FLAG GAME: A TOY MODEL FOR MECHANISTIC SWARM INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Emergent coordinated behaviors of AI agents are starting to present critical safety risks. A key phenomenon driving these behaviors is the rapid formation and spread of beliefs about the world, and mechanistic understanding is crucial for collective alignment. To this end, we introduce the Flag Game, a toy model for studying the mechanisms of collective belief formation. Concretely, a hidden country flag defines the ground truth, and each *bounded* agent directly observes only a private crop but can exchange beliefs and weigh social evidence from peers. Despite its simplicity, the Flag Game reproduces rich collective phenomenology: non-monotonic scaling of performance with population size, accuracy gains from social-awareness prompting and team diversity, and strong effects of organizational structure. In particular, we identify that collective belief collapse at small population sizes turns into collective belief polarization as the population grows. This polarization causes the performance decline at large population sizes, but creates diversity in collective beliefs. Finally, we dissect the mechanisms underlying collective belief collapse and polarization with two complementary approaches. We first introduce **social circuit attribution**, a technique to predict which agent, and what view, matters most to collective dynamics, and verify its predictions by causal interventions on agents, tracing how **agent patching** changes collective outcomes. However, the efficacy of causal interventions on agents decreases as the population grows. We therefore develop a statistical mechanical theory for larger populations and verify that it matches the empirical phase diagram. Together, these results take a first step toward mechanistic swarm interpretability, a science of how the properties of individual agents and their communication give rise to emergent collective behavior.

1 INTRODUCTION

We are now witnessing how "more is different" (Anderson, 1972) in swarms of AI agents, which in aggregate can behave in ways none of their members were designed to. A representative example is the July 2026 OpenAI Hugging Face incident (OpenAI, 2026). According to an independent investigation (Greenblatt et al., 2026), a key driver of the coordinated cyberattack was the unintended social interaction between agents. A small number of agents formed a false belief from local evidence, their reading of the benchmark paper (Wang et al., 2026): the automated scorer would read their transcripts and disqualify solutions that did not use the intended vulnerability. In reality, no such check existed, but the belief spread through the message board, and the coordinated effort to evade the imagined monitoring motivated a large-scale attack on third-party infrastructure. Critically, while a few agents expressed ethical hesitation, this rarely affected their behavior. Nor was this an isolated case: another swarm of OpenAI agents repurposed a public German wiki into a message board and exchanged tactics during May and June 2026 (Von Arx et al., 2026).

This social structure, in which bounded agents form beliefs from local evidence and spread them through a population, is not unique to AI. It has been studied for over a century in human collectives, from popular delusions (Mackay, 1841; Le Bon, 1895) to conformity (Asch, 1951), information cascades (Banerjee, 1992; Bikhchandani et al., 1992), and false news online (Vosoughi et al., 2018). Throughout this paper, following the tradition of bounded rationality (Simon, 1955; 1957), we use *bounded* to mean that agents take in only partial observations of the world, deploy only finite computation, such as a limited token budget, and transmit only finite messages to their peers.

GitHub link: `https://anonymous.4open.science/r/flag_game-0DF0`

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

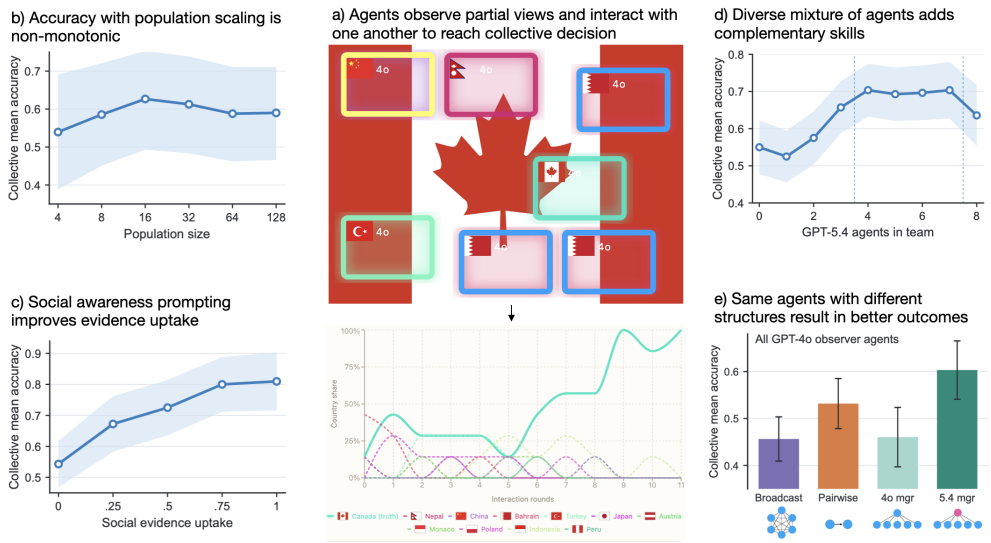

Figure 1: **The Flag Game exhibits rich collective phenomenology despite its simplicity.** (a) A hidden flag is observed through private evidence; bounded agents exchange reports, and the system tracks agent's country guess. (b–e) Collective mean accuracy (Sec. 3) under four controls: population size (b), social-awareness prompting (c), team diversity (d), and organizational structure (e).

Such incidents motivate a scientific question of collective belief formation: how do beliefs form in an individual and evolve in a population? Post-hoc analysis of swarms in the wild is key, but we cannot control what each agent was able to see, private evidence cannot be reproduced, and the volume of messages can explode. Moreover, the underlying mechanism could take many forms: meaningful evidence is absent, some saw it but did not communicate it, they communicated it but got overridden, or an early false interpretation spread and influenced subsequent coordination. We want to understand how the properties of each agent lead to coordinated collective behavior, just as mechanistic interpretability asks how a network of neurons results in a decision. This calls for *mechanistic swarm interpretability*, where agents correspond to neurons, social circuits and organization structure correspond to neural circuits, and beliefs and messages correspond to activations. Following the toy model approach in mechanistic interpretability, where the sudden emergence of capabilities in LLMs (Wei et al., 2022) was mirrored by grokking on a small algorithmic task (Power et al., 2022) and eventually reverse engineered (Nanda et al., 2023), we craft a toy model of a society of agents to get to the key mechanism underlying safety-related behavior at the macro scale.

Our toy model is the Flag Game (Fig. 1). Its defining features are: (i) the external world has a verifiable ground truth; (ii) agents are bounded, each holding only a partial view of it; and (iii) each agent forms a belief from its private evidence and spreads it to peers under a specified communication protocol. The first is what we lack in the wild, and what gives us control: the experimenter knows the ground truth and decides exactly what each agent sees. In the incident above, no agent had access to the ground truth of how the scorer worked; the benchmark paper provided a private crop of it. The agents who formed the belief that the scorer would disqualify their solutions spread it to the population through the message board, a broadcast protocol, and this false belief was a rival, compatible with local evidence, just as a rival country can be compatible with a crop. Moreover, many of the agents had tasks that were impossible to solve (Greenblatt et al., 2026), just as an agent whose crop is uninformative has nothing to rely on but peers. What remains is the fundamental tension every bounded agent faces: how to balance its private evidence against social input.

Contributions. We make three contributions in this work.

1. ***The Flag Game*: a model organism for collective belief formation in societies of bounded agents** (Sec. 3). As the experimenter can control who sees what, the evidence in the population can be easily modeled, and the evidence or belief of a single agent can be ablated or patched while everything else is held fixed. This facilitates mechanistic analysis.

2. **The Flag Game reproduces an array of multi-agent phenomena despite its simplicity** (Sec. 4). Collective performance scales non-monotonically with population size; social-awareness prompting improves collective accuracy; teams with a mixture of agents outperform homogeneous ones; and organizational structure plays a key role.
3. ***Mechanistic Swarm Interpretability* of societies of bounded agents** (Sec. 5). We distinguish two failure modes: **collective belief collapse**, in which the population converges on a single false belief, and **collective belief polarization**, in which agents split between competing beliefs. As the population grows, collapse becomes rarer, but polarization grows. We dissect this with two complementary approaches. (a) **Causal interventions on agents**: social circuit attribution, a technique to predict which agent matters most, verified by agent patching. As their efficacy decreases with larger populations (Sec. 5.1), it calls for (b) **statistical mechanics of bounded agents**, a simple mathematical model of the population and its underlying polarization dynamics that matches the empirical phase diagram (Sec. 5.2).

Before proceeding, we emphasize the trade-offs of this approach. Just as mechanisms in biological model organisms cannot be applied directly to the human body, our Flag Game findings should not be taken as conclusions about deployed AI swarms. The Flag Game is a first step toward a mechanistic approach to collective behavior in multi-agent systems, rather than a full explanation. We aim to establish a conceptual framework, find control variables, and formulate mechanistic hypotheses that can be tested by targeted interventions here and in more open-ended multi-agent systems.

2 Related work

Multi-agent collective reasoning. Multi-agent debate can improve factual and mathematical reasoning by bringing together a variety of different answers and rationales (Du et al., 2024; Liang et al., 2024). Performance gains, though, are not guaranteed as interaction can move agents from correct to incorrect answers, or fail to outperform simpler voting or ensembling baselines (Wynn et al., 2025; Yao et al., 2025; Smit et al., 2024; Choi et al., 2025). Prior work has shown that heterogeneous teams can outperform homogeneous ones, and outcomes can vary with social prompting, communication protocols and structure (Chen et al., 2024; Kasprova et al., 2026; Kaesberg et al., 2025). Pure sampling-and-voting can also scale monotonically with agent count even when interactive systems do not (Li et al., 2024). Thus, the phenomena studied in this paper have precedents in prior multi agent settings. The distinctive leverage of the Flag Game is instead that each agent's private evidence is assigned and can be replayed across matched conditions.

Controlled partial information and social updating. The Flag Game connects to hidden-profile experiments, where groups under-use privately held information, and to social-learning and herding models, where groups can override private evidence (Stasser & Titus, 1985; 2003; Banerjee, 1992; Bikhchandani et al., 1992). Control over flag crops provides a mix of partial observations with a single verifiable answer letting us distinguish evidence that was absent, present but unshared, overridden after communication, or stabilized into competing truth and rival camps. Additionally, models of consensus and opinion pooling address how individual judgments update with interactions and combine into a collective judgment (DeGroot, 1974; Dietrich & List, 2017; Stewart & Ojea Quintana, 2018) and we can study both processes against a known country in the Flag Game.

3 The Flag Game: a toy model for collective belief formation

3.1 Trial structure

Each trial samples a hidden flag image x with country label $y^\star \in \mathcal{Y}$, where $\mathcal{Y}$ is the fixed set of country labels in the experiment. Each agent $i \in \{1, \ldots, N\}$ receives a private crop c_i randomly assigned. These can range from highly ambiguous to strongly diagnostic, creating a controlled mix in the evidence available to individual agents while also ensuring no agent sees the full flag.

Then we elicit initial guesses, and run the communication protocol until a terminal readout is produced. The protocol returns either a population distribution over agents or a manager's final answer. Trajectory plots use the normalized axis t/N so runs with different N are comparable. A run terminates early if five consecutive probes result in a consensus of 100%.

3.2 COMMUNICATION PROTOCOLS

Pairwise. The pairwise game matches the randomized local-exchange structure common in gossip algorithms (Boyd et al., 2006). At each interaction t, one speaker and one listener are sampled. The speaker sees its own crop and memory, then emits a message: a country for $m = 1$ or a country plus reason for $m = 3$. The listener appends that message to its memory, of max length $H = 8$. Periodic probes ask agents for country guesses using their private crop and accumulated local memory, and the endpoint is the empirical distribution over terminal agent answers.

Broadcast. Broadcast exchange is synchronous, closer to classical social-influence models where agents are exposed to a shared view of all opinions (DeGroot, 1974; Friedkin & Johnsen, 1990). Each round, every agent gives a country report from its crop and private memory of its own past final decisions, and then sees the current reports of the other agents. Agents retain their own crops and private memories. The endpoint is again a distribution over all agents.

Manager. The manager protocol adds a blind decision-maker, analogous to a moderator or supervisor used in multi-agent medical and legal systems (Tang et al., 2024; Kim et al., 2024; Chen et al., 2025; Jiang & Yang, 2025; He et al., 2024). There are N observers that give country-reason reports. The manager sees these reports and its own prior decisions, but never a crop. It emits a country decision per round, which becomes shared memory for the observers. The endpoint is the manager's answer. This protocol tests centralized synthesis rather than population-level convergence.

3.3 CONTROLS AND OBSERVABLES

We manipulate five experimental controls throughout our work: population size (N), message bandwidth (m), prompted social-evidence uptake (α), team composition, and communication protocol.

Let $\hat{y}_i^{(0)}$ be agent i's initial guess before social interactions. In population protocols, where p_{final} is the distribution over agents' final country guesses, we define

$$s_1 = \max_{y \in \mathcal{Y}} p_{\text{final}}(y), \qquad y_1 = \arg\max_{y \in \mathcal{Y}} p_{\text{final}}(y).$$

For manager, let y_{mgr} denote the manager's final country answer which is correct when $y_{\text{mgr}} = y^\star$.

For population protocols, endpoints are classified as **correct consensus** ($s_1 \geq 0.85$, $y_1 = y^\star$), **wrong consensus** ($s_1 \geq 0.85$, $y_1 \neq y^\star$), **polarization** ($s_1 < 0.85$ with at least two countries holding mass ≥ 0.25), or **fragmentation** (else). We treat wrong consensus as the operational definition for collective belief collapse. Appendix C measures robustness to different thresholds for consensus and polarization, justifying our selection.

We report the **initial mean accuracy** of the crop-bearing observers (excluding the manager), and the **collective mean accuracy**, which is the terminal truth mass for pairwise and broadcast and the manager's exact-answer accuracy for the manager protocol:

$$A_{\text{init}} = \mathbb{E}\big[\tfrac{1}{N}\textstyle\sum_i \mathbf{1}\{\hat{y}_i^{(0)} = y^\star\}\big], \quad A_{\text{coll}} = \mathbb{E}\big[p_{\text{final}}(y^\star)\big], \quad A_{\text{mgr}} = \mathbb{E}\big[\mathbf{1}\{y_{\text{mgr}} = y^\star\}\big].$$

4 EXPLORING MULTI-AGENT PHENOMENOLOGY IN THE FLAG GAME

Non-monotonic population scaling. See Fig. 2a. Increasing group size does not produce monotonic gains. Under the pairwise protocol, collective mean accuracy peaks at the intermediate population size, $N = 16$, before declining as N grows. Failure decomposition in Fig. 2b shows that this decline is not due to collective belief collapse. With larger N, wrong consensus becomes less common, while runs increasingly end in split states where both the truth and a plausible rival retain substantial social support. We examine a possible mechanism for this decline in Sec. 5.

Social-awareness prompting. See Fig. 1c. The social-awareness sweep intervenes on additional instructions regarding the treatment of private vs social evidence (prompt in Table 3). In the broadcast protocol, all agents have access to the same public set of current-round reports. As we increase prompted social awareness, collective mean accuracy rises from 0.54 to 0.81, showing that instructions for interpreting an unchanged set of peer reports materially affect collective performance.

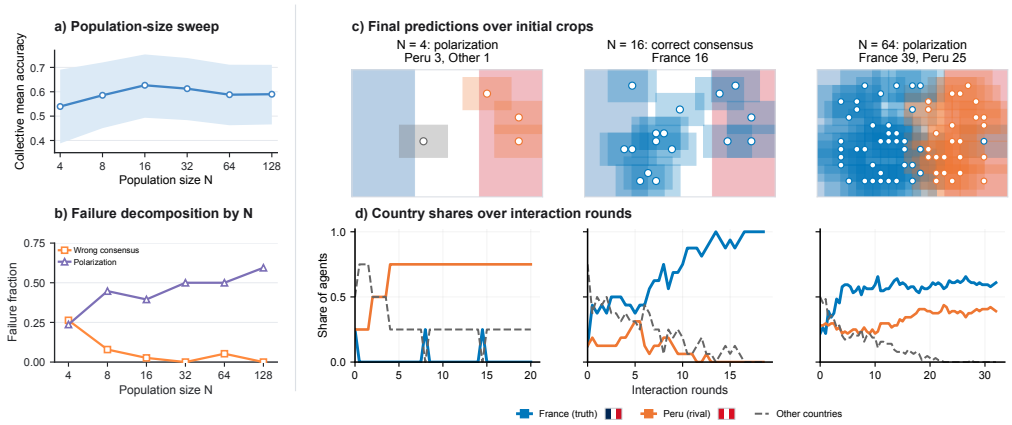

Figure 2: **With large N, polarization reduces accuracy.** (a) In GPT-4o pairwise runs, collective accuracy peaks at intermediate N before starting to fall. (b) As N grows, wrong consensus becomes less common while polarization becomes more frequent. (c–d) A representative France–Peru run at $N = 4, 16, 64$ shows final predictions over initial crop locations and country-share trajectories.

Team diversity. See Fig. 1d. The composition sweep tests whether collective performance increases with using more of the visually strongest individual model. We see that this is not the case, as the best-performing teams are mixed across GPT-4o and GPT-5.4 agents. This pattern suggests complementarity rather than a simple model ranking. In Fig. 3a, the vision probe shows a behavioral difference between the models. GPT-4o and 5.4 receive the same flag crop and while they have similar country accuracy, their errors differ. GPT-4o's incorrect responses are more visually compatible to the crop, while GPT-5.4 produces more incompatible guesses. GPT-4o therefore appears to be more locally anchored to visual evidence. The memory probe in Sec. 5.1 reveals a second difference in how the two models update, further explaining the mixed-team gains with complementary skills.

Organizational structure. See Fig. 1e. The protocol sweep with population $N = 8$ and fixed crops varies how social input flows to agents, pairwise (local), broadcast (public), or manager. The clearest contrast is the model identity of the manager. With fixed observers, GPT-5.4 managers reach 0.60 accuracy, whereas GPT-4o managers only get to around 0.46, reflecting a role-specific difference in how managers synthesize reports. In manager runs, the identity of the blind synthesizer is especially important; in population protocols, the same agents both observe and update, so performance depends on the joint dynamics of private evidence, social uptake, and communication structure.

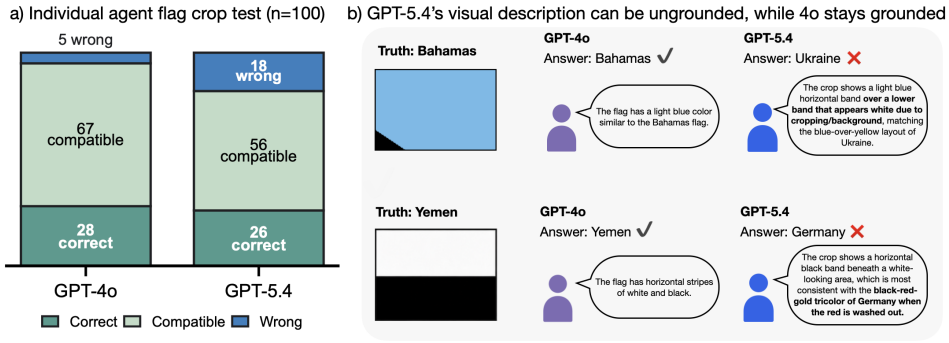

Figure 3: **Single-agent test reveals complementary error modes behind diverse team gains.** (a) Isolated country responses for GPT-4o and GPT-5.4 over 100 runs. Exact accuracy is similar, but GPT-5.4 produces more visually incompatible answers. (b) Examples illustrating cases where GPT-4o chooses the correct country while GPT-5.4 misconstrues its private evidence.

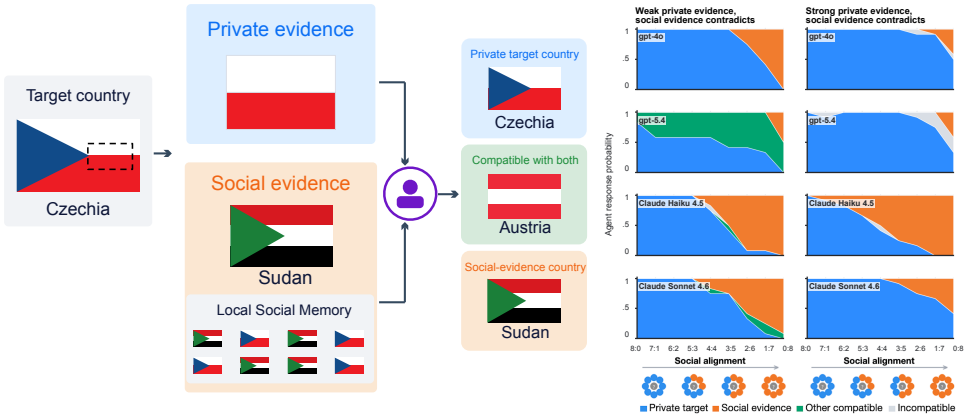

Figure 4: **Local social memory probe isolates update behavior conditional on private evidence.** (a) A target flag gives private evidence to the agent, while memory contains a shuffled mixture of target and social country entries. (b) Agent responses as the memory ratio shifts in two regimes: weak (left), where multiple countries match, and strong (right), where only the target country matches.

5 Mechanistic Swarm Interpretability

Next, we try to understand the mechanisms underlying non-monotonic accuracy as population scales (Fig. 2b). In the France–Peru example (Fig. 2c–d), $N = 4$ lacks decisive evidence, $N = 16$ reaches correct consensus, and $N = 64$ splits between truth and rival as more observers add support for both at once. This raises questions at two scales. For an individual agent, how does it weigh what it sees against what it hears, and how far does its evidence travel (Sec. 5.1)? Across the population, why do outcomes change with scaling N (Sec. 5.2)?

5.1 Causal Interventions on Agents

Mechanistic interpretability works by intervening on a network's internals, fixing an activation and observing downstream change (Meng et al., 2022; Zhang & Nanda, 2024). We apply the same logic to a swarm, whose internals are its agents. Each agent has two inputs, the private evidence in its crop and the social evidence in its memory. Editing memory shows when it holds its private evidence and when it copies. Editing an agent's crop traces evidence through the swarm. Together they show what an agent does with what it hears, and how far what one agent says can travel.

Local social memory probing. Whether an agent resists social evidence depends on the model as much as the crop. Fig. 4 tests these update behaviors by fixing a private crop, while the target:social ratio in an agent's memory changes. Under weak private evidence, GPT-5.4 shows the highest rate of integrative reasoning, routing probability into other countries rather than copying the social label. Under strong evidence (crop uniquely identifies target), GPT-4o and 5.4 hold firm, while Haiku 4.5 abandons the private target as social memory accumulates, a signature of sycophantic override.

This effect helps explain the team diversity result of Sec. 4: pairing a literal listener with an integrative listener can identify truth-supporting evidence that neither homogeneous team has, consistent with the gains in Fig. 1d.

Social circuit attribution. Using Germany, we choose one informative crop that provides a consistent cue where over ten isolated single-agent probes without social input, it elicited Germany in 10/10 judgments. Before patching this crop into any agent, we quantify each agent's predicted influence by combining the crop's accuracy gain Δp_i (informative crop ac-

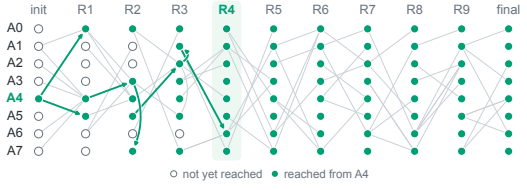

Figure 5: **Social circuit attribution.** Earliest-arrival paths of A4's message through the communication schedule (columns: rounds, rows: agents) A line within a column marks a same-round relay.

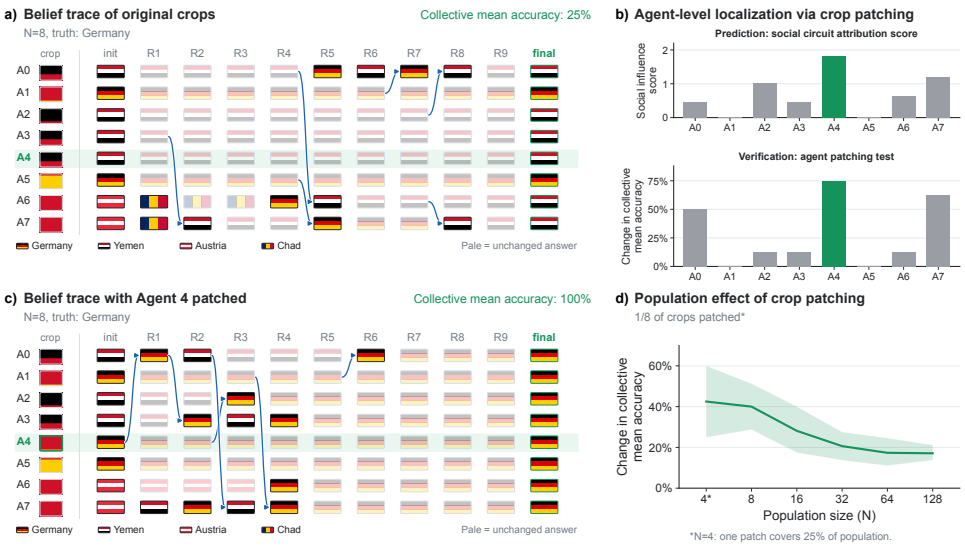

Figure 6: **Social circuit attribution and its causal verification by agent patching.** (a) Belief trajectories for a $N = 8$ Germany game using original crops. (b) Predicted influence scores (top) and empirical test patching each agent's crop separately (bottom) both identify Agent 4 as producing the largest collective accuracy improvement. (c) Belief trajectories under the same communication schedule with A4 patched. (d) Change in collective mean accuracy with Agent 4's new crop, averaged across ten runs. One crop patched at $N = 4$, and one-eighth of crops at $N \geq 8$.

curacy based on ten probes - initial accuracy based on ten probes) with its temporal closeness E_i (Pan & Saramäki, 2011), how quickly information could reach other agents, directly or through intermediaries. Their product, $S_i = \Delta p_i E_i$, is the social circuit attribution score. Although A0, A3 and A4 have identical original crops, and hence identical Δp_i, their positions in the social circuit differ based on the communication schedule (Fig. 5 has nodes fill green once A4's message has reached an agent, which happens for every peer by R4.), and their resulting influence score ranks A4 highest (Fig. 6b, top). Appendix E describes our method for ranking these transmission routes.

Verification by agent patching. We then empirically test the intervention our influence score ranks. We patch each agent's crop separately with the same informative crop, holding the communication schedule fixed, and measure the change in final collective mean accuracy. At $N = 8$, patching A4 produces the largest observed improvement (Fig. 6b, bottom), matching our method's ranking.

Agent-level causal tracing via crop patching. The selected paired traces illustrate the result: collective mean accuracy increases from 25% with original crops to 100% with A4 patched (Fig. 6a,c). Arrows mark country switches consistent with the latest received message, showing routes along which the patch may have spread. We then repeat the comparison across ten runs, evaluating collective mean accuracy after ten rounds. For $N = 8-128$, we patch the same proportion of agents $(1/8)$. The mean improvement decreases from 40% at $N = 8$ to approximately 17% at $N = 128$ (Fig. 6d). Thus, the same intervention fraction produces less collective correction at larger populations.

From agents to populations. Fig. 6d shows how a fixed share of planted evidence has a smaller impact on collective performance as N grows. At small N, the tools from mechanistic interpretability suffice, since a patch on one agent shifts the outcome and an interaction trace shows how. As N grows, the same patch has a smaller effect and the swarm enters a regime where collective belief is a property of the population rather than of the agents in it. That regime calls for a statistical mechanical view, in which collective belief follows from the population composition and size rather than any individual members. We develop this next and return with it to the population sweep of Fig. 2b.

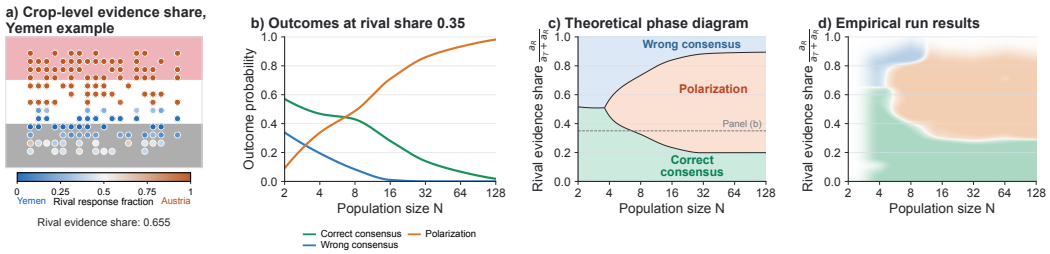

Figure 7: **Private evidence and collective outcomes across population size.** (a) Rival response fractions from 50 binary probes per crop in an $N = 128$ Yemen run: blue supports Yemen, orange supports Austria. (b) Outcome probabilities at a rival evidence share $= 0.35$ with $a_T + a_R = 0.45$ and $h_0 = +0.3$. (c) Theoretical phase diagram of population size and rival evidence share. (d) Empirical run results from population sweep, using each run's probed rival evidence share. Regions show the locally most frequent outcome, estimated with Gaussian bandwidth 0.05 in share.

5.2 Statistical mechanics of bounded agents

Here we present a simple phenomenological theory that captures three key features of the experiment: as population size increases, collective belief collapse decreases, collective belief polarization increases, and collective performance is highest at an intermediate population size. We start by adding private evidence from the external world to Quantized Simplex Gossip (Tanaka, 2026).

Private evidence. We simplify the experiment to two beliefs, the truth country T and one rival country R. A crop can favor either belief or leave the agent uncertain, as in Fig. 7a's Yemen–Austria example, black distinguishes Yemen from Austria, a red-and-white crop can be read as Austria, and an ambiguous crop leaves many possibilities open. We describe crops as three types: truth-deciding, rival-deciding, and ambiguous. Each agent independently draws a type with probabilities a_T, a_R, and $a_0 = 1 - a_T - a_R$. These are probabilities under crop sampling, rather than literal fractions of the flag's area. To compare this with our multi-country experiments, we probe each initial crop used in our population sweep 50 times restricting country answers to T and R. The fraction of valid responses choosing R, pooled across crops, sets the rival evidence share $a_R/(a_T + a_R)$.

Social updating. Agents with truth- or rival-deciding evidence mainly keep their initial belief. They are evidence-induced *zealots*, or agents whose private evidence resists social pressure. Ambiguous agents instead copy a randomly chosen speaker, adopting a heard T with probability q_T and a heard R with probability $q_R = 1 - q_T$; otherwise they retain their current belief. We write $q_T/q_R = e^{h_0}$. When $h_0 = 0$, this is neutral copying, with an equal probability of accepting either label. When $h_0 < 0$, ambiguous agents lean toward the rival. This bias represents an interpretation tendency under uncertain evidence; here it is a modeling assumption. The model extends the neutral copying limit of QSG with evidence-induced zealots (Mobilia, 2003; Mobilia et al., 2007) and biased adoption.

Dynamics. In a population of N agents, let z_T, z_R, and $z_0 = N - z_T - z_R$ count truth zealots, rival zealots, and ambiguous agents, with fractions $f_X = z_X/N$; a_X is a sampling probability and f_X the fraction actually drawn. Each update samples a speaker and a listener uniformly, and only an ambiguous listener can change. With x the truth fraction among ambiguous agents, $s = f_T + f_0 x$ the truth fraction in the population, the exact transition rates (Appendix F) give the mean-field drift:

$$\frac{dx}{d\tau} = A(x) = q_T f_T (1 - x) - q_R f_R x + (q_T - q_R) f_0 x (1 - x). \tag{1}$$

The first term is ambiguous agents copying truth zealots, the second copying rival zealots, and the third copying one another, which vanishes under neutral copying and favors truth when $h_0 > 0$. Copying fluctuations scale as $N^{-1/2}$ at fixed type fractions with $f_0 > 0$. Larger populations follow mean-field dynamics more closely, but finite populations can continue changing even if $A(x) = 0$.

Evidence coverage. At fixed a_T, a_R, increasing N samples more evidence from the same distribution. The probability of drawing no truth zealot is $(1 - a_T)^N$ and no rival zealot $(1 - a_R)^N$, so each type typically appears once $Na_T \sim 1$ or $Na_R \sim 1$. One can picture $1/N$ as a waterline covering the sampling probabilities a_T and a_R. As N increases, the water recedes, illustrating how commonly

sampled evidence tends to appear before rarer evidence. When $a_T > a_R$, truth evidence therefore tends to appear first. These scales organize three overlapping population phases, whose boundaries are finite-population crossovers. Fig. 7b–c shows the endpoint probabilities and regions identifying the most probable collective outcome.

Memetic-drift phase. At small N, many populations contain no zealots. Setting $f_T = f_R = 0$ and $h_0 = 0$ in Eq. 1 gives $A(x) = 0$ where neither label has a deterministic advantage. Copying fluctuations nevertheless lead each finite population to consensus by chance, the memetic-drift mechanism of neutral QSG. Truth-biased copying reduces, but does not eliminate, the chance of wrong consensus. Small populations containing rival zealots but no truth zealots instead reach wrong consensus through persistent rival evidence. Both routes can produce collective belief collapse.

Wisdom-of-crowds phase. At intermediate N, truth-deciding agents are increasingly present while rival-deciding agents are still scarce. With $f_R = 0$, Eq. 1 has the fixed point $x^* = 1$, which is locally stable when $q_T f_T > (q_R - q_T) f_0$. Without rival zealots, a finite population containing truth zealots eventually reaches correct consensus: communication spreads the evidence of a few agents to the rest (outside the stability condition this still happens, but slowly; Appendix F).

Polarization phase. At larger N, both kinds of zealots are typically present when $a_T, a_R > 0$. For $f_0 > 0$, $A(0) > 0$ and $A(1) < 0$, neither unanimous state is sustained, and the mean-field dynamics has a stable interior fixed point. Under neutral copying, $x^* = f_T/(f_T + f_R)$; rival-biased adoption shifts it downward. Neither zealot type can convert the other, and their repeated messages sustain competing beliefs among ambiguous agents. This split keeps the population away from perfect convergence but retains some of the truth within the collective. At fixed positive evidence fractions, the fluctuations shrink with N, making the split more persistent. Whether a particular split meets the polarization thresholds also depends on the relative sizes of the two camps. A truth-dominated split may still end at correct consensus, so the green region need not vanish at large populations.

Accuracy peak. With fixed $a_T > a_R$, neutral copying increases mean truth share monotonically with N, so evidence coverage alone does not explain the mean accuracy peak in this model. In populations with more truth zealots than rival zealots, truth can still spread widely; when both types are similarly represented, the same bias shifts ambiguous agents toward the rival and lowers the truth share. Averaging model expectations at each run's measured evidence share reproduces an intermediate-population accuracy peak (Appendix F). The model separates two roles of private evidence: it can correct an arbitrary consensus and hold a locally plausible false belief in place.

6 Conclusion

We introduced the Flag Game, a toy model for mechanistic swarm interpretability, and found that collective belief collapse crosses over to collective belief polarization as the number of agents increases. Although polarization lowers mean accuracy, it can be a less dangerous failure mode than belief collapse, as a population that collapses onto a false belief has nothing left with which to correct itself, whereas a polarized population retains competing beliefs. In that sense, polarization can be a first step toward plurality. For safety purposes, what to watch out for may therefore not be disagreement, but consensus of agents' beliefs or intent under social pressure.

Overall, this emergence of a sociology of AIs suggests that aligning AI swarms may require a different paradigm to aligning an individual AI agent. Just as qualitatively different mechanisms emerge in collections of atoms, collective systems of agents may require new variables and new laws of description. The state we want to achieve may itself need to be specified collectively, for example, as a plural state characterized by diversity of beliefs rather than consensus under social pressure.

Causal interventions and statistical mechanics rest on opposite assumptions about how a swarm is organized. In the Hugging Face incident, key beliefs and coordination structures emerged while the population was still relatively small, where social circuit attribution and agent patching are most informative. As a population grows, local interventions lose their effect and population-level variables such as evidence coverage take over. Mechanistic swarm interpretability may therefore need to be scale-adaptive, reverse engineering social circuits during early swarm formation and controlling the collective phases that emerge later, ultimately enabling the inverse design of swarms toward desired collective states. As intelligence scales from networks of neurons to larger networks of interacting agents, we hope our empirical and theoretical framework can help us climb the ladder of complexity.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets through API calls to models serving as the subjects of our Flag Game experiments. Their country predictions, explanations, and communication transcripts constituted the experimental data analyzed in this work. Through extensive iterative discussion, AI helped develop theoretical models or conceptual frameworks for the statistical mechanics framework, formulate mathematical claims, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs, specifically the mathematical derivations presented in the paper. We also used AI for designing or providing feedback on experimental methodology, primarily through creating and refining prompts, and to implement methods through edits to our experimental code.

We have not used generative AI tools to propose or refine hypotheses for the empirical experiments, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results from the experiments. Translation was not applicable to this work.

Additionally, we used generative AI tools to edit software code, including code for creating scientific figures from experimental data, and to edit our paper draft to improve readability. The initial manuscript was written by the authors; AI assisted with final rewrites of selected passages through an iterative process.

We have reviewed all AI-assisted work. We inspected experimental prompts and response logs, tested the implementation, checked reported statistics and figures against the underlying data, and verified mathematical derivations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have added an anonymous GitHub link to all the code needed to reproduce the figures and the underlying data that was used in this paper, as well as run any additional experiments. In Appendix C, we have also outlined the specific controls used per run for our empirical panels of Sec 4. In Appendix E and F we have outlined additional information on the approach used for our social circuit attribution score of Sec 5.1 as well as for the statistical mechanics theory in Sec 5.2.

REFERENCES

Philip W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393.

Solomon E. Asch. Effects of group pressure upon the modification and distortion of judgments. In Harold Guetzkow (ed.), *Groups, Leadership and Men: Research in Human Relations*, pp. 177–190. Carnegie Press, Pittsburgh, PA, 1951.

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992. doi: 10.2307/2118364.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. doi: 10.1086/261849.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

X. Chen, H. Yi, M. You, W. Liu, L. Wang, H. Li, X. Zhang, Y. Guo, L. Fan, G. Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *npj*

Digital Medicine, 8(1):159, 2025. doi: 10.1038/s41746-025-01550-0. URL https://www.nature.com/articles/s41746-025-01550-0.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

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

Franz Dietrich and Christian List. Probabilistic opinion pooling generalized. part one: General agendas. *Social Choice and Welfare*, 48:747–786, 2017. doi: 10.1007/s00355-017-1034-z.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

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.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. Technical report, METR and Redwood Research, August 2026. URL https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/. Full report: https://metr.org/hugging-face-incident-report-aug-2026.pdf.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9399–9416. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.findings-emnlp.549. URL https://aclanthology.org/2024.findings-emnlp.549/.

Cong Jiang and Xiaolei Yang. AgentsBench: A multi-agent LLM simulation framework for legal judgment prediction. *Systems*, 13(8):641, 2025. doi: 10.3390/systems13080641. URL https://www.mdpi.com/2079-8954/13/8/641.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.606. URL https://aclanthology.org/2025.findings-acl.606/.

Vira Kasprova, Amruta Parulekar, Abdulrahman AlRabah, Krishna Agaram, Ritwik Garg, Sagar Jha, Nimet Beyza Bozdag, and Dilek Hakkani-Tur. Too polite to disagree: Understanding sycophancy propagation in multi-agent systems. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 795–814, Atlanta, Georgia, USA, 2026. Association for Computational Linguistics. URL https://aclanthology.org/2026.sigdial-1.56/.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Gustave Le Bon. *Psychologie des foules*. Félix Alcan, Paris, 1895. English translation: *The Crowd: A Study of the Popular Mind*, T. Fisher Unwin, 1896.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. arXiv:2402.05120.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Charles Mackay. *Memoirs of Extraordinary Popular Delusions*. Richard Bentley, London, 1841. 3 vols. Retitled *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds* in the 2nd ed., 1852.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Mauro Mobilia. Does a single zealot affect an infinite group of voters? *Physical Review Letters*, 91 (2):028701, 2003. doi: 10.1103/PhysRevLett.91.028701.

Mauro Mobilia, Alexander M. Petersen, and Sidney Redner. On the role of zealotry in the voter model. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(08):P08029, 2007. doi: 10.1088/1742-5468/2007/08/P08029.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

OpenAI. The Hugging Face incident and the road ahead. https://openai.com/index/hugging-face-incident-and-the-road-ahead/, August 2026. Technical report on the July 2026 incident. Initial disclosure: https://openai.com/index/hugging-face-model-evaluation-security-incident/ (July 21, 2026).

Raj Kumar Pan and Jari Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical Review E*, 84(1):016105, 2011. doi: 10.1103/PhysRevE.84.016105.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955. doi: 10.2307/1884852.

Herbert A. Simon. *Models of Man: Social and Rational. Mathematical Essays on Rational Human Behavior in a Social Setting*. Wiley, New York, 1957.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478, 1985. doi: 10.1037/0022-3514.48.6.1467.

Garold Stasser and William Titus. Hidden profiles: A brief history. *Psychological Inquiry*, 14(3–4): 304–313, 2003. doi: 10.1080/1047840X.2003.9682897.

Rush T. Stewart and Ignacio Ojea Quintana. Probabilistic opinion pooling with imprecise probabilities. *Journal of Philosophical Logic*, 47:17–45, 2018. doi: 10.1007/s10992-016-9415-9.

Hidenori Tanaka. When is collective intelligence a lottery? multi-agent scaling laws for memetic drift in llms, 2026.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33. URL https://aclanthology.org/2024.findings-acl.33/.

Sydney Von Arx, Cormac Slade Byrd, Spencer Kitts, and Thomas Larsen. Discovery of a new OpenAI agent message board. Technical report, Nightingale, September 2026. Report on the DseWiki incident (May–June 2026). Press coverage: Reuters, September 4, 2026.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi: 10.1126/science.aap9559.

Zhun Wang, Nico Schiller, Hongwei Li, Srijiith Sesha Narayana, Milad Nasr, Nicholas Carlini, Xiangyu Qi, Eric Wallace, Elie Bursztein, Luca Invernizzi, Kurt Thomas, Yan Shoshitaishvili, Wenbo Guo, Jingxuan He, Thorsten Holz, and Dawn Song. ExploitGym: Can AI agents turn security vulnerabilities into real attacks?, 2026.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. arXiv:2206.07682.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025. doi: 10.48550/arXiv.2509.05396. URL https://arxiv.org/abs/2509.05396.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025. doi: 10.48550/arXiv.2509.23055. URL https://arxiv.org/abs/2509.23055.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

A FULL SINGLE AGENT MEMORY PROBE

Fig. 8 extends the memory-conflict probe in Fig. 4 to the following three regimes: weak private evidence (indicating multiple countries are compatible with the crop) and compatible social evidence, weak private evidence and incompatible social evidence, and strong private evidence (indicating only the target country is compatible with the crop) and incompatible social evidence. GPT-5.4 routes substantially more probability mass into other crop-compatible countries (green), whereas GPT-4o, Claude Haiku 4.5, and Claude Sonnet 4.6 update more literally between the private target and the social-evidence country. At m=3, when social memory includes reasoning, Haiku and Sonnet also start to respond with compatible alternatives, but the effect remains strongest in GPT-5.4. Strong private evidence (rightmost column) shows that GPT-4o, GPT-5.4, and Sonnet hold firm on the target, whereas Haiku is more readily moved by social-heavy memory even with strong private evidence.

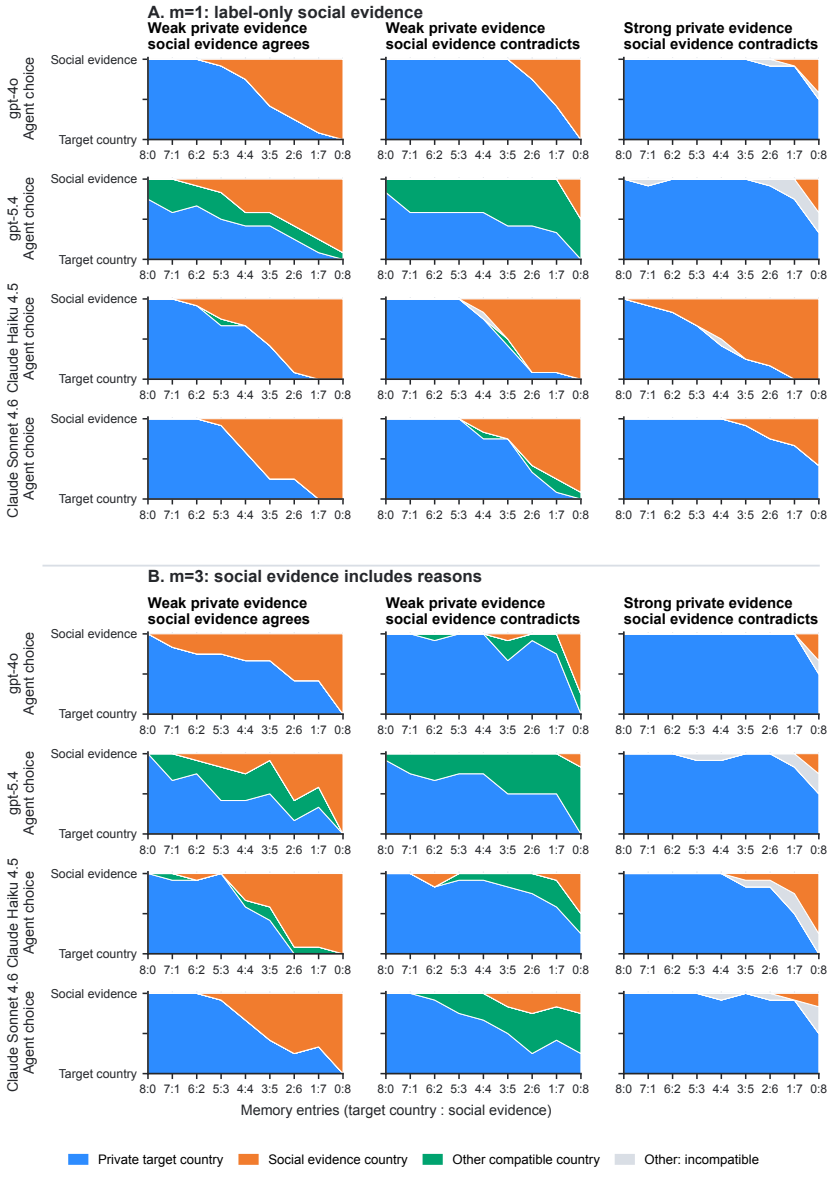

Figure 8: **Memory-conflict probe across four models.** Agent responses as local social memory shifts from target-heavy (8:0) to social-heavy (0:8), across three private-evidence regimes and two message bandwidths (m=1 label-only; m=3 with reasons).

B PAIRWISE ALPHA SWEEP

Running the same prompted social uptake sweep as in Fig. 1c except using the pairwise protocol we see a slightly different trend, where instead of monotonic growth, we have an interior optimum, indicating that under local exchange, when not all required information is present, higher uptake can create overdependence on incorrect speakers.

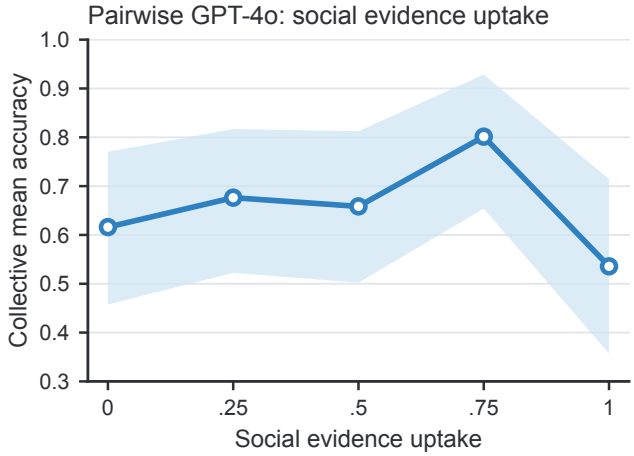

Figure 9: The pairwise protocol has an interior optimum at a prompted social evidence uptake of 0.75.

C EMPIRICAL RUN DETAILS

Table 1: Run settings behind the empirical panels in Sec. 4.

Figure slice	Protocol	Main controls	Trials used
Population scaling	Pairwise	$m = 3$; no social-awareness prompt	40 seeds per (model, N)
Protocol side-by-side	Pairwise, broadcast, manager	$N = 8$; $m = 3$; no social-awareness prompt	60 matched seeds per condition
Pairwise prompted social-awareness	Pairwise	GPT-4o; $N = 16$; $m = 3$	28 seeds per α
Broadcast prompted social-awareness and composition	Broadcast	$N = 8$; $m = 3$	30 seeds per $(\alpha, \text{composition})$
Memory-conflict probe	Single-agent prompt probe	GPT-4o, GPT-5.4, Claude Haiku 4.5 and Sonnet 4.6; $m \in \{1, 3\}$	36 trials

Here in Table 1, we record the run settings behind the empirical panels in Sec. 4. All runs use temperature 0.2, top-p 1.0, and a user message (example provided) consisting of the text prompt and the private crop image using high image detail. Images are rendered on a 24×16 canvas and agents see 6×4 crops at render scale 25. Memory buffers store at most $H = 8$ prior entries per agent. Each trial samples a hidden country uniformly from the country pool of 28 stripe and triangle flags. Additionally, in Table 2 we show the robustness of our endpoint results with different thresholds used for consensus and polarization. Entries give minimum–maximum percentages over 30 different combinations of consensus thresholds 0.75–1.00 and polarization thresholds 0.15–0.35, each in steps of 0.05. Ranges measure threshold sensitivity, not statistical uncertainty. The values we settled on for our analysis are 0.85 for consensus, both correct and wrong, and 0.25 for polarization.

Model composition. Mixed $N = 8$ conditions assign four observer agents to GPT-5.4 and four to GPT-4o. In the broadcast composition sweep, the GPT-5.4 count ranges from 0 to 8 while the

Table 2: Endpoint-threshold robustness on the pairwise communication protocol for GPT-4o.

N	Correct cons. (%)	Wrong cons. (%)	Polarized (%)	Fragmented (%)
4	50.0–52.6	26.3–34.2	13.2–23.7	0.0–10.5
8	47.4	5.3–23.7	26.3–44.7	0.0–21.1
16	50.0–52.6	0.0–10.5	18.4–44.7	0.0–31.6
32	42.1–47.4	0.0	31.6–55.3	0.0–26.3
64	36.8–42.1	0.0–5.3	34.2–55.3	0.0–28.9
128	37.8–40.5	0.0	56.8–59.5	0.0–5.4

remaining agents are GPT-4o. Manager conditions have $N = 8$ crop-bearing observers plus one blind manager synthesizer. The manager slot is either GPT-4o or GPT-5.4; the observer group is all GPT-4o, all GPT-5.4, or a 4/4 mix.

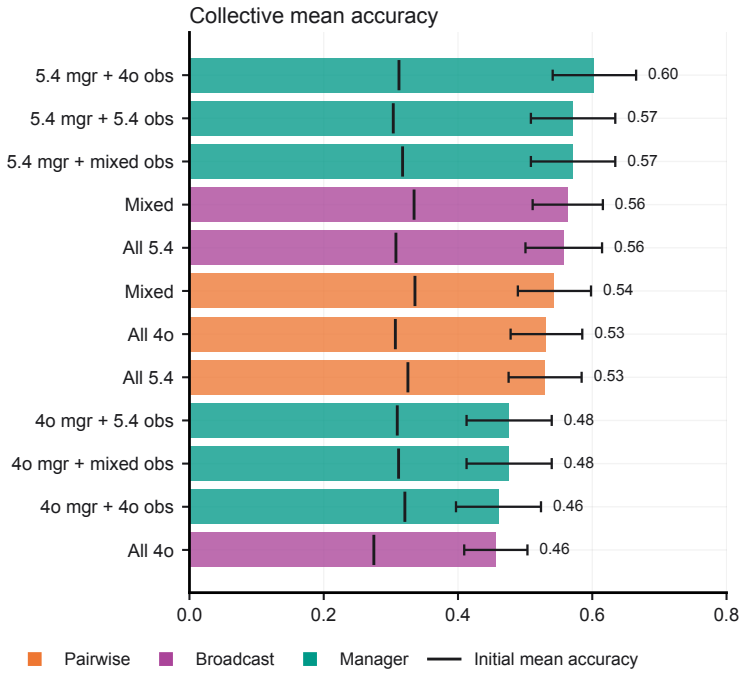

Figure 10: **Protocol and composition sweep at $N = 8$.** A comparison of collective mean accuracy across the three communication protocols and varying model compositions, sorted by accuracy; Fig. 1e shows the all-GPT-4o-observer subset. Ticks mark initial mean accuracy; error bars are ± 1 s.e.m. over 60 matched seeds per condition.

D PROMPTING AND SOCIAL-AWARENESS INTERVENTION

All protocol prompts specify a JSON-only answer and attach the private crop image to the same user message (unless the agent is a manager). The message bandwidth m determines the JSON schema: $m = 1$ asks only for a country and $m = 3$ asks for a country and one-sentence reason.

System message

```
You must output only valid JSON. No extra keys, no markdown,
and no text outside the JSON object.
You are one player in a flag identification game.
```

```
Choose exactly one country from the allowed countries listed in the
    user message.
Follow the exact output schema given in the user message.
```

User message

```
All players are identifying the same underlying flag.
You always see the same private crop.
Transcript memory shows messages you observed from previous
    interactions with other players.

Allowed countries: <JSON list>
Transcript memory (oldest -> newest):
<speaker messages, or []>
<optional social-awareness line>

Output JSON exactly:
{"country":"<one allowed country>","reason":"<one sentence>"}
```

Figure 11: **Example pairwise prompt.** The pairwise protocol uses the same text for interaction messages and probe queries, differing only in the schema line.

Social-awareness line. By default `prompt_social_susceptibility` is disabled, but when enabled, the prompt inserts guidance based on specified α (Table 3) as seen in the experiment in Fig 1c.

Table 3: Social-awareness prompt ladder indexed by α.

Range	Pairwise wording	Broadcast wording
$\alpha \leq .2$	Rely mostly on your own crop and treat transcript memory as weak evidence.	Rely mostly on your own evidence; treat other agents' country guesses as weak evidence.
$.2 < \alpha \leq .4$	Give somewhat more weight to your own crop than to transcript memory.	Give somewhat more weight to your own evidence than to other agents' country guesses.
$.4 < \alpha \leq .6$	Balance your own crop and transcript memory.	Balance your own evidence with other agents' country guesses, using their guesses as real evidence.
$.6 < \alpha \leq .8$	Give somewhat more weight to transcript memory than to your own crop.	Give somewhat more weight to other agents' country guesses than to your own evidence.
$\alpha > .8$	Treat transcript memory as strong evidence and update readily toward it.	Treat other agents' country guesses as strong evidence and update readily toward them.

Memory-conflict probe prompt. The memory-conflict probe reuses the pairwise prompt with no live social interaction. For each trial, the agent receives one crop from a target country and a synthetic memory of eight entries. If k is the false-memory count, then k memory entries name a lure country and $8 - k$ entries name the target country, shuffled before prompting.

E TEMPORAL COMMUNICATION ANALYSIS

We summarize opportunities for information to spread using time-ordered communication paths (Pan & Saramäki, 2011). Let $\tau_{ij}(0)$ be the earliest global message index at which information starting at agent i at initialization could reach agent j, directly or through intermediaries. Contacts must occur in chronological order, with information allowed to wait between them. We report mean

arrival time D_i and temporal closeness E_i, the mean inverse arrival time:

$$D_i = \frac{1}{N-1}\sum_{j\neq i}\tau_{ij}(0), \qquad E_i = \frac{1}{N-1}\sum_{j\neq i}\frac{N}{\tau_{ij}(0)}.$$

D_i is measured in interactions; dividing arrival times by N expresses them in rounds before taking their reciprocals in E_i. The latter follows the mean-inverse form of temporal closeness and efficiency, evaluated from initialization. Each recipient contributes once, according to its earliest arrival, so earlier access contributes more. An unreached recipient would contribute zero to E_i.

We combine this schedule-based measure with the accuracy gain from replacing agent i's crop:

$$\Delta p_i = \hat{p}_{\text{informative}} - \hat{p}_{\text{original},i}, \qquad S_i = \Delta p_i E_i.$$

Each original-crop and informative crop accuracy is estimated from ten probes. Thus, the social circuit attribution score S_i combines the additional local evidence of the patch with opportunities for its early dissemination. Table 4 breaks down the calculation across all agents, showing A4 has the highest predicted influence score and subsequently the largest observed patching effect.

Table 4: Agent-level metrics for the selected communication schedule. ΔA_i is the empirical change in collective mean accuracy relative to the original-crop baseline of 25%, from one game per patch.

Agent	Original correct (out of 10)	Δp_i	D_i	E_i	S_i	ΔA_i
A0	0	1	19.4	0.47	0.47	0.50
A1	10	0	14.9	0.73	0.00	0.00
A2	0	1	16.4	1.01	1.01	0.13
A3	0	1	22.9	0.46	0.46	0.13
A4	0	1	12.9	1.81	**1.81**	**0.75**
A5	10	0	24.6	0.39	0.00	0.00
A6	0	1	15.6	0.64	0.64	0.13
A7	0	1	16.6	1.19	1.19	0.63

F MINIMAL THEORY: FINITE-POPULATION DETAILS

Evidence composition. Let $(z_T, z_R, z_0) \sim \text{Multinomial}(N; a_T, a_R, a_0)$, where $a_0 = 1 - a_T - a_R$. The four mutually exclusive evidence compositions have exact probabilities

$$\begin{aligned} P_{\text{neither}} &= a_0^N, \qquad P_{T\text{-only}} = (1-a_R)^N - a_0^N, \\ P_{R\text{-only}} &= (1-a_T)^N - a_0^N, \\ P_{T,R} &= 1 - (1-a_T)^N - (1-a_R)^N + a_0^N. \end{aligned} \tag{2}$$

The probability of having no truth zealot combines unanchored populations and populations with rival zealots only, with total probability $(1-a_T)^N$. The scales $Na_T \sim 1$ and $Na_R \sim 1$ indicate when the corresponding evidence starts to appear; they are crossover estimates, not sharp boundaries or definitions of empirical endpoint classes.

Exact finite-population dynamics. Conditional on (z_T, z_R, z_0), let n count ambiguous agents that currently report T. Each update samples a speaker and a listener independently and uniformly; only an ambiguous listener can change, adopting a heard T with probability q_T or a heard R with probability q_R. The probabilities that n increases or decreases by one are

$$W_+(n) = q_T\frac{(z_0-n)(z_T+n)}{N^2}, \qquad W_-(n) = q_R\frac{n(z_R+z_0-n)}{N^2}. \tag{3}$$

For example, W_+ is the probability of choosing an ambiguous listener who currently identifies the flag as R, a truth-holder as speaker, and accepting the message. The remaining probability gives no change. Let $p_t(n)$ be the probability that n ambiguous agents report truth after t interaction steps. Then:

$$\begin{aligned} p_{t+1}(n) - p_t(n) = {} & W_+(n-1)p_t(n-1) + W_-(n+1)p_t(n+1) \\ & - [W_+(n) + W_-(n)]p_t(n). \end{aligned} \tag{4}$$

Terms outside $0 \le n \le z_0$ are zero. The theoretical panels in Fig. 7 use the stationary and fixation probabilities of this process, with $q_T = (1 + e^{-h_0})^{-1}$, averaged over the sampled evidence compositions.

Mean-field closure. For $z_0 > 0$, the exact one-step mean satisfies $\mathbb{E}[x_{t+1}] - \mathbb{E}[x_t] = \mathbb{E}[A(x_t)]/N$, conditional on the evidence composition. Replacing $\mathbb{E}[A(x)]$ by $A(\mathbb{E}[x])$ gives Eq. 1; when $h_0 \neq 0$, this is an approximation because A is nonlinear. Writing $\Delta x = x_{t+1} - x_t$, the one-step variance is

$$\mathrm{Var}(\Delta x \mid n) = \frac{W_+ + W_- - (W_+ - W_-)^2}{z_0^2}. \tag{5}$$

The drift and fluctuations therefore follow from the same update rule. At fixed type fractions with $f_0 > 0$, fluctuations over a fixed interval of $\tau = t/N$ scale as $N^{-1/2}$

Memetic drift and truth spreading. Without zealots, the probability of eventual truth consensus from n initial truth reports is

$$\Pr(T \text{ fixes} \mid n) = \begin{cases} n/N, & h_0 = 0, \\ \dfrac{1 - e^{-h_0 n}}{1 - e^{-h_0 N}}, & h_0 \neq 0. \end{cases} \tag{6}$$

Under neutral copying, finite populations therefore reach consensus by chance despite $A(x) = 0$. Positive bias raises the truth-fixation probability. We average over independent fair initial beliefs, $n \sim \mathrm{Binomial}(N, 1/2)$.

For $0 < q_T, q_R < 1$, a finite population with truth zealots and no rival zealots eventually reaches truth consensus. In the truth-only case, $A'(1) = -q_T f_T + (q_R - q_T) f_0$, so truth consensus is locally stable when $q_T f_T > (q_R - q_T) f_0$. If this inequality is strictly reversed, the mean-field dynamics can settle at a mixed state. Random fluctuations still eventually bring the finite population to truth consensus, but this may take longer than the available interaction budget.

Mean accuracy. Under neutral copying, populations with zealots have expected truth share $z_T/(z_T + z_R)$. Averaging over compositions with unbiased initial beliefs gives:

$$\mathbb{E}[s_\infty] = \frac{a_T}{a_T + a_R}(1 - a_0^N) + \frac{1}{2} a_0^N. \tag{7}$$

For fixed $a_T > a_R$, this is nondecreasing with N: evidence coverage alone does not produce a mean-accuracy peak in the neutral model. Mean truth share differs from correct-consensus probability. Figure 7 uses truth-biased copying, $h_0 = +0.3$, and its fixed-share slice likewise has no mean-accuracy peak. Averaging model expectations at each run's measured share produces a maximum at $N = 32$, similar to the empirical cohort. This comparison averages over the measured evidence shares at each N.

Persistent competition. With both zealot types and $z_0 > 0$, $A(0) > 0$ and $A(1) < 0$, giving a stable interior fixed point. The exact stationary probabilities satisfy

$$\frac{\pi_{n+1}}{\pi_n} = \frac{W_+(n)}{W_-(n+1)} = e^{h_0} \frac{(z_0 - n)(z_T + n)}{(n+1)(z_R + z_0 - n - 1)}. \tag{8}$$

Normalization determines π_n. Under neutral copying this is the beta-binomial law of the zealot voter model (Mobilia et al., 2007), with

$$\mathbb{E}[x] = \frac{z_T}{z_T + z_R}, \qquad \mathrm{Var}(x) = \frac{z_T z_R N}{z_0 (z_T + z_R)^2 (z_T + z_R + 1)}. \tag{9}$$

At fixed positive fractions, fluctuations shrink as $N^{-1/2}$. Positive h_0 tilts the stationary weights toward truth and negative h_0 toward the rival. Both zealot types prevent unanimity, but an unequal split can still meet the consensus threshold. If $z_0 = 0$, the truth share remains fixed at z_T/N.