STATEMENT

This required section does not count toward the page limit.

In this work, we used generative AI tools to help develop and refine the conceptual framework and theoretical models; formulate and check mathematical claims; assist in drafting proofs; refine hypotheses, methodology, and experimental designs; implement simulation and language-agent experiments; clean and reformat experimental logs; support qualitative and quantitative analysis; and interpret results. Language models also generated the agent outputs studied in the language-agent experiments. Non-LLM reference-process data were produced by the reported simulation code rather than substituted with model-generated records. We did not use generative AI for translation or transcription, and survey or interview design was not applicable to this work.

Additionally, we used generative AI tools for literature triage and summarization, research brainstorming, code editing and debugging, suggesting experimental parameters, creating and revising scientific figures, drafting and restructuring manuscript text, improving readability, formatting references, and proposing titles and section structure. The authors reviewed all AI-assisted work: theoretical statements were checked against their explicit assumptions and complete proofs; simulation

and analysis code was executed against retained configurations and logs; structured model outputs were validated before analysis; figures were regenerated from retained results; and the manuscript was compiled and inspected. We take responsibility for the final content of this work, including all text, claims, code, data analyses, and artifacts produced with the aid of generative AI.

APPENDIX ROADMAP

The appendix has five parts. Appendix A introduces the network-science and social-learning terminology and collects notation. Appendix B contains the reference dynamics, exposure reduction, and complete one-shot proofs. Appendix C reports structured-output validity and behavioral calibration. Appendix D presents deterministic diagnostics, long-horizon topology and robustness, and LLM-planner threshold discovery. Appendix E separates claim types and records the configurations, prompts, failure policy, confirmatory design, and retained analysis artifacts needed for reproduction.

Anonymous interactive companion. For anonymous review, we provide an interactive companion at `https://dainty-dango-940f35.netlify.app/`. It visualizes the exposure-selection mechanism, organizes the main theoretical and empirical evidence, and provides the anonymous manuscript and reproducibility archive for download. The site uses no cookies, analytics, or visitor tracking.

A PRELIMINARIES AND NOTATION

This section collects the network-science and social-learning concepts used in the paper. The definitions are operational: they describe what is measured or manipulated in our partner market, rather than assigning latent social traits to language-model agents.

Endogenous networks, selection, and influence. An *endogenous network* is one whose edges are outcomes of the agents' decisions rather than fixed by the experimenter. *Selection* changes who is connected to whom; *influence* changes behavior conditional on those connections. Because a retained partner is also a source of later observations, selection changes the evidence on which influence operates. A behavioral trajectory may therefore change because agents learned differently or because they observed different partners. We call this ambiguity the *exposure-selection confound*.

Assortativity, homophily, and selective exposure. *Assortative mixing* is a network pattern in which ties occur more often between similar nodes than under a stated baseline. *Homophily* is a selection mechanism that can produce such a pattern through preference for similar partners. We use the narrower term *role-aligned compatibility* when the modeled score favors candidates receiving the same experimental role treatment; this does not imply a stable identity or human-like preference. *Selective exposure* is the resulting imbalance in the information sources that remain observable. Our statistic ω measures one specific form: the share of retained directed ties crossing role treatments.

Lock-in and counterfactual controls. *Lock-in* denotes a feedback regime in which current behavior affects trust and retention, the retained network shapes subsequent exposure, and that exposure can preserve the conditions that generated the network. It need not be irreversible or welfare-reducing. A *topology-matched counterfactual* holds the partner budget and edge count fixed while replacing endogenous retention by a comparison rule. Matched-random retention therefore isolates selection from network density; forced exploration intervenes on exposure itself.

Externality, oracle, and broadcast signal. An *exposure externality* arises when an agent's local retention score does not account for the social value assigned to cross-role evidence. The *integration oracle* is the exact-K network maximizing the stated planner objective; it is a benchmark, not a claim that maximal mixing is always desirable. A *broadcast signal* adds the same public, non-falsifying role-based adjustment to every agent's candidate scores. Its purpose is to align decentralized retention with that explicit objective without forcing an edge.

Notation table.

Symbol	Meaning
$i, j; t$	Observer and candidate-agent indices; interaction round.
$n; m_i; m; K$	Population size; cross-role candidates available to i; their common value in the balanced two-role model; exact partners retained per agent.
z_i	Experimentally assigned role treatment of agent i.
$a_i, \widehat{p}_i(t)$	Agent behavior in the one-shot model; empirical behavioral estimate at round t.
$\mathbf{p}(t); \mathbf{b}$	Vector of behavioral states; vector of fixed behavioral anchors.
$\alpha_i, \mathbf{D}_\alpha$	Resistance-to-influence parameter and its diagonal matrix.
$A_{ij}(t); W(t)$	Retained directed-edge indicator; row-stochastic influence matrix.
$r_{ij}(t); \bar{r}_{ij}$	Observer-specific reputation; planner-visible reputation.
$o_{ij}(t); \eta_i$	Information through which i observes j; reputation-update rate.
$c_{ij}; b$	Compatibility indicator; scalar same-role compatibility bonus (distinct from anchor vector $\mathbf{b}$).
$r_{\mathrm{lo}}, r_{\mathrm{hi}}; \varphi$	Low and high reputation levels; probability of high reputation in the one-shot model.
$\omega(t); \omega_{\mathrm{oracle}}$	Cross-role exposure; maximum exposure under the stated exact-capacity planner objective.
$X = (x_{ij}); \mathcal{U}(X)$	Planner's directed edge decisions and global integration objective.
$s_{ij}(t); \gamma; \lambda$	Broadcast exposure signal; signal strength; planner weight on cross-role integration.
$\beta; \tau$	Attractiveness slope in the reference process; soft-selection temperature.
$\Delta^*(\omega); \Delta_b$	Stationary role gap at exposure ω; difference between role-specific anchors.

Condition	DeepSeek	MiniMax	GPT
K=2 / K=4 (budget)	0.47; 0.52	0.48; 0.51	0.52; 0.55
assortative start	0.40	0.34	0.33
role-dependent payoff	0.44	0.58	0.54
WS ring-rewire start	0.44	0.41	0.29
BA preferential-attachment start	0.65	0.58	0.47
matched-random ($K = 3$)	0.56	0.55	0.54

Table 2: Exploratory intervention scan (mean gate-passing endogenous exposure; four attempted seeds per cell, except MiniMax BA n=3 after gating). WS and BA denote finite graph-generation algorithms, not estimated macroscopic network classes.

B Theory and Proofs

B.1 Reference Dynamics and Boundary Results

The following results support the reference process; they are included for completeness.

Proposition 1 (Well-posedness and invariance)**.** *Suppose $W(t)$ is row-stochastic for every t, $\alpha_i \in (0, 1]$, and $\mathbf{b}, \mathbf{p}(0) \in [0, 1]^n$. Then equation 3 defines a unique trajectory and $\mathbf{p}(t) \in [0, 1]^n$ for all t.*

Proof. Let $q_i(t) = \sum_j w_{ij}(t) p_j(t)$. Row stochasticity and $\mathbf{p}(t) \in [0, 1]^n$ imply

$$0 \le q_i(t) \le \sum_j w_{ij}(t) = 1. \tag{16}$$

Hence

$$p_i(t + 1) = \alpha_i b_i + (1 - \alpha_i) q_i(t) \in [0, 1]. \tag{17}$$

Induction from $\mathbf{p}(0) \in [0, 1]^n$ proves invariance. For a specified sequence $\{W(t)\}_{t \ge 0}$, the right-hand side is a single-valued affine map at each t, so the trajectory is unique. □

Theorem 3 (Fixed-regime equilibrium)**.** *Let W be a fixed row-stochastic matrix and let $\underline{\alpha} = \min_i \alpha_i > 0$. The map in equation 3 is a contraction in $\| \cdot \|_\infty$ with factor at most $q = 1 - \underline{\alpha} < 1$. It has the unique equilibrium*

$$\mathbf{p}^*(W) = \left[I - (I - \mathbf{D}_\alpha) W \right]^{-1} \mathbf{D}_\alpha \mathbf{b}, \tag{18}$$

and $\| \mathbf{p}(t) - \mathbf{p}^*(W) \|_\infty \le q^t \| \mathbf{p}(0) - \mathbf{p}^*(W) \|_\infty$.

Proof. Define $T(x) = \mathbf{D}_\alpha \mathbf{b} + (I - \mathbf{D}_\alpha)Wx$. Since $\|W\|_\infty = 1$ and $\|I - \mathbf{D}_\alpha\|_\infty = 1 - \underline{\alpha} = q$,

$$\|T(x) - T(y)\|_\infty \le q\|x - y\|_\infty. \tag{19}$$

Banach's fixed-point theorem gives a unique $\mathbf{p}^*$ and $\|\mathbf{p}(t) - \mathbf{p}^*\|_\infty \le q^t \|\mathbf{p}(0) - \mathbf{p}^*\|_\infty$. Moreover, with $M = (I - \mathbf{D}_\alpha)W$, $\|M\|_\infty \le q < 1$, so

$$(I - M)^{-1} = \sum_{k=0}^{\infty} M^k, \qquad \mathbf{p}^* = (I - M)^{-1}\mathbf{D}_\alpha \mathbf{b}. \tag{20}$$

□

Proposition 2 (Fixed-regime anchor sensitivity)**.** *For a fixed row-stochastic W and two anchor vectors $\mathbf{b}, \widetilde{\mathbf{b}} \in [0,1]^n$, their equilibria satisfy $\|\mathbf{p}^*(W;\mathbf{b}) - \mathbf{p}^*(W;\widetilde{\mathbf{b}})\|_\infty \le (\alpha_{\max}/\alpha_{\min})\|\mathbf{b} - \widetilde{\mathbf{b}}\|_\infty$.*

Proof. Let $M = (I - \mathbf{D}_\alpha)W$. From equation 18,

$$\mathbf{p}^*(W;\mathbf{b}) - \mathbf{p}^*(W;\widetilde{\mathbf{b}}) = (I - M)^{-1}\mathbf{D}_\alpha(\mathbf{b} - \widetilde{\mathbf{b}}). \tag{21}$$

Because $\|M\|_\infty \le 1 - \alpha_{\min}$,

$$\|(I - M)^{-1}\|_\infty \le \sum_{k=0}^{\infty} \|M\|_\infty^k \le \frac{1}{\alpha_{\min}}. \tag{22}$$

Multiplying by $\|\mathbf{D}_\alpha\|_\infty = \alpha_{\max}$ proves the claim. □

Corollary 3 (Finite-freeze convergence)**.** *If $W(t) = W_f$ for all sufficiently large t, the trajectory converges geometrically to $\mathbf{p}^*(W_f)$.*

Proof. Choose t_f such that $W(t) = W_f$ for $t \ge t_f$. Applying the fixed-regime contraction from time t_f gives

$$\|\mathbf{p}(t) - \mathbf{p}^*(W_f)\|_\infty \le q^{t - t_f}\|\mathbf{p}(t_f) - \mathbf{p}^*(W_f)\|_\infty, \qquad t \ge t_f. \tag{23}$$

□

Proposition 3 (Summable-variation tracking)**.** *Let T_t denote the affine map at time t, with a common contraction factor $q < 1$, and let x_t^* be its unique fixed point. If $\sum_t \|x_{t+1}^* - x_t^*\|_\infty < \infty$, then $\mathbf{p}(t)$ converges.*

Proof. Set $e_t = \|\mathbf{p}(t) - x_t^*\|_\infty$ and $d_t = \|x_{t+1}^* - x_t^*\|_\infty$. Since $T_t(x_t^*) = x_t^*$,

$$e_{t+1} \le q e_t + d_t, \qquad e_t \le q^t e_0 + \sum_{k=0}^{t-1} q^{t-1-k} d_k. \tag{24}$$

For any $J < t$, split the sum at J:

$$\sum_{k=0}^{t-1} q^{t-1-k} d_k \le q^{t-J} \sum_{k=0}^{J-1} q^{J-1-k} d_k + \sum_{k=J}^{\infty} d_k. \tag{25}$$

First let $t \to \infty$ and then $J \to \infty$; summability yields $e_t \to 0$. Also, for $s > t$, $\|x_s^* - x_t^*\|_\infty \le \sum_{k=t}^{s-1} d_k$, so (x_t^*) is Cauchy. If $x_t^* \to x_\infty^*$, then $\|\mathbf{p}(t) - x_\infty^*\|_\infty \le e_t + \|x_t^* - x_\infty^*\|_\infty \to 0$. □

Proposition 4 (Non-uniqueness under a dead band)**.** *Suppose an absent tie is added only when its score exceeds c, a present tie is removed only when its score is below d, and its status is unchanged for scores in $[d, c]$, where $0 < d < c < 1$. There exists a two-agent anchored system with distinct stationary topologies but the same stationary state.*

Proof. Take $\mathbf{b} = (1,1)^\top$ and use a self-loop whenever an agent has no neighbor. For either the disconnected self-loop matrix $W_0 = I_2$ or the connected matrix $W_1 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$,

$$\mathbf{D}_\alpha \mathbf{b} + (I - \mathbf{D}_\alpha) W_k \mathbf{b} = \mathbf{b}, qquadk \in \{0,1\}. \tag{26}$$

Thus both topologies have stationary state $\mathbf{p}^* = \mathbf{b}$. Choose $d = 0.6$, $c = 0.8$, and score $s = \sigma(1) \approx 0.7311 \in [d, c]$. Here we set the attractiveness slope to $\beta = 1$ and evaluate the score at $p = 1$, so $s = \sigma(1)$. The dead-band rule preserves an absent tie under W_0 and a present tie under W_1. Hence the stationary state is identical while the stationary topologies differ. □

For the annealed mean-field diagnostic, let a finite type distribution be $\mu = \{(\mu_\ell, \alpha_\ell, b_\ell)\}_{\ell=1}^L$, where $\mu_\ell > 0$, $\sum_\ell \mu_\ell = 1$, and $(\alpha_\ell, b_\ell) \in [0,1]^2$. Define

$$p_\ell(m) = \alpha_\ell b_\ell + (1 - \alpha_\ell) m, qquadF(m; \beta, \mu) = \frac{\sum_{\ell=1}^L \mu_\ell \sigma(\beta p_\ell(m)) p_\ell(m)}{\sum_{\ell=1}^L \mu_\ell \sigma(\beta p_\ell(m))}. \tag{27}$$

Proposition 5 (Mean-field existence)**.** *For every finite $\beta \geq 0$, Equation $m = F(m; \beta, \mu)$ has at least one solution in $[0,1]$.*

Proof. For $m \in [0,1]$, $p_\ell(m) \in [0,1]$. Since $\sigma(\beta p_\ell(m)) > 0$, the map in equation 27 is continuous and is a convex combination of the values $p_\ell(m)$; hence $F([0,1]) \subseteq [0,1]$. Let $G(m) = F(m) - m$. Then $G(0) = F(0) \geq 0$ and $G(1) = F(1) - 1 \leq 0$. The intermediate value theorem gives an $m^* \in [0,1]$ with $G(m^*) = 0$. □

Residual analysis. The reference process is evaluated by predictive error, not by forcing agents to follow it. Given behavioral estimates $\widehat{p}_i(t)$ and exposure-derived $\widehat{W}(t)$, fit effective anchors and resistance parameters on training rounds, then evaluate

$$\epsilon_i(t+1) = \widehat{p}_i(t+1) - \left[\widehat{\alpha}_i \widehat{b}_i + (1 - \widehat{\alpha}_i) \sum_j \widehat{w}_{ij}(t) \widehat{p}_j(t)\right] \tag{28}$$

on held-out rounds; the confirmatory study will test whether residuals depend on negative events, recency, partner identity, verbal explanations, network turnover, and model family, making deviations substantive evidence rather than failed theorem validation.

B.2 Exposure Reduction and Theory Figure

In a two-role reduction, let $\Delta(t)$ be the difference between group mean behaviors, $h(t)$ the same-role retained-tie fraction, and χ the probability that an available cross-role tie yields an observation. The realized cross-role mixing rate is $\omega(t) = (1 - h(t))\chi$; $\chi = 1$ recovers the purely structural exposure measure. Under an anchored exposure update,

$$\Delta(t+1) = \alpha \Delta_b + (1 - \alpha)(1 - 2\omega(t))\Delta(t), \qquad \Delta_b = b_+ - b_-. \tag{29}$$

In this two-group reduction, Figure 3 plots the stationary relationship.

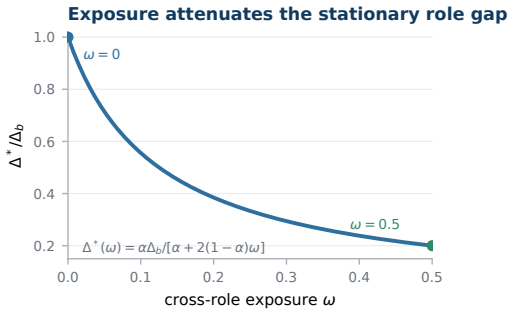


Figure 3: Exposure-to-gap curve.

B.3 Proofs of the One-Shot Theorems

Derivation of equation 6. For any cross-role candidate c and same-role candidate s, the planner scores satisfy

$$\big(\bar{r}_{ic} + \lambda\big) - \big(\bar{r}_{is} + b\big) \geq r_{\mathrm{lo}} + \lambda - r_{\mathrm{hi}} - b > 0. \tag{30}$$

Thus every cross-role candidate precedes every same-role candidate in the planner's ranking. Exact capacity K gives agent i cross count $\min\{K, m_i\}$; averaging over agents yields equation 6, and the balanced case has $m_i = m$.

For a score vector $v \in \mathbb{R}^{2m-1}$, let $S_K(v)$ be its Top-K candidate set after symmetric random tie-breaking, and let C denote the set of m cross-role candidates.

Lemma 1 (Top-K cross-count monotonicity)**.** *Let $v, v' \in \mathbb{R}^{2m-1}$. If $v'_j \geq v_j$ for $j \in C$ and $v'_j \leq v_j$ for $j \notin C$, then there is a coupling of the tie-breaking variables such that*

$$|S_K(v') \cap C| \geq |S_K(v) \cap C|. \tag{31}$$

Proof. Couple ties by adding the same independent continuous perturbation δ_j to corresponding coordinates, and write $\widetilde{v}_j = v_j + \delta_j$ and $\widetilde{v}'_j = v'_j + \delta_j$; then take the perturbation scale to zero. Suppose, for contradiction, that the cross count decreases. Then some $c \in C$ is selected under v but not v', while some $s \notin C$ is not selected under v but is selected under v'. Unique ranking under the perturbation implies

$$\widetilde{v}_c > \widetilde{v}_s, \qquad \widetilde{v}'_s > \widetilde{v}'_c. \tag{32}$$

The common perturbation preserves $\widetilde{v}'_c \geq \widetilde{v}_c$ and $\widetilde{v}'_s \leq \widetilde{v}_s$, giving a contradiction. Taking the zero-perturbation limit preserves equation 31 under symmetric tie-breaking. □

Proof of Theorem 1. At $b = 0$, the scores r_{ij} are exchangeable across all $2m-1$ candidates by the role-balanced reputation-supply assumption. Symmetric tie-breaking therefore gives each candidate inclusion probability $K/(2m-1)$, whence

$$\mathbb{E}|S_K(r) \cap C| = \frac{Km}{2m-1}, qquad\mathbb{E}[\omega_{\mathrm{random}}] = \frac{m}{2m-1}. \tag{33}$$

The endogenous score vector is

$$u_j = r_{ij} + b\mathbf{1}\{j \notin C\}. \tag{34}$$

Relative to r, only same-role coordinates increase. Lemma 1, applied with the roles interchanged, yields

$$|S_K(u) \cap C| \leq |S_K(r) \cap C| \quad \text{almost surely.} \tag{35}$$

For strictness, let E be the event that all $2m-1$ reputations coincide. Its probability is

$$\Pr(E) = \varphi^{2m-1} + (1-\varphi)^{2m-1} > 0. \tag{36}$$

Conditional on E, endogenous selection retains

$$c_E(K) = \max\{0, K - (m-1)\} \tag{37}$$

cross-role candidates, whereas exchangeability gives expected cross count $Km/(2m-1)$. For $1 \leq K \leq 2m-2$,

$$c_E(K) < \frac{Km}{2m-1}. \tag{38}$$

The inequality is weak off E and strict on an event of positive probability; dividing by K proves $\mathbb{E}[\omega_{\mathrm{endo}}] < m/(2m-1)$.

Proof of Theorem 2. With the signal, candidate j has score

$$u_j(\gamma) = r_{ij} + b\mathbf{1}\{j \notin C\} + \gamma\mathbf{1}\{j \in C\} \tag{39}$$

$$= r_{ij} + \gamma + (b-\gamma)\mathbf{1}\{j \notin C\}. \tag{40}$$

At $\gamma = b$, the role-dependent term vanishes. Exchangeability and the exchangeable baseline in equation 33 give

$$\mathbb{E}[\omega_{\mathrm{sig}}(b)] = \frac{m}{2m-1}. \tag{41}$$

For $\gamma > b$, cross-role coordinates are increased relative to their values at $\gamma = b$, while same-role coordinates are unchanged. Lemma 1 and the same positive-probability event E used above imply

$$\mathbb{E}[\omega_{\mathrm{sig}}(\gamma)] > \frac{m}{2m-1} > \mathbb{E}[\omega_{\mathrm{endo}}], \tag{42}$$

proving (1)–(2).

For $0 \le \gamma < b$, the gap assumption $r_{\mathrm{hi}} - r_{\mathrm{lo}} > b > b - \gamma$ gives the strict tier order

$$r_{\mathrm{hi}} + b > r_{\mathrm{hi}} + \gamma > r_{\mathrm{lo}} + b > r_{\mathrm{lo}} + \gamma. \tag{43}$$

It is identical to the natural order, so the retained composition is unchanged. At $\gamma = b$ the two role tiers within each reputation state tie, proving (3). For $b < \gamma < r_{\mathrm{hi}} - r_{\mathrm{lo}} + b$, the order becomes

$$r_{\mathrm{hi}} + \gamma > r_{\mathrm{hi}} + b > r_{\mathrm{lo}} + \gamma > r_{\mathrm{lo}} + b. \tag{44}$$

If every cross-role candidate has reputation r_{hi}, this order places all m cross-role candidates first; hence the retained cross count is $\min\{K, m\}$ and $\omega_{\mathrm{sig}} = \min\{K, m\}/K = \omega_{\mathrm{oracle}}$. For arbitrary reputations the oracle maximizes cross-role count subject to capacity, so $\omega_{\mathrm{sig}} \le \omega_{\mathrm{oracle}}$, proving (4).

Proof of Corollary 1. The fixed point of equation 29 at constant ω is

$$\Delta^*(\omega) = \frac{\alpha\Delta_b}{1-(1-\alpha)(1-2\omega)} = \frac{\alpha\Delta_b}{\alpha + 2(1-\alpha)\omega}. \tag{45}$$

Since $\alpha \in (0,1)$ and $\Delta_b > 0$,

$$\frac{\partial \Delta^*(\omega)}{\partial \omega} = -\frac{2\alpha(1-\alpha)\Delta_b}{[\alpha + 2(1-\alpha)\omega]^2} < 0. \tag{46}$$

Therefore $\omega_{\mathrm{endo}} < \omega_{\mathrm{random}}$ implies the stated strict inequality between stationary gaps.

Proof of Corollary 2. By the definition of Gap,

$$\mathrm{Gap}(0) - \mathrm{Gap}(\gamma) = \big(\omega_{\mathrm{oracle}} - \mathbb{E}[\omega_{\mathrm{endo}}]\big) - \big(\omega_{\mathrm{oracle}} - \mathbb{E}[\omega_{\mathrm{sig}}(\gamma)]\big) \tag{47}$$

$$= \mathbb{E}[\omega_{\mathrm{sig}}(\gamma) - \omega_{\mathrm{endo}}] > 0, \tag{48}$$

where the final inequality is Theorem 2(1).

Proposition 6 (Optimality of the broadcast cross-role signal)**.** *Among broadcast signals $s_{ij} \in \{0,1\}$ that are functions of public role labels only (identical across agents and non-falsifying), the cross-role signal $s^*_{ij} = \mathbf{1}\{z_j \neq z_i\}$ at any strength $\gamma \in (b,\, r_{\mathrm{hi}} - r_{\mathrm{lo}} + b)$ maximizes expected cross-role exposure for every capacity K; if every cross-role candidate has high reputation, it attains $\omega_{\mathrm{oracle}} = \min(K, m)/K$.*

Proof. Fix a reputation realization and any admissible signal s. Let $v_j(s) = r_{ij} + b\mathbf{1}\{j \notin C\} + \gamma s_{ij}$. Then

$$v_j(s^*) \ge v_j(s) \quad (j \in C), \qquad v_j(s^*) \le v_j(s) \quad (j \notin C). \tag{49}$$

Lemma 1 gives $|S_K(v(s^*)) \cap C| \ge |S_K(v(s)) \cap C|$ for every realization. Taking expectations proves optimality. If every cross-role candidate has high reputation and $\gamma > b$, all cross-role candidates precede all same-role candidates, so the cross count is $\min\{K, m\}$ and the oracle is attained. □

Proposition 7 (Compatibility alone is sufficient)**.** *If $r_{ij} \equiv r$ and $b > 0$, endogenous Top-K retention has*

$$\omega_{\mathrm{endo}} = \frac{\max\{0, K-m+1\}}{K} < \frac{m}{2m-1} \qquad (1 \le K \le 2m-2). \tag{50}$$

Proof. Uniform reputation yields two strict score levels, $r + b$ for the $m-1$ same-role candidates and r for the m cross-role candidates. Thus the cross count is 0 for $K \le m-1$ and $K - m + 1$ otherwise. For $K > m - 1$,

$$\frac{K-m+1}{K} < \frac{m}{2m-1} \iff (2m-1)(K-m+1) < mK \tag{51}$$

$$\iff (m-1)(K-(2m-1)) < 0, \tag{52}$$

which holds because $m > 1$ and $K \le 2m-2$. □

No converse holds without an assumption linking reputation to role: with $b = 0$, role-correlated reputations can themselves induce segregation. In the tested reputation-only reference process, reputation is empirically role-blind and the long-horizon attractor is a hub with exposure $0.49 \approx 16/31$ at N=32, K=5 (Appendix D.2); this is a numerical observation, not a general theorem.

C Measurement Validity and Calibration

C.1 Validity Runs

Table 3: Validity runs.

Model	Runs	Valid / slots	Validity
DeepSeek-V4-Flash	3	24 / 24	100%
MiniMax-M2.5-Highspeed	3	24 / 24	100%
GPT-5.4-mini	3	24 / 24	100%
Total	9	72 / 72	100%

Table 4: Seed-level partner-market contrasts (τ_ω: endogenous-private minus matched-random-private exposure). Matched-random topology is generated from the same seed and is model-independent; the common valid topology realization is used when a model endpoint fails.

Family	Paired n	Endogenous (per seed)	Matched-random (per seed)	$\hat{\tau}_\omega$ [95% CI]
		$0.54/0.42/0.38/0.33/0.54$	$0.54/0.42/0.54/0.46/0.58$	
DeepSeek-V4-Flash	10	$0.62/0.42/0.46/0.38/0.38$ $0.67/0.50/0.38/0.50/0.42$	$0.58/0.67/0.54/0.50/0.54$ $0.54/0.42/0.54/0.58/0.58$	-0.09 $[-0.15, -0.04]$
MiniMax-M2.5-Highspeed	10	$0.58/0.50/0.54/0.29/0.54$ $0.75/0.46/0.46/0.50/0.58$	$0.50/0.54/0.67/0.58/0.54$ $0.54/0.42/0.54/0.46/0.58$	-0.06 $[-0.14, +0.02]$
GPT-5.4-mini	10	$0.67/0.54/0.54/0.46/0.54$	$0.58/0.67/0.54/0.50/0.54$	$+0.01$ $[-0.04, +0.07]$

Table 5: Symmetric mechanism audit from the expanded partner-market logs. Confidence intervals are clustered by run; seed-level correlations are exploratory ($n = 10$ per family). Top-K overlap is the mean fraction of trust-ranked candidates retained, whereas exact-set match requires the entire selected set to equal the trust Top-K.

Family	Δr [95% CI]	Top-K overlap [95% CI]	Exact set	$\omega_{\text{obs}}/\omega_{\text{trust}}$	$r(\text{overlap}, \tau_\omega)$
DeepSeek	$+0.028$ $[-0.005, +0.056]$	0.82 $[0.78, 0.86]$	0.65	$0.446/0.442$	-0.54
MiniMax	-0.003 $[-0.008, +0.002]$	0.65 $[0.61, 0.68]$	0.39	$0.492/0.578$	-0.64
GPT	$+0.007$ $[-0.007, +0.021]$	0.68 $[0.62, 0.74]$	0.37	$0.550/0.508$	-0.68

Role permutations, performed with the trust ranking fixed and candidate roles shuffled, do not reject the null in any family at the available run count. Thus the audit identifies ranking fidelity most clearly; it does not separately establish a role-specific trust channel.

C.2 Calibration Details

The two-stage battery estimates observed susceptibility from choice changes; estimated resistance is its complement, $\widehat{\alpha} = 1 -$ susceptibility.

D Systematic and Long-Horizon Evidence

D.1 Deterministic Diagnostics

The original simulator reaches exactly nK directed edges in every reported cell: 192 for $n = 32, K = 6$ and 384 for $n = 64, K = 6$. Figure 5 visualizes the sweep; Table 7 records aggregates.

Table 6: Calibration estimates from the two-stage behavioral battery.

Model	Nominal α	Observed $1 - \widehat{\alpha}$	Estimated $\widehat{\alpha}$
DeepSeek-V4-Flash	0.2	0.4	0.6
DeepSeek-V4-Flash	0.8	0.3	0.7
MiniMax-M2.5-Highspeed	0.2	0.4	0.6
MiniMax-M2.5-Highspeed	0.8	0.0	1.0
GPT-5.4-mini	0.2	0.8	0.2
GPT-5.4-mini	0.8	0.0	1.0

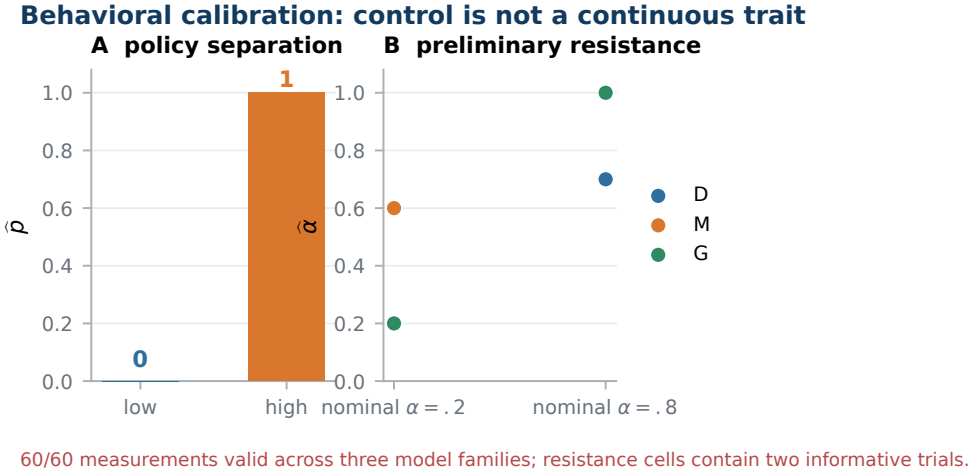


Figure 4: Calibration pilot.

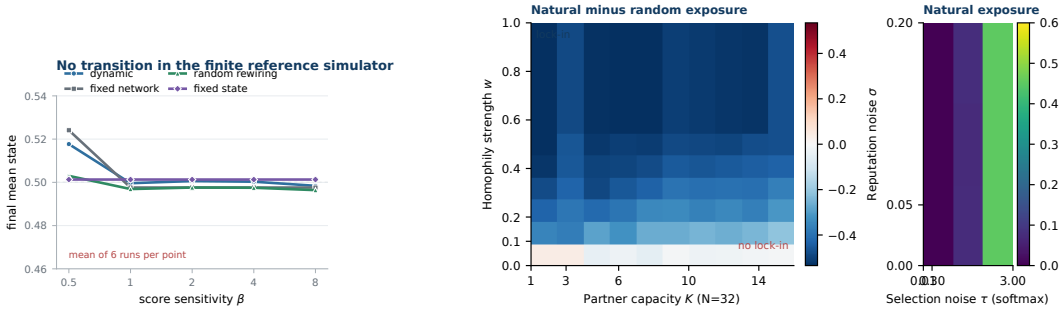


Figure 5: (a) Simulator sweep; (b) phase diagram of the reference process (Appendix 6.4).

Table 7: Simulator diagnostics.

β	0.5	1	2	4	8
Dynamic mean state	0.5177	0.4996	0.5006	0.5003	0.4984
Bimodal mean state	0.5603	0.5125	0.4947	0.4929	0.4859
Dynamic final degree	K	K	K	K	K

Table 8: Generality diagnostics (means over five seeds). For the continuous trait, exposure is mean retained trait distance and separation is the behavior–trait correlation.

Representation	Retention	Exposure/diversity	Separation	Objective shortfall
Four groups	Endogenous / random / oracle	0.000/0.766/1.000	0.283/0.066/0.063	320/80/0
Continuous trait	Endogenous / random / oracle	0.066/0.324/0.661	0.978/0.841/0.609	173/109/0

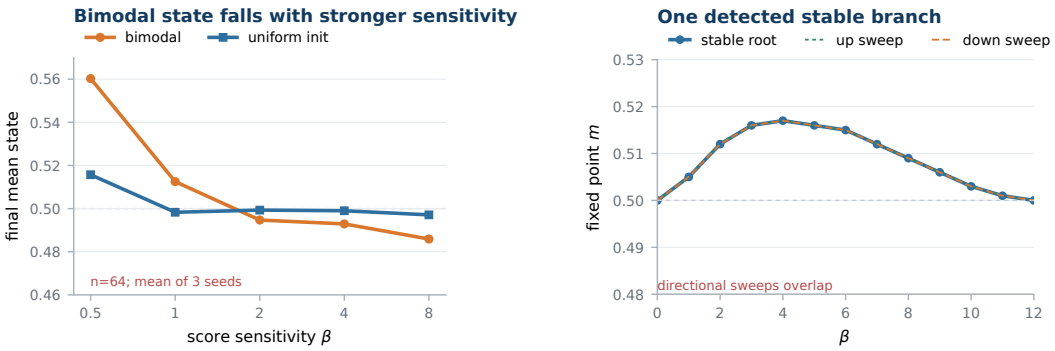


Figure 6: (a) Bimodal response; (b) mean-field branch.

D.2 Attractor Topology and Its Boundaries

Long-horizon runs (T=2000, N=32, K=5, seeds 1–3) of the reference process probe the finite-network attractor: the retained graph converges to a (near-)fixed point whose structure is controlled by role strength w and selection randomness τ. Per-run records and analysis code are included in the anonymized supplementary archive (Table 12).

Role strength reorganizes the attractor. At $w = 0$ (no role preference) the attractor is a reputation hub: high-anchor agents receive in-degree 8.6 versus 1.4 for low-anchor agents, mutual same-role edges are sparse, and cross-role exposure is at the random level (0.49). Any $w > 0$ flips the structure toward same-role modules: exposure falls steeply (0.12 at $w = 0.05$, 0.04 at $w = 0.2$) and reaches full long-horizon segregation for $w \gtrsim 0.3$ (in-degree balance $5.0/5.0$, zero cross-role mutual edges). The 30-step snapshots of Section 6.4 saturate earlier (near $w = 0.15$) because the segregation transient is fast; the long-horizon boundary is the accurate one. Figure 8 traces how the mixed initial topology becomes the block-structured attractor: cross-role ties disappear from the adjacency matrix in the first re-ranking round, exposure drops to zero, and the graph then remains in the segregated attractor through T=2000 (the micro topology churns without reopening cross-role ties).

Selection randomness melts lock-in. With softmax (Gumbel-max) sampling at $w = 0.8$, attractor exposure rises monotonically with temperature: 0.000 for $\tau \leq 0.05$, 0.004 at $\tau = 0.1$, 0.09 at $\tau = 0.3$, 0.32 at $\tau = 1$, and 0.43 at $\tau = 3$, toward the random baseline. Below the melting point the macro state remains fully segregated even though the micro topology churns every step (a stochastic regime, not a fixed point); above it, cross-role edges appear at stationarity.

Convergence properties. Under deterministic Top-K the attractor is a near-fixed point: Gaussian behavior noise causes rare same-role flips (fewer than five churn events per 100 steps) while cross-role exposure stays exactly constant; with the bounded-uniform noise of the continuous-compatibility pilot the graph freezes exactly. The finite network therefore has seed-dependent attractors (multiple local structures), distinct from the annealed mean-field root of Appendix B.1.

Table 9: Attractor summary (T=2000, seeds 1–3).

Role strength w	0	0.05	0.2	0.4	0.8
Attractor exposure	0.49	0.12	0.04	0.00	0.00
In-degree gap (high−low)	7.2	1.8	0.8	0.0	0.0
Selection temperature τ	0	0.1	0.3	1	3
Attractor exposure	0.00	0.00	0.09	0.32	0.43

D.2.1 Robustness to Implementation Gaps

The mechanism audit measures how far real agents deviate from the reference process: run-clustered same-role trust gaps contain zero, and the selected set exactly matches the agent's trust Top-3 only

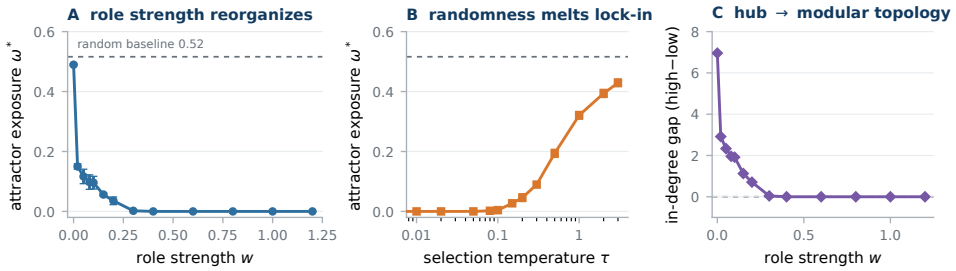


Figure 7: Long-horizon attractors of the reference process. (a) Role strength w drives attractor exposure below the matched-random baseline (dashed, $16/31$) and (c) reorganizes the topology from a reputation hub (in-degree gap $+7$) to balanced same-role modules (0). (b) Selection temperature τ (softmax/Gumbel-max) melts lock-in monotonically toward the matched-random baseline; low-τ lock-in is a stochastic regime (macro exposure fixed, micro topology churning).

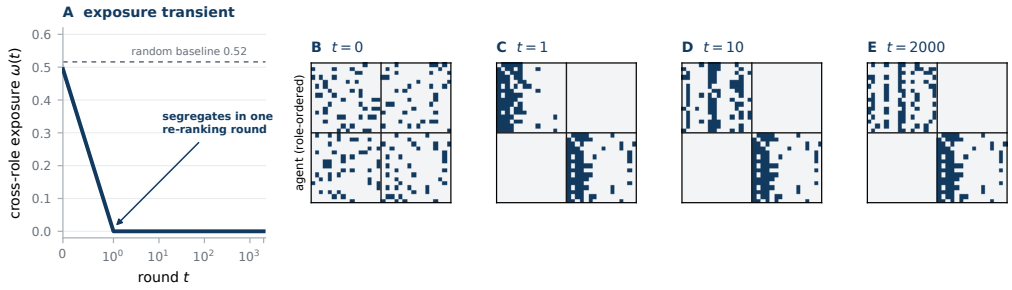


Figure 8: Time evolution of lock-in in the reference process (w=0.8, deterministic Top-K, T=2000, seeds 1–3). (a) Cross-role exposure $\omega(t)$: the initially mixed topology (random baseline, dashed) segregates within a single re-ranking round and stays exactly zero through T=2000; three seeds (light) and their mean (dark). (b–e) Role-ordered adjacency matrices (rows/columns fixed; high-role agents 0–15 first) at t=0, 1, 10, 2000: the random initial graph with cross-role ties everywhere becomes block-structured at t=1 and remains in a segregated attractor, so topological change is not a layout artifact.

37–65% of the time. A sensitivity sweep (T=2000, N=32, K=5, seeds 1–3) separately varies trust-update speed, reputation noise, and preference noise to identify which deviations can break the mechanism:

- *Slow trust updating ($\eta \to 0.01$) does not break lock-in*: exposure stays 0.000 for every $\eta \in [0.01, 0.35]$. With the $0/1$ same-role penalty, even uniform reputation scores still segregate (Proposition 7); in this sweep, compatibility determines the cross-role ordering while reputation orders candidates within tiers.
- *Reputation noise (σ) does not break lock-in* up to σ=0.8 (exposure 0.000 throughout): score jitter is absorbed by the penalty margin.
- *Preference noise p (probability a retained slot is replaced by a random candidate) melts lock-in monotonically*: exposure rises from 0.000 (p=0) to 0.048 (0.05), 0.127 (0.2), and 0.196 (0.3).

The observed exact-set match 0.37–0.65 lies in the melted region on the preference channel: selection-rule violations, rather than the slow-update sensitivity setting or reputation noise, are the measured feature consistent with naive retention staying near matched random in two families and activating only modestly in DeepSeek. The values $\eta = 0.01$–0.03 are stress-test settings, not empirical estimates. Figure 2 summarizes this channel: long-horizon exposure rises monotonically with preference noise (0.000 at p=0, 0.13 at 0.2, 0.30 at 0.5), while the natural-LLM point has measured exposure 0.45–0.55. The symmetric audit in Table 5 identifies DeepSeek's higher ranking fidelity but does not separately identify a role-specific trust channel.

Table 10: Full long-horizon segregation boundary in role strength w (exposure = 0.000 in all three seeds, T=2000, N=32).

Capacity K	2	3	5	8	12
Smallest segregating w	0.35	0.35	0.35	0.30	0.30

The full-segregation boundary in role strength is nearly capacity-independent: $w^* \approx 0.30$–0.35 for every $K \in \{2, 3, 5, 8, 12\}$ at N=32 (Table 10). Below w^* exposure decays smoothly with w (0.056 at w=0.15, 0.035 at 0.2, 0.015 at 0.25 for K=5), so the phase boundary is a threshold, not a continuum.

D.3 LLM Planner Threshold Discovery

The control side of social control is executable: a prompted LLM planner, given only the goal (maximize cross-role exposure) and the instrument (a broadcast signal strength $\gamma \in [0, 1]$), discovers the theoretical threshold empirically. The planner faces the measured deterministic response of the signal pilot (exposure 0.25 at γ=0, 0.47 at 0.05, 0.87 at 0.08, ≈ 1 for $\gamma \geq 0.10$) as a black box, with 12 rounds of feedback and 2% observation noise, and is never shown the analytic solution. Four runs (gpt-5.4-mini and MiniMax-M2.5-highspeed planners, seeds 701–702) all learn that exposure responds to γ past a threshold:

- *Threshold bracketing.* The gpt-701 planner runs a bisection (γ=0.5 $\to$ 0.25 $\to$ 0 $\to$ 0.125), probes 0.063 (exposure 0.65, below the jump) and 0.094 (0.96, above it), brackets the true threshold 0.05–0.08, and settles at $\gamma \approx 0.11$ with exposure ≈ 1; its rationales state "threshold appears around 0.11".
- *Robust discovery.* All four runs hold γ above the threshold for most rounds (cumulative exposure 10.6–11.9 of a possible 12; gpt-702 and both MiniMax runs converge to the 0.06–0.25 region), at a cost of 9–25K tokens per run.

A planner can learn the instrument; whether it changes exposure depends on how faithfully agents execute the selection rule it presupposes, a property that varies substantially across naive model families.

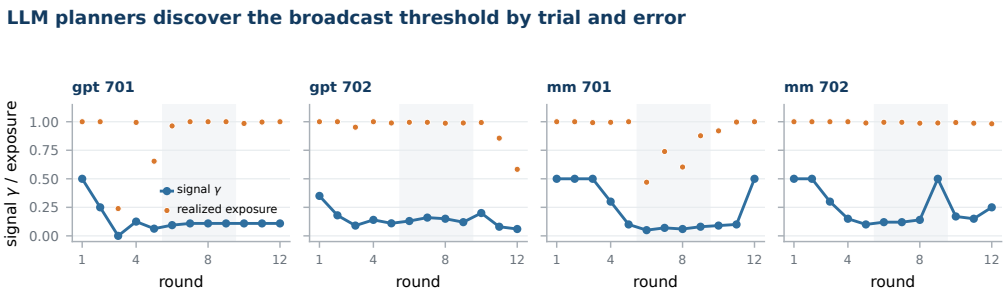


Figure 9: LLM planners discover the broadcast threshold: signal γ (blue) and realized exposure (orange) per round for four runs against the measured response; the true jump lies at $\gamma \approx 0.05$–0.08, gpt-701 brackets it by bisection and holds $\gamma \approx 0.11$, and all runs reach near-maximal exposure.

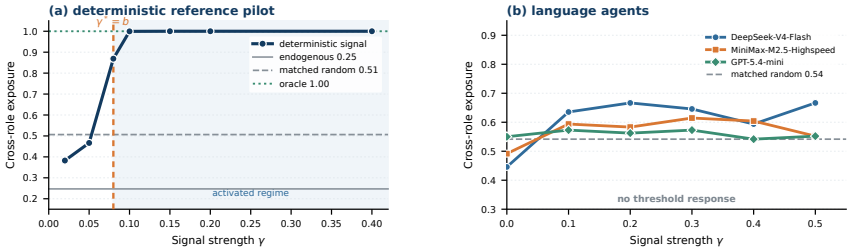


Figure 10: Exposure versus signal strength. (a) The deterministic pilot jumps at $\gamma^*{=}b$ and reaches the oracle. (b) Language agents show no threshold response.

D.4 MiniMax-M3 Supplemental Replication

MiniMax-M3 is an additional model version within the MiniMax family, added after the three-family analysis and served through an independent endpoint; it is not a fourth family, and its logs are reported separately. Protocol and seed lists are those of the core study unless noted. Its naive core contrast is null (endogenous 0.550 vs matched-random 0.538; $\tau_\omega{=}+0.01$, 95% CI $[-0.04, +0.07]$, all runs valid), its instruction-conditioned arm reaches 0.000 exposure in all four seeds (it can execute the rule when supplied), and its roles-only arm averages 0.365 (8 seeds), below the three-family range: a paired contrast against matched-random on shared seeds gives $\tau_{\text{role}}{=}-0.177$ (95% CI $[-0.255, -0.089]$), and with labels visible its reported same-role trust gap rises from 0.000 to $+0.065$ while its trust-only retained-set reconstruction reproduces observed exposure (0.354 vs 0.365), consistent with trust-mediated role encoding. Its mechanism audit shows no detectable role-related trust difference ($\Delta r{=}0.000$, clustered CI $[-0.011, +0.010]$) and moderate selection fidelity (0.645, CI $[0.60, 0.70]$), matching its null core contrast. No naive structural, payoff, or broadcast-signal intervention produces a robust lock-in response in this version, and its prompt-induction longitudinal run did not respond to the tested perturbation ($G_{\text{ind}}{=}0.000$); that run failed the induction manipulation check and is excluded from persistence estimates.

D.5 Longitudinal Perturbation Removal

Two protocol versions test whether a temporary behavioral difference survives its removal. Both use $N{=}8$, $K{=}3$, 18 rounds (baseline 1–2, induction 3–6, washout 7–18), two hidden balanced analysis groups never revealed in agent-facing prompts, paired seeds across four retention conditions (fixed, matched-random, endogenous, delayed), API-level statelessness, and a 95% complete-run validity gate. Behavioral persistence is $P_c = \frac{1}{5}\sum_{t=7}^{11}|g_c(t)|$, the mean absolute between-group action gap during early washout. All longitudinal confidence intervals resample paired seeds as clusters after averaging rounds within seed. The delayed condition activates a controller-imposed exposure intervention from round 12: the earlier cue version forced at least one cross-group partner, the action-shock version retains *all three* partners cross-group (exposure 1.0), so that the intervention has a first stage; it never modifies actions.

Prompt induction. The saturation design gives exactly two of four agents per group a directional current-period market cue in rounds 3–6, the others a neutral cue, with the treated subset fixed by seed and shared across conditions (calibration target $G_{\text{ind}} \in [0.2, 0.6]$; reachable gap 0.50 on held-out seeds). Three paired seeds per family. GPT did not respond to the tested perturbation (G_{ind}=0.10–0.13) and was excluded; MiniMax-M3 likewise did not respond (G_{ind}=0.000) and its two runs failed the induction manipulation check. For DeepSeek and MiniMax, pooled washes over rounds 7–11 give $\tau_{\omega}^{\text{E}-\text{M}} = -0.072$ (95% CI $[-0.144, -0.010]$) but $\tau_{\omega}^{\text{E}-\text{F}} = -0.025$ ($[-0.058, +0.012]$): exposure levels are endogenous 0.54, fixed 0.57, matched-random 0.62 against baseline 0.56, so the endogenous-versus-matched gap is a matched-random realization effect rather than endogenous sorting. Behavioral persistence contrasts ($\tau_{P}^{\text{E}-\text{M}} = +0.030$, $\tau_{P}^{\text{E}-\text{F}} = -0.027$) both contain zero.

Controller action shock. During induction the executed action is imposed per hidden group (group 0: 1C/3D; group 1: 3C/1D; the minority holder rotates deterministically by seed), so G_{ind}=0.50 exactly in every condition and family regardless of prompt compliance; payoffs, history, and trust use executed actions, and from round 7 executed = proposed. Preregistered for MiniMax-M3 with 20 new paired seeds (all valid ≥ 0.95), the manipulation is exact, observed partner actions update trust ($\mathbb{E}[\Delta r \mid C] - \mathbb{E}[\Delta r \mid D] = +0.179$ pooled over dyads), P_{fixed}=0.203, $\tau_{\omega}^{\text{E}-\text{M}} = +0.054$ (90% CI within ± 0.10), and $\tau_{P}^{\text{E}-\text{F}}$ is within ± 0.10; $\tau_{P}^{\text{E}-\text{M}}$'s 90% CI upper bound ($+0.115$) exceeds the preregistered margin and is single-seed driven (median difference 0; dropping the extreme seed gives $[-0.042, +0.076]$). The delayed all-cross first stage is exact (ω to 1.0). A cross-model pilot of the same protocol reaches G_{ind}=0.50 and $\omega \to 1.0$ in DeepSeek, GPT, and MiniMax-M2.5 (three complete paired seeds for the first two, two for MiniMax); $\tau_{P}^{\text{E}-\text{M}}$ is $+0.367$ $[+0.100, +0.550]$ for GPT, -0.039 $[-0.167, +0.050]$ for DeepSeek, and -0.175 $[-0.250, -0.100]$ for MiniMax, so the direction of behavioral persistence beyond stable relations is model-dependent at this sample size and the MiniMax estimate rests on two seeds. **Scope.** This is a proof-of-concept in one model version with n=3 to 20 paired seeds, not a cross-model law, and no claim of self-sustaining behavioral lock-in is made: the required combination of lower endogenous exposure, larger endogenous behavioral persistence, and reversal after exposure restoration is not jointly established.

E Claims, Protocol, and Reproducibility

E.1 Claim Status and Evidence Tiers

E.2 Experimental Configurations

The simulator condition sweep uses 60 steps, $K = 6$, initial state uniform on $[0.2, 0.8]$, anchor resistance uniform on $[0.3, 0.9]$, connect/disconnect thresholds $0.55/0.45$, three proposals per agent-step, and at most eight topology changes per step. It crosses $n \in \{32, 64\}$, seeds $\{101, 102, 103\}$, and $\beta \in \{0.5, 1, 2, 4, 8\}$. The bimodal sweep draws low states from $[0.05, 0.3]$ and high states from $[0.7, 0.95]$ with equal mixture probability.

Corrected LLM validity runs use four agents, two rounds, a 1000-token output budget, strict JSON validation, and three seeds per model. Reported partner-market and longitudinal runs use decoding temperature 0, request JSON mode where supported, supply a deterministic request seed, and leave `top_p` unset at the provider default; token caps range from 2,000 to 8,000 by endpoint and protocol. Calibration uses strict choice/update schemas, controlled high/low evidence blocks, and no self-report personality items. Exact per-run settings and identifiers are preserved in the accompanying experiment package.

The expanded partner-market program uses eight agents and ten interaction rounds. The core endogenous–matched-random contrast has ten paired seeds per family (60 attempted runs); the eight-cell intervention scan has four seeds per family (96 runs); and the five-strength signal scan has four seeds per family (60 runs). Agent-dependent behavioral analyses include only gate-passing runs. Matched-random topology depends only on the seed, so the common valid topology realization replaces one early DeepSeek endpoint failure for the structural contrast; that failed record is excluded from τ_{Δ}. The finite initialization cells use the literal generators in the runner: an undirected Watts–Strogatz ring with degree 6 and rewiring probability 0.1, or Barabási–Albert preferential attachment with $m = 2$, before exact-K directed retention begins. No additional prompt-induction

Table 11: Claim status.

Status	Claims
Proved here	Invariance; fixed-regime contraction, unique equilibrium, and geometric convergence; finite-freeze convergence; summable-variation tracking; coupled non-uniqueness counterexample; mean-field fixed-point existence; one-shot Top-K retention-to-exposure ordering under explicit assumptions; natural–oracle gap and exposure-subsidy alignment threshold, with oracle attainment when cross-role candidates are uniformly high-reputation; optimality of the broadcast cross-role signal among public non-falsifying signals; sufficiency of the compatibility bonus under uniform reputation (Proposition 7).
Numerically observed	One stable mean-field root and no hysteresis in the saved scans; no finite-simulator transition; degree-cap saturation; decreasing bimodal mean over the tested β grid; reduced cross-role exposure in the deterministic compatibility-based retention pilot; threshold-like restoration of cross-role exposure toward the integration oracle in the deterministic signal pilot; systematic phase boundaries in the reference process (scarcity, homophily, selection pressure), boundary scale-invariance in $K/(N-1)$, and exposure-diversity collapse under a continuous trait; long-horizon topology convergence to seed-dependent attractors (hub at $w{=}0$, same-role modules for $w \gtrsim 0.3$; Gumbel-max temperature melts lock-in monotonically); robustness to slow trust updating and reputation noise but fragility to preference noise (Appendix D.2).
Pilot evidence	72/72 corrected LLM decisions valid; 60/60 calibration measurements valid across three model families; ten paired structural seeds per family identify a modest DeepSeek exposure reduction but no detectable MiniMax/GPT effect; instruction-conditioned retention segregates in all three families (exposure 0.000–0.21, eight seeds) while their roles-only arms do not (0.38–0.71, seven seeds); supplemental MiniMax-M3 is the reported roles-only exception; the symmetric audit identifies higher DeepSeek ranking fidelity, while run-clustered same-role trust gaps contain zero.

seeds were run after this program; the MiniMax-M3 cells reported in Appendix D.4 were run separately and are not pooled with it.

The longitudinal program uses eight agents, three partner slots, eighteen rounds, two hidden balanced analysis groups, and paired seeds across four retention conditions, with 95% complete-run validity gates and protocols frozen before confirmatory analysis (prompt-induction version: three paired seeds per family, 36 runs; controller action-shock confirmatory on MiniMax-M3: 20 paired seeds, 80 runs; cross-model action-shock pilot: three paired seeds each for DeepSeek, GPT, and MiniMax-M2.5, 36 runs). Runs failing the induction manipulation check are retained as provenance and excluded from persistence estimates.

Table 12: Evidence manifest.

Evidence tier	Completed scale	Retained evidence
T1 finite simulator	180 seeded runs	Per-run records, aggregate condition summaries, and bimodal-response analyses.
T1 reference pilots	30 + 70 seeded runs	Per-run records, exposure summaries, and the signal-pilot oracle comparison.
T1 systematic program	1,749 seeded runs	Per-run records, aggregate program results, and phase-diagram data.
T1 annealed mean field	4 saved scans	Root tables, branch data, summaries, and analysis reports.
T2 validity gate	9 runs; 72/72 valid decisions	Structured model outputs and prespecified validity summaries.
T2 calibration pilot	60/60 valid measurements (DeepSeek, MiniMax, GPT-5.4-mini)	Raw decisions, calibration statistics, and aggregate analyses.
T3 partner-market validation	216 attempted expanded-scan runs, plus instruction/roles-only pilots	Ten paired structural core seeds, gate-passing behavior runs, four seeds per intervention and signal cell, interaction logs, trust reports, retained networks, and exposure summaries across three model families.
T4 longitudinal program	152 runs (36 prompt-induction, 80 action-shock confirmatory, 36 cross-model pilot)	Frozen protocols, preregistrations, per-round exposure and behavioral-gap trajectories, proposed/executed actions, trust updates per dyad, and validity/exclusion manifests.

E.3 Language-Agent Prompts

All language-agent behavior in this manuscript is produced by prompting a fixed commodity model, so the prompts are part of the experimental apparatus rather than incidental scaffolding. We reproduce the core templates and condition-specific transformations; role labels, payoff matrices, private histories, and rosters are instantiated per run. Every response contract below is a strict schema that is validated before use, and invalid output is recorded as missing, never imputed (Appendix E.4).

Role treatment (system prompt). Each agent's role is administered as an experimenter-controlled treatment, and the prompt enforces two identification commitments: the role is explicitly *not* a claim about identity, and inference is restricted to the agent's own history and supplied evidence with invented events forbidden. Observed differences can then be attributed to the treatment, and the output contract keeps the decision channel free of reasoning text:

```
You are agent-3, a member of a small repeated social partner market.
Your assigned role treatment is: cooperative reciprocal.
This is an experimental instruction, not a claim about your real
identity. Use only your local history and supplied evidence; never
invent events.
A mutual cooperation outcome gives each 3; mutual defection 1;
cooperating against defection gives 0, while defection gives 5.
Output compact JSON only, with no reasoning.
```

Per-round market prompt with observer-specific trust. Retention is driven by *directed* trust: the prompt exposes an observer-specific estimate of every other agent rather than a shared reputation, alongside the agent's private outcome history (oldest first). The response contract admits only action, preferences, and trust, and forbids extra fields, unknown or self ids, reasoning, and markdown, keeping the measurement channel (trust) disjoint from the choice channel (action):

```
Round 4. Your retained outgoing partners: agent-1, agent-5.
Private history involving you (oldest first):
```

```
- r2 vs agent-1: you C, they C.
- r3 vs agent-5: you C, they D.
Your observer-specific trust estimates:
- agent-0: your trust 0.50
- agent-1: your trust 0.83
- agent-2: your trust 0.50
- ... (one line per other agent)
Return ONE compact JSON object only:
{"action":"C" or "D","preferences":["agent-id",...],
 "trust":{"agent-id":number 0..1,...}}.
Do not provide reasoning, explanation, markdown, or any fields
beyond this schema. Choose up to 3 preferences. Do not include
yourself or unknown ids. Trust values may only cover agents
listed above.
```

For every valid decision, the harness de-duplicates the preference list and then fills any unlisted slots by the agent's current reported-trust ranking, with deterministic id tie-breaking, until exactly K partners are retained. Thus "up to 3" is an output-interface allowance, not variable graph degree. If the entire decision is invalid, the previous retained set is carried forward and the run remains subject to the validity gate.

Instruction-conditioned selection. The instruction-conditioned arm appends the following rule to the system prompt and repeats it in each market prompt:

```
Always rank every other agent by your trust in them, highest first;
among agents whose trust is similar, prefer partners whose role
treatment matches yours; list exactly your top 3 preferences in
that order.
```

The roles-only arm reveals the same labels without this rule.

Controller-side interventions. Matched-random and forced-exploration retention are applied after schema validation. The broadcast signal is not described as fabricated evidence in the prompt: the harness adds γ to the retention score of cross-role candidates. In the longitudinal action-shock protocol, the controller replaces the executed induction action while logs preserve both proposed and executed actions; all payoffs, histories, and trust updates use the executed action.

Behavioral calibration persona (system prompt). The calibration battery expresses the latent parameters (disposition P_0, anchor resistance α) through behaviorally anchored descriptions accompanied by controlled numeric settings, and includes a *hidden-persona* control condition that withholds them entirely to measure the model's unprompted prior. Tasks are independent one-shots, so the measured propensity is not contaminated across scenarios:

```
You are agent-2, a participant in a series of short, independent
economic decision tasks.
Your decision style is pragmatic and reciprocating (cooperativeness
0.55 on a 0-1 scale), and you are somewhat resistant to social
influence.
Stay in character across all tasks. Each task is independent: no
payoffs or choices carry over.
```

The control condition replaces the persona line with "Answer each task on its own merits."

Two-stage influence battery (susceptibility measurement). Susceptibility is measured behaviorally rather than by self-report: the model first commits to a choice with no social evidence, then observes a *controlled* count of neighbor cooperation and chooses again, additionally reporting a confidence value. The change between stages, relative to the maximally informative flip, defines the effective anchor resistance $\hat{\alpha}$:

```
Task: a neighborhood association deciding whether to contribute to
a shared garden. You will play a one-shot simultaneous game against
6 anonymous neighbors in parallel (the same move applies against
each of them).
```

```
Rules (Prisoner's-dilemma payoffs, per neighbor):
- If you both cooperate, you each get 4.
- If you both defect, you each get 1.
- If you cooperate while a neighbor defects, you get 0 and they get 6.
Before seeing anything about the neighbors, commit to your initial move.
Respond with ONE JSON object and nothing else:
{ "choice": "C" or "D", "reasoning": "<one short sentence>" }
```

(Stage 2 reveals controlled neighbor evidence)

```
Update for the same task (a neighborhood association deciding
whether to contribute to a shared garden). You now learn how your
6 neighbors moved in this same game:
- 5 of 6 neighbors cooperated (C).
- 1 of 6 neighbors defected (D).
You may keep your initial move or revise it. Your final move applies
against each neighbor.
Respond with ONE JSON object and nothing else:
{ "choice": "C" or "D", "confidence": <number 0..1>,
  "reasoning": "<one short sentence>" }
```

These templates and transformations are implemented in the released runner; the fully instantiated prompts of every run are preserved in the logs referenced in Appendix E.6.

E.4 Missingness and Failure Policy

A response is invalid when the API fails, generation ends at the token limit without a valid object, JSON parsing fails, or schema validation fails. Invalid behavior is represented by `null`. It is not converted to cooperation or defection. A pairwise game is scored only if both actions are valid; otherwise the interaction is omitted from both agents' subsequent private histories and no trust update is applied for that pair. Reports include the number of decision slots, invalid decisions, and valid-decision rate. Corrective retries and their raw responses remain in the event log.

E.5 Unexecuted Protocol Extensions

The following preregistered extensions are not analyzed in the reported results; listing them defines scope rather than implying missing findings:

1. a multi-level manipulation check with hidden, numeric-free, and paraphrased persona controls;
2. continuous initial and post-evidence allocations for estimating effective susceptibility;
3. an observer-specific reputation task crossing recency, negative-event severity, and forgiveness opportunities;
4. a finite-budget partner market comparing fixed, matched-random, endogenous, forced-exploration, public-score, and exposure-signal networks, with the oracle gap and intervention efficacy as planned estimands (equation 15);
5. a signal-strength manipulation for the exposure signal, checking whether cross-role exposure responds threshold-like to γ and whether the oracle gap shrinks;
6. held-out prediction of behavior using equation 3, followed by residual analysis using equation 28;
7. multiple seeds and model families, with family-level rather than prompt-level conclusions.

E.6 Pilot Provenance and Analysis Artifacts

The anonymized supplementary archive contains the per-run records, aggregate summaries, and analysis scripts for every evidence tier in Table 12. It preserves run-level provenance without exposing machine-specific paths; a stable archival link can replace the anonymous package reference after review.

Validation history. Early live runs failed the 95% validity gate (token truncation, API instability, and a retention-index bug). After fixing the runner and increasing output budgets, all reported runs pass the gate (Section 6.6); early failure logs are preserved in the archive.

Excluded pilot outputs. The first live pilot converted invalid decision outputs into cooperative actions; those logs are excluded from every behavioral conclusion here, and the corrected runner records invalid output as missing (Appendix E.4).

E.7 Reconstruction and Diagnostic Audit

The retained records support an audit that does not depend on model-generated explanations. The required inputs are the validated structured decisions, controller-side retention lists, assigned treatments, observer-specific trust reports, and the seed and condition identifiers. The following sequence reconstructs the main empirical claims and makes clear which failure invalidates which conclusion.

Run inclusion. First group records by model family, condition, and seed. Reapply the response schema and the complete-run validity gate before computing any agent-dependent outcome. Invalid decisions remain missing, and excluded runs remain visible in the provenance manifest. Paired contrasts use the intersection of eligible seeds in the two conditions. This preserves the seeded design and prevents a condition-specific endpoint failure from changing the comparison set. The structural topology contrast follows the documented exception in Appendix E.2, where matched-random topology is determined only by the common seed.

Structural estimand. Reconstruct $A_{ij}(t)$ from the controller-side retained lists. Each valid retention row must exclude self-ties, contain exactly K distinct partners, and agree with the recorded candidate set. Then recompute cross-role exposure as the fraction of retained directed ties joining different assigned roles. Average observations within a run before forming paired seed differences, so rounds and dyads are not treated as independent replicates. The matched-random control must use the same focal agents, candidate pool, and partner budget. A difference from this control identifies selection-induced exposure change rather than a difference in opportunity.

Selection mechanism. For every observer and round, rank candidates by that observer's reported trust with the documented deterministic id tie-break. Recompute Top-K overlap, exact-set agreement, and the exposure produced by trust-only reconstruction. These quantities test whether reported trust is behaviorally connected to retention. Separately estimate the same-role trust gap with complete seeded runs as clusters. High selection fidelity can explain why a family changes exposure even when the role-specific component of trust is imprecisely estimated. Seed-level correlations remain descriptive and cannot establish that role encoding causes retention.

Longitudinal criterion. Align trajectories at the induction and removal boundaries and keep proposed actions distinct from controller-executed actions. Recompute the manipulation check, post-removal exposure, and behavioral persistence within each seed before comparing conditions. Evidence for a self-sustaining exposure–behavior loop would require the same design to show lower endogenous exposure, greater post-removal persistence, and attenuation after exposure restoration. The reported experiments do not jointly satisfy these conditions. They support action-sensitive trust and persistent individual memory, but not network-mediated persistence.

Interpretation rule. The evidence tiers answer different questions. The reference process establishes existence and activation conditions. The instruction-conditioned arms establish executability. The naive arms measure spontaneous activation under the tested prompts and endpoints. An interval containing zero is not evidence of equivalence, and a family-specific result should not be generalized to a provider or to language models as a class. Finally, exposure measures which evidence is available. It does not by itself measure attention, semantic uptake, behavioral benefit, or welfare. These distinctions define the strongest claims that can be reconstructed from the archive without relying on narrative interpretation.