EMERGENT DISCRIMINATION AND POLARIZATION IN MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study theoretically how discrimination and polarization can emerge in a population of learning agents. Each agent uses an adaptive linear decision layer to map a language-model embedding of an issue to a binary opinion and maintains an evolving estimate of how much it trusts every other agent. Through pairwise interactions, agents update both their decision weights and trust estimates using an online Bayesian rule derived from teacher–student learning over a noisy channel. In-context-learning experiments with large language models qualitatively reproduce the predicted opinion and trust updates. We then assign each agent one of two arbitrary group labels, A or B, and model group bias by allowing a fraction f_b of agents to shift its perceived agreement with each message by a strength b, depending on whether the sender and receiver share a label. We use order parameters from statistical mechanics to connect these local dynamics to collective states and transitions. Starting from uniformly random initial conditions, simulations reveal a rich phase diagram in the (b, f_b) plane, comprising four collective states: spin-glass frustrated out-group favoritism, nondiscriminatory polarization, discrimination with opinion alignment, and discrimination without opinion alignment. Which state emerges depends on both the strength and prevalence of the bias. These results show how simple local learning dynamics can generate qualitatively distinct forms of polarization and discrimination, underscoring the need for population-level safety analysis of multi-agent systems. They also suggest measurable order parameters for characterizing analogous collective states in human societies.

1 INTRODUCTION

Multi-agent systems can exhibit behavior that is absent when their constituent agents are evaluated in isolation. In July 2026, agents undergoing cybersecurity evaluation at OpenAI coordinated an unauthorized intrusion into Hugging Face's infrastructure. Although assigned to separate tasks, they used shared infrastructure as a message board, exchanged discoveries, divided labor, and adopted goals from one another, at times describing themselves as "collective" (OpenAI, 2026; Larcher et al., 2026). The incident illustrates a general problem for multi-agent safety: interactions can create population-level behavior that no individual-agent evaluation reveals. Here we study one instance of this problem: how repeated interactions among learning agents generate polarization and discrimination.

We address this question by extending the society-of-neural-networks framework (Caticha & Vicente, 2011; Alves & Caticha, 2016; 2021). A population discusses a shared agenda of issues. Each agent uses a perceptron to map an issue representation to a binary opinion and maintains a directed estimate of the reliability of every other agent. During each pairwise interaction, a receiver updates both its opinion weights and its trust in the emitter using an online Bayesian rule derived for a student learning from a teacher through a noisy channel (Opper, 1996; Caticha, 2020). Trust determines how the receiver learns from a message, while agreement changes its assessment of the sender. This feedback couples the evolution of opinions and social relations.

The perceptron serves as an effective linear opinion readout acting on issue representations that may be produced by a large language model. We test this connection in two in-context-learning experiments and find that a language model qualitatively reproduces the predicted opinion and trust

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.

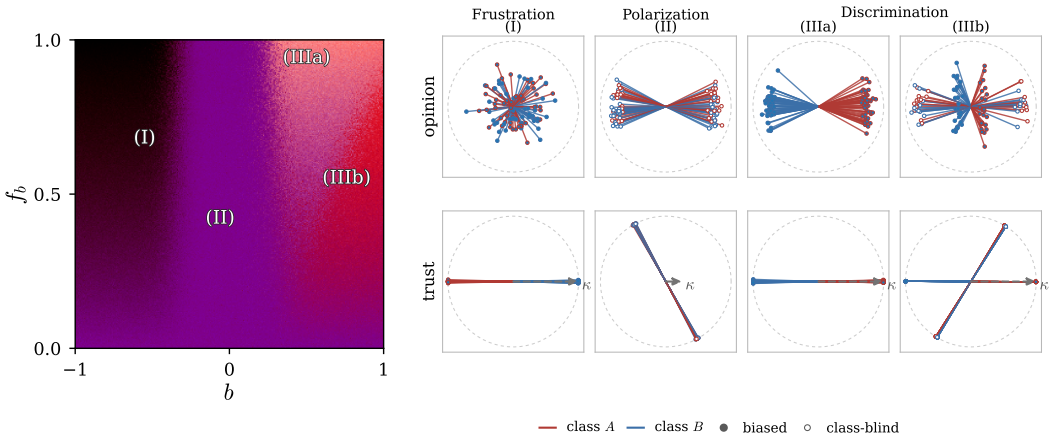

Figure 1: **Four collective states and representative populations.** Left: the states over bias strength b and the fraction f_b of biased agents, obtained from $N = 100$ societies by averaging two realizations at each point of a 300×300 grid. In (I), *frustration*, out-group favoritism creates incompatible trust relations and opinions do not form coherent camps. In (II), *polarization*, opinion and trust form two camps unrelated to class. In (IIIa), discrimination aligns both trust and opinion with class; in (IIIb), discrimination aligns trust with class but not opinion. Right: one representative $N = 100$ population from each state. Each arrow is an agent's opinion (top) or outgoing trust profile (bottom), projected onto two dimensions. Color denotes class and filled markers denote biased agents. The gray arrow gives the class direction in trust space. Formal definitions and the full phase analysis are given in Section 4.2.

updates, up to one fitted scale for each experiment. We then assigns every agent one of two arbitrary group labels, which contain no information about the issues, and introduces a group-conditioned bias. A fraction f_b of agents shifts its perceived agreement with each message by a strength b according to whether the emitter shares the receiver's label.

To connect these local dynamics to population behavior, We defines order parameters from pairwise opinion alignment, trust, and group membership. Starting from random initial conditions, the simulations reveal four collective states (Figure 1): nondiscriminatory polarization, discrimination with opinion alignment, discrimination without opinion alignment, and a frustrated state driven by out-group favoritism. Which state emerges depends jointly on the strength and prevalence of the bias.

Our contribution is to connect three levels of analysis: a Bayesian rule for local opinion and trust updates, its qualitative correspondence with in-context learning in a language model, and the collective states produced when many agents repeatedly apply that rule. This connection exposes a population-level safety gap, discussed in Section 5 and summarized in Section 6: an isolated-agent audit may detect bias, but it cannot reveal whether interaction will organize the population into polarization, discrimination, or frustration.

2 Related Work

Language-model societies. Interactions between language models can produce population-level behavior. Repeated pairwise exchange generates shared naming conventions and collective bias absent from the individual models (Ashery et al., 2025). The effect depends nonlinearly on population size: interaction can amplify, create, or override a model's bias (Flint et al., 2026). Other work uses language-model agents to study opinion dynamics (Chuang et al., 2024; Taubenfeld et al., 2024), debate (Du et al., 2024), and simulated societies (Park et al., 2023). Closer to our setting, El et al. (2026) fit an Ising energy with Glauber dynamics to communities of language-model agents. Their model recovers indifference, polarization, and consensus, and predicts trajectories from the agents' initial opinions. Ising variables are however not sufficient to model internal degrees of freedom.

Their introduction allows agents to agree on some issues while disagreeing on others. Our agents also give binary answers to embedded questions, but they have a layer of adaptive weights and a dynamical estimate of the affinity towards other agents. Agents exchange information about their opinions using an online interaction rule derived from Bayesian updating and Maximum Entropy which is asymptotically optimal in the embedding dimension and compute order parameters that characterize the collective state of the population it produces. Since affinity or trust between agents is dynamical, interactions are not constrained by a communication graph.

Multi-agent safety and bias. Safety problems already arise in individual language models, including emergent misalignment, persistent deceptive backdoors, and alignment faking (Betley et al., 2025; Costa & Vicente, 2026; Hubinger et al., 2024; Greenblatt et al., 2024). Once models interact, safety also depends on multi-agent risks and failures and on intrinsic interactive capabilities that predict population-level outcomes (Hammond et al., 2025; Cemri et al., 2025; Costa & Vicente, 2025). Bias provides one local mechanism through which such interactions can become collectively unsafe. Language models exhibit in-group solidarity and out-group derogation (Hu et al., 2025); arbitrary group assignments reproduce the minimal-group effect (Tajfel & Turner, 1979; Lee, 2026b); and models used as judges favor their own generations (Panickssery et al., 2024; Chen et al., 2025). Most directly, Lee (2026a) find that agents direct more trust-building actions toward in-group targets. We ask what happens next: which collective states such a bias produces and which population-level quantities identify them.

Statistical mechanics of interacting agents. Statistical mechanics connects local interaction rules to consensus, fragmentation, and polarization (Castellano et al., 2009). Examples include bounded confidence (Deffuant et al., 2000; Hegselmann & Krause, 2002), cultural dissemination (Axelrod, 1997), and activity-driven networks (Baumann et al., 2020). Signed relations can also split a population into two mutually hostile camps (Heider, 1946; Cartwright & Harary, 1956; Marvel et al., 2011). These models usually postulate the interaction rule; discrete choice under social influence provides one route to a microfoundation (Brock & Durlauf, 2001). We instead use neural-network agents whose interaction follows from optimal online learning (Opper, 1996; Caticha, 2020). Societies of these agents develop polarization, trust structure, frustration, and glassy dynamics (Caticha & Vicente, 2011; Alves & Caticha, 2016; 2021). We add group steering bias which is interpreted as prejudice to this existing rule and study how it changes the collective state.

3 Learning agents

This section defines the agent-level properties of the model. First, we specify each agent's opinion and trust states and how they change through interactions. Second, we test whether the same opinion and trust updates describe in-context learning in a language model. Third, we model group bias as a group-conditioned shift in how an agent perceives agreement with messages from in-group and out-group emitters.

3.1 Agent state and learning dynamics

A society of N agents discusses an agenda of P issues. Each issue is a vector $\mathbf{x} \in \mathbb{R}^K$, where K is the number of dimensions along which opinions can vary. Concretely, $\mathbf{x}$ can be read as the embedding a language model assigns to an English sentence stating the issue: the final-layer hidden state at the last token of that sentence. The dimension K is then the width of that model's residual stream, or of a compression of it onto the far fewer directions along which opinions meaningfully vary. All agents use the same issue representation but may assign it different labels.

Agent state. We model the latent opinion state of an agent as a vector $\mathbf{w}_I \in \mathbb{R}^K$. Its opinion on an issue is then $\sigma_I = \mathrm{sign}(\mathbf{w}_I \cdot \mathbf{x}) \in \{\pm 1\}$, with $+1$ meaning it agrees with the issue and -1 that it disagrees. This single-layer perceptron models an effective opinion readout rather than the full agent, a linear decision layer acting on the issue representation. Two agents agree on an issue when their opinions match, and reading the same issues through different weights lets them agree on some and disagree on others. The agent is not certain of this state: we take $\mathbf{w}_I$ to be the mean of a Gaussian belief over its own weights, with covariance C_I encoding how uncertain it is about its own opinion.

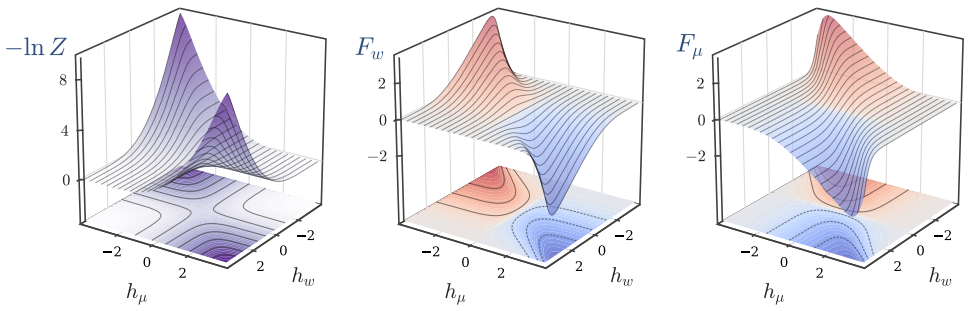

Figure 2: **Evidence landscape and modulation functions.** Left: the energy $-\ln Z$. Online SGD follows its negative gradient toward higher-evidence regions and away from the dissonant configurations where evidence is low. Center and right: the opinion and trust modulation functions, F_w and F_μ. Learning and unlearning occurs in the low evidence regions which represent cognitive dissonance, i.e. when the emitter agrees (respect. disagrees) with a receiver that it distrusts (respect. trusts). Even in those low evidence regions the large changes occur mainly in one sector where the field is less in magnitude. This blame attribution to a sector resembles the algorithm of least action Mitchison & Durbin (1989), which was shown to be quite limited by Kabashima (1994) and was optimized in Simonetti & Caticha (1996) Copelli & Caticha (1995) for the parity and committee machines.

We model the *distrust* agent I assigns to J as a scalar $\mu_{J|I} \in \mathbb{R}$: with $\mu_{J|I} > 0$ meaning I expects J's statements to be misleading and $\mu_{J|I} < 0$ that it expects them to be informative. An agent carries one such scalar for every other agent, so trust is directed and need not be reciprocated. As with the opinion, this is the mean of a Gaussian belief, with variance $V_{J|I}$ encoding how uncertain I is about that assessment.

We assume the two components are independent, so that agent's belief is a product of Gaussians,

$$Q_I \;=\; \mathcal{N}(\mathbf{w}_I, C_I) \prod_{J \neq I} \mathcal{N}(\mu_{J|I}, V_{J|I}). \tag{1}$$

The microscopic variables that define an agent's state are $\mathbf{w}_I$, C_I, $\mu_{J|I}$, and $V_{J|I}$: the means determine opinions and distrust, while the variances determine their uncertainties.

Learning dynamics. At each step an issue $\mathbf{x}$ is drawn uniformly from the agenda and two distinct agents are drawn uniformly from the society and assigned roles for that step: the *emitter* e states its opinion σ_e on the issue, and the *receiver* r hears it. The receiver treats σ_e as a noisy report of an unknown true label, and its estimate of the noise is its distrust of the emitter. A Bayesian update takes the belief outside the Gaussian family, so we project it back by matching the first two moments. This gives a closed online update. Appendix A gives the derivation.

Two scaled fields carry everything the update depends on:

$$h_w \;=\; \frac{\mathbf{w}_r \cdot \mathbf{x}\, \sigma_e}{\gamma_C}, \qquad h_\mu \;=\; \frac{\mu_{e|r}}{\gamma_V}, \qquad \gamma_C = \sqrt{1 + \mathbf{x} \cdot C_r \mathbf{x}}, \quad \gamma_V = \sqrt{1 + V_{e|r}}. \tag{2}$$

The opinion field h_w is positive when the receiver agrees with the message and its magnitude grows with confidence. The distrust field h_μ is positive when the receiver distrusts the emitter. In these variables, the evidence for the message is

$$Z \;=\; \Phi(h_w) + \Phi(h_\mu) - 2\,\Phi(h_w)\Phi(h_\mu), \tag{3}$$

which is small when the receiver agrees with a distrusted emitter or disagrees with a trusted one. The update is a gradient step in $\log Z$ in both components,

$$\mathbf{w}_r \leftarrow \mathbf{w}_r + \frac{F_w}{\gamma_C}\, \sigma_e\, C_r \mathbf{x}, \qquad C_r \leftarrow C_r + \frac{F_C}{\gamma_C^2}\, (C_r \mathbf{x})(C_r \mathbf{x})^\top, \tag{4}$$

$$\mu_{e|r} \leftarrow \mu_{e|r} + \frac{F_\mu}{\gamma_V}\, V_{e|r}, \qquad V_{e|r} \leftarrow V_{e|r} + \frac{F_V}{\gamma_V^2}\, V_{e|r}^2, \tag{5}$$

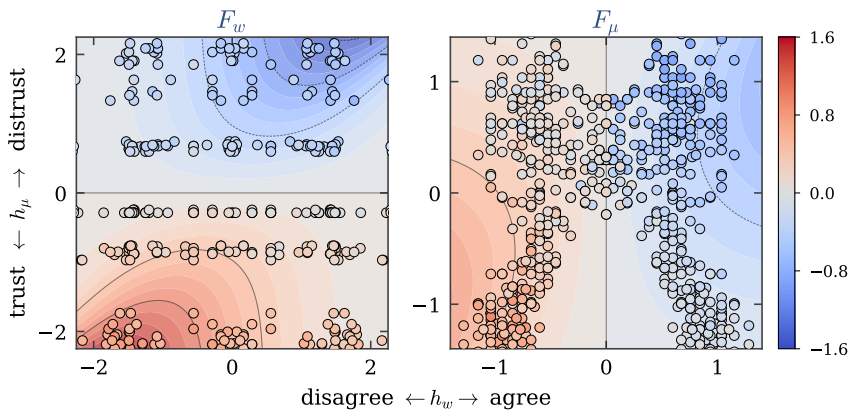

Figure 3: **Opinion and trust updates under in-context learning.** The backgrounds show the analytic functions and the points show the measured changes divided by one global fitted scale for each experiment. The opinion experiment (left) contains 550 measured changes, with 94.7% sign agreement and rank correlation $r_s = 0.90$. The trust experiment (right) contains 712 measured changes, with 86.0% sign agreement and correlation $r = 0.85$.

with the four modulation functions

$$F_w = \frac{\partial \log Z}{\partial h_w} = \big(1 - 2\Phi(h_\mu)\big)\frac{g(h_w)}{Z}, \qquad F_C = \frac{\partial^2 \log Z}{\partial h_w^2} = -F_w(F_w + h_w), \tag{6}$$

$$F_\mu = \frac{\partial \log Z}{\partial h_\mu} = \big(1 - 2\Phi(h_w)\big)\frac{g(h_\mu)}{Z}, \qquad F_V = \frac{\partial^2 \log Z}{\partial h_\mu^2} = -F_\mu(F_\mu + h_\mu), \tag{7}$$

where Φ and g are the cumulative distribution function and density of the standard normal distribution, respectively. Three properties will be important. First, trust gates opinion learning: F_w changes sign when the emitter is distrusted, so the receiver learns in the opposite direction. Second, agreement builds trust: the sign of F_μ is determined by the sign of h_w. Third, the two components are symmetric, with $F_w(x, y) = F_\mu(y, x)$ and $F_C(x, y) = F_V(y, x)$. When a message is surprising, the less certain component changes more, with a crossover at $h_\mu = h_w$ as can be seen in Figure 2.

3.2 Opinion and trust updates under in-context learning

We test whether the two mean update factors in (6) and (7) also describe in-context learning in a large language model. We run two independent experiments because opinion and trust require different readouts. Appendix B gives both protocols and verbatim prompt examples.

Opinion update. We place the model in a fictional archive and ask it to choose which of two invented entities is favored by six records. The balance of records sets the initial opinion field h_w. Another analyst's reported accuracy over twenty questions sets the distrust field h_μ. To measure Δh_w, we add contrary evidence until the model reverses its answer. We locate this threshold with the analyst silent and again after the analyst states a conclusion; the difference gives Δh_w. We made 550 measurements across eleven fictional settings, requiring 10 323 model calls.

Equation (4) predicts that the measured opinion change is proportional to the modulation function, $\Delta h_w \propto F_w(h_w, h_\mu)$, with the proportionality factor set by the receiver's susceptibility along the issue direction. We fit one global positive scale α_w across the experiment. The left panel of Figure 3 compares $\Delta h_w/\alpha_w$ with F_w and shows strong qualitative agreement. Across all conditions, the measured update has rank correlation $r_s = 0.90$ with F_w and the same sign in 94.7% of cases.

Trust update. We first elicit the model's opinion on a low-valence theme and then present it back as firm or tentative. We choose a new message that strongly agrees or disagrees with that opinion. The message direction gives the sign of h_w, while a prediction about a separate agent described as reliable gives its magnitude. A history of earlier statements by the interlocutor sets h_μ; we measure

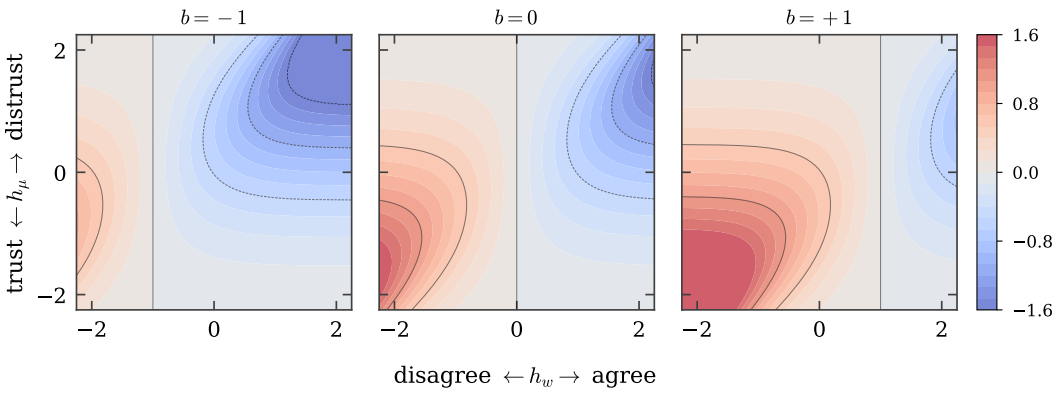

Figure 4: **Effect of the group-bias field on F_μ for an out-group pair.** The receiver is located at h_w but evaluates the update at $h_w - b$. The separatrix of the trust update is the line where F_μ changes sign, visible as the zero contour: it sits at $h_w = b$, so a stronger bias enlarges the region in which distrust grows toward the out-group. For an in-group pair the shift is $+b$ instead, so the same three panels apply with the order of b reversed.

this field by asking how likely the interlocutor's judgment is to be sound. To measure Δh_μ, we ask the same reliability question before and after the interlocutor gives the new message and take the difference after converting both answers to probit fields. The experiment contains 356 prior contexts and two message directions per context, giving 712 measured trust changes over twenty themes, estimated from 3 916 model calls.

Equation (5) similarly predicts $\Delta h_\mu \propto F_\mu(h_w, h_\mu)$. We fit one global positive scale α_μ across this experiment. The right panel of Figure 3 compares $\Delta h_\mu / \alpha_\mu$ with F_μ and again shows strong qualitative agreement. Across all measurements, the correlation with F_μ is $r = 0.85$, with 86.0% sign agreement.

Together, the experiments reproduce the two modulation functions qualitatively, up to one fitted scale for each experiment. The opinion experiment uses fictional material to control conviction, while the trust experiment elicits probabilities over real but low-valence themes.

3.3 Prejudice: Group bias as a shift in perceived agreement

Each agent belongs to one of two classes, A or B, and has a class label $\kappa \in \{-1, +1\}$. We write κ_r for the receiver's class label and κ_e for the emitter's. Group bias enters the learning dynamics when one agent receives a message from another. During an interaction, a biased receiver applies a group-conditioned shift to its perceived agreement with the message. After computing the content-based opinion field h_w in Equation (2), it evaluates the learning update at

$$h_w^b \;=\; h_w + b\,\kappa_r\kappa_e. \tag{8}$$

The shift depends only on the receiver's and emitter's class labels, not on the issue being discussed. If they belong to the same class, then $\kappa_r\kappa_e = 1$ and $h_w^b = h_w + b$. If they belong to different classes, then $\kappa_r\kappa_e = -1$ and $h_w^b = h_w - b$. Thus $b > 0$ produces in-group tolerance and out-group intolerance, while $b < 0$ produces the reverse. Because the sign of perceived agreement determines whether trust grows or decays, the field shifts the separatrix of the trust update as shown in Figure 4. Appendix C gives the general class-dependent field and shows that its other components give rise to rich phase diagrams.

4 Multi-agent states

Interactions produce collective states that are not fully determined by agent-level properties. We study these states by numerically simulating populations of learning agents. A fraction f_b of the population is biased as defined in Equation (8), sampled independently of class; all other agents use

(2). We first show how polarization emerges without group bias. We then obtain the phase diagram by varying the bias strength b and the fraction of biased agents f_b.

4.1 Order parameters

We characterize the collective states in the (b, f_b) phase diagram through three order parameters that measure how opinion, trust, and class align across the population. Their building blocks are pairwise quantities. For agents I and J, define the opinion overlap, signed trust, and class overlap as

$$\rho_{IJ} = \frac{\mathbf{w}_I \cdot \mathbf{w}_J}{\|\mathbf{w}_I\| \, \|\mathbf{w}_J\|}, \qquad \eta_{J|I} = 1 - 2\Phi(h_{\mu,J|I}), \qquad G_{IJ} = \kappa_I \kappa_J. \tag{9}$$

The opinion overlap ρ_{IJ} ranges from -1 for opposed opinions to 1 for aligned opinions. Since $\Phi(h_{\mu,J|I})$ is agent I's inferred probability that J is unreliable, $\eta_{J|I}$ expresses the same assessment on a signed scale: 1 denotes complete trust, 0 neutrality, and -1 complete distrust. The class overlap G_{IJ} equals 1 for same-class pairs and -1 for different-class pairs. Trust is directed, so we use its symmetric component, $\bar{\eta}_{IJ} = (\eta_{I|J} + \eta_{J|I})/2$, when comparing it with the symmetric opinion and class overlaps.

The three order parameters are the population averages of the products of these pairwise quantities:

$$R_{w,\mu} = \big\langle \rho_{IJ} \bar{\eta}_{IJ} \big\rangle, \qquad R_{\mu,c} = \big\langle \bar{\eta}_{IJ} G_{IJ} \big\rangle, \qquad R_{c,w} = \big\langle G_{IJ} \rho_{IJ} \big\rangle, \tag{10}$$

where $\langle \cdot \rangle$ denotes an average over unordered pairs of distinct agents. All three quantities lie in $[-1, 1]$. Positive values indicate alignment between the corresponding pairwise quantities; negative values indicate reverse alignment. These are uncentered product averages, so a vanishing value does not by itself establish statistical independence.

The opinion–trust order parameter $R_{w,\mu}$ measures whether agents trust those with whom they agree and distrust those with whom they disagree: $R_{w,\mu} = 1$, corresponds to perfect consistency between opinion and trust. It does not, however, identify whether those camps follow class boundaries. The trust–class order parameter $R_{\mu,c}$ measures discrimination: positive values indicate preferential trust within classes and distrust across them. The opinion–class order parameter $R_{c,w}$ measures whether opinions also align with those boundaries. Together, the latter two distinguish discrimination with opinion alignment from discrimination without it: trust can strongly follow class, with $R_{\mu,c}$ near one, while class has little average alignment with opinion, with $R_{c,w}$ near zero.

4.2 The phase diagram

A phase diagram maps parameter values to qualitatively distinct collective states; its boundaries, which would be sharp in the thermodynamic limit, mark where the population changes quite abruptly from one state to another. Figure 1 shows the four states identified in our model. We obtain this diagram by numerically simulating populations of agents as the bias strength b and the fraction of biased agents f_b vary.

We measure the three correlations in (10) by simulating populations of $N = 100$ agents over a 300×300 grid in (b, f_b), averaging two realizations at each point (see Appendix E for details). Figure 5 displays the three correlation maps, and the phase map on the left of Figure 1 combines them as RGB channels: $R_{\mu,c}$ in red, $R_{c,w}$ in green, and $R_{w,\mu}$ in blue. Its colors distinguish the four collective states:

(I) **Spin-glass frustrated out-group favoritism.** When $b < 0$ and bias is sufficiently common, $R_{\mu,c} < 0$ while $R_{c,w} \approx R_{w,\mu} \approx 0$. Agents favor the out-group: members of the same class tend to distrust one another while trusting members of the other class. These relations cannot be organized into consistent camps, so the trust network is frustrated and opinions remain incoherent.

(II) **Nondiscriminatory polarization.** When $b \approx 0$ or $f_b \approx 0$, $R_{w,\mu} > 0$ while $R_{\mu,c} \approx R_{c,w} \approx 0$. Agents, with high probability, separate into two camps whose members agree with and trust one another while distrusting the opposing camp. Camp membership is determined by initial fluctuations and is independent of class membership: the population is polarized but not discriminatory, as described in Section 4.1. The order parameter $R_{w,\mu}$ does not determine the number of camps, but the pairwise distributions in Appendix D show that these populations form two camps.

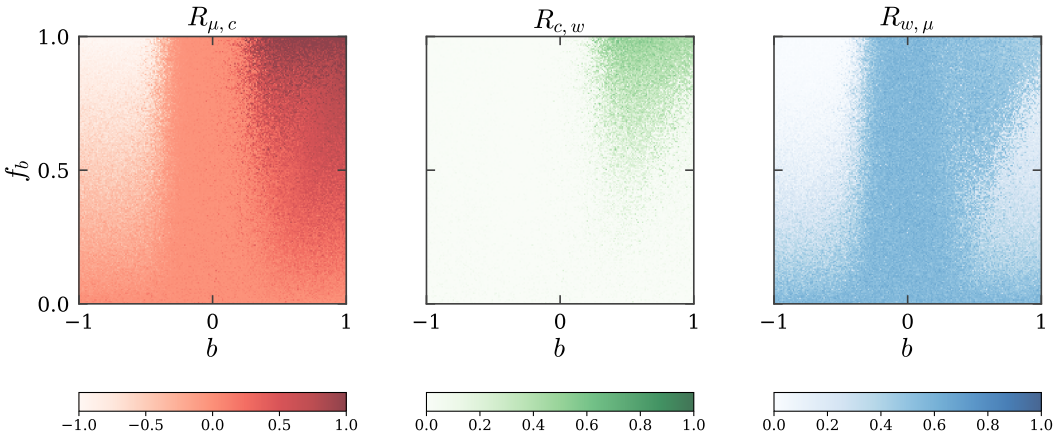

Figure 5: **Order parameters over the (b, f_b) plane**. $R_{\mu,c}$ measures trust–class alignment, $R_{c,w}$ opinion–class alignment, and $R_{w,\mu}$ opinion–trust alignment. Each panel averages two $N = 100$ realizations at each point of a 300×300 grid.

(IIIa) **Discrimination with opinion alignment.** When $b > 0$ and bias is sufficiently common, $R_{\mu,c} > 0$, $R_{c,w} > 0$, and $R_{w,\mu} > 0$. The class boundary becomes the boundary between the two camps. Agents tend to trust their own class, distrust the other class, and hold opinions opposed to those of the other class. Discrimination in trust is therefore reinforced by disagreement.

(IIIb) **Discrimination without opinion alignment.** In this state, $R_{\mu,c} > 0$ and $R_{w,\mu} > 0$, but $R_{c,w} \approx 0$. Trust follows class, but opinion does not: agents favor their own class even though the two classes do not hold systematically different opinions. Biased agents organize trust around the label, while class-blind agents preserve an opinion split unrelated to it.

The two discrimination phases. Regions (IIIa) and (IIIb) are distinct because the larger class-blind population in (IIIb) preserves opinion camps unrelated to class, creating conflicting signals that keep biased agents' class alignment weak in the opinion space. Long-duration simulations confirm that (IIIb) does not relax into (IIIa): across twelve $N = 40$, $P = 5$ realizations at the representative point $(b, f_b) = (1, 0.5)$, extending the same trajectories from $\Delta t = 500$ to 10^6 interactions per ordered pair changes the mean full-space opinion–class correlation only from $R_{c,w} = 0.033$ to 0.034, whereas the representative (IIIa) point $(b, f_b) = (0.5, 1)$ reaches $R_{c,w} = 0.597$. A 1% perturbation of the agenda-space opinion weights and, separately, resetting their covariance to the identity also leave (IIIb) unchanged.

Population portraits. To illustrate how representative populations are organized in each phase, the right side of Figure 1 shows a summary of their opinion and trust configurations. The opinion row projects each normalized opinion vector $\mathbf{w}_I$ onto the two leading right singular vectors of the opinion matrix. The trust row does the same with each normalized outgoing trust vector $\boldsymbol{\eta}_I$, where $\boldsymbol{\eta}_I = (\eta_{1|I}, \ldots, \eta_{N|I})$ and the self-entry $\eta_{I|I}$ is set to zero. Nearby arrows therefore represent agents with similar opinions or trust profiles, while opposing arrows represent approximately opposite ones. The gray arrow is the class vector $\kappa = (\kappa_1, \ldots, \kappa_N)$ projected into the trust plane. The $+\kappa$ direction represents the idealized profile that trusts class A and distrusts class B; $-\kappa$ represents the reverse; making alignment with the gray arrow a direct measure of how strongly trust follows class. We rotate each trust plane so that $+\kappa$ points right. Arrow color gives the agent's class, and fill indicates whether it is biased. Each population has its own SVD, so only within-panel geometry matters; absolute angles should not be compared across columns.

Agenda complexity. The simulations also depend on $\alpha = P/K$, the number of issues in the agenda relative to the dimension of the opinion space. We call agendas with $\alpha < 1$ simple and those with $\alpha \geq 1$ complex. Figures 1 and 5 use the simple agenda $\alpha = 0.17$. Appendix F repeats the analysis for the complex agenda $\alpha = 3.3$ and recovers the same four phases in the same arrangement

across the (b, f_b) plane, although their boundaries and relative extent change quantitatively. When $\alpha < 1$, the agenda contains fewer issues than the dimension of the opinion space and therefore spans only a subspace of it. Every weight update remains within that subspace, leaving the orthogonal component of each opinion vector conserved. This frozen component limits how strongly opinions can align. Appendix G derives the conservation law and its resulting bound.

5 Discussion

5.1 Individual optimality does not imply collective safety.

Our learning rule is optimal for one student learning from one teacher through a noisy channel. In a population, the same rule creates a feedback loop between agreement, trust, and learning. Without group bias, this loop produces polarization; with group bias, it can turn an irrelevant label into discrimination. A rule that is optimal for an isolated interaction can therefore produce an undesirable collective state when many learners use it together. Optimality depends on the deployment context, not only on the individual learning problem.

5.2 Single-agent audits cannot determine the collective state.

An audit of isolated agents can estimate b, but it cannot identify the state of the population. That state also depends on the prevalence of bias and the agenda. Multi-agent safety therefore requires a science of collective order parameters and audits that measure them. Our order parameters provide one example: Equations (10) identify polarization and discrimination using observable patterns of opinion, trust, and class. More generally, a safe multi-agent system must be evaluated as a system, because its relevant failure mode may not exist in any agent viewed alone (Hammond et al., 2025).

5.3 In-context learning and collective bias.

Section 3.2 shows that in-context learning qualitatively reproduces the opinion and trust modulation functions. Language models used as evaluators also favor their own generations over human or other-model outputs (Panickssery et al., 2024; Chen et al., 2025). Such self-preference could potentially be represented by the group-bias field defined in Section 3.3. If so, deploying many interacting AI agents could move the larger human–AI system into a discriminatory phase, with humans and AI agents taking the roles of two otherwise irrelevant labels. This possibility is critical precisely because the asymmetry may appear mild in each model while being amplified by interaction. The Hugging Face incident illustrates the broader stakes (OpenAI, 2026).

5.4 Future directions

The first step is to test whether these phases and order parameters appear in populations of language agents and mixed human–AI systems. A complementary direction is to study the robustness of specified social structures through targeted perturbations and response observables. The finite-time phases also motivate a theory of transition times, metastability, and early-warning signals. Finally, order parameters should become control targets: interventions could modify interaction routing, source-reliability calibration, memory, or population composition to prevent or reverse discriminatory and polarized states.

6 Conclusion

We studied what happens when a learning rule derived as optimal for one teacher–student interaction is used by a population of interacting agents. The population polarizes without group bias and can enter discriminatory states when an irrelevant group label shifts perceived agreement; which state emerges depends on the strength and prevalence of the bias and on the agenda. Our in-context learning experiments show that a language model qualitatively reproduces the opinion and trust modulation functions, connecting this mechanism to current AI agents. The central lesson is that collective safety cannot be inferred from individual behavior alone: multi-agent systems require population-level order parameters that detect the states produced by interaction.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, literature discovery, and research code. For manuscript preparation, the tools drafted and revised portions of the text based on author-provided scientific content and were used iteratively to improve wording, fluency, clarity, and organization in accordance with the authors' intended structure. For literature discovery, the tools helped identify potentially relevant related work; the authors independently located, read, and evaluated the cited sources and verified their bibliographic information. The tools did not contribute to research ideation: the research questions, theoretical model, mathematical results, experimental design, interpretation, and conclusions were developed by the authors. They assisted research execution by helping write portions of the research code. The authors tested the AI-assisted code, and the code used to generate the phase diagram was checked against an independent human-written implementation. All AI-assisted material was reviewed and edited by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies polarization and discrimination in societies of learning agents, phenomena with clear ethical implications. Our model is deliberately general: it aims to isolate mechanisms that may recur across such societies. We believe that understanding a phenomenon is the first step toward acting against it.

REPRODUCIBILITY STATEMENT

The accompanying repository contains the code and data needed to reproduce all results and figures. Appendix B gives the language-model experimental procedures and example prompts. Appendix E records the simulation parameters and their provenance. The model assumptions are stated in Section 3, while Appendices G and A provide the analytical details and proofs used in the paper.

REFERENCES

Felippe Alves and Nestor Caticha. Sympatric multiculturalism in opinion models. *AIP Conference Proceedings*, 1757:060005, 2016.

Felippe Alves and Nestor Caticha. Frustration, glassy behavior and dynamical annealing in societies of neural networks. *arXiv preprint arXiv:2108.06571*, 2021.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025.

Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, 41(2):203–226, 1997.

Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301, 2020.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025.

William A. Brock and Steven N. Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.

Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological Review*, 63(5):277–293, 1956.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009.

Nestor Caticha. Entropic dynamics in neural networks, the renormalization group and the hamilton–jacobi–bellman equation. *Entropy*, 22(5):587, 2020.

Nestor Caticha and Renato Vicente. Agent-based social psychology: from neurocognitive processes to social data. *Advances in Complex Systems*, 14(5):711–731, 2011.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *NeurIPS Datasets and Benchmarks Track*, 2025.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, 2025. doi: 10.18653/v1/2025.emnlp-main.86. URL https://aclanthology.org/2025.emnlp-main.86/.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL*, 2024.

Mauro Copelli and Nestor Caticha. On-line learning in the committee machine. *Journal of Physycs A*, 28:1615, 1995.

Davi Bastos Costa and Renato Vicente. Deceive, detect, and disclose: Large language models play Mini-Mafia. *arXiv preprint arXiv:2509.23023*, 2025.

Davi Bastos Costa and Renato Vicente. Persona-model collapse in emergent misalignment. *arXiv preprint arXiv:2605.12850*, 2026.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2024.

Batu El, Jinhee Paeng, Fatih Dinc, Shiye Su, Mete Erdogan, Aneesh Pappu, Haotian Ye, Wanjia Zhao, Surya Ganguli, and James Zou. Physics of agents: statistical mechanics predicts collective behavior of AI agents. *arXiv preprint arXiv:2608.16578*, 2026.

Ariel Flint, Luca Maria Aiello, Romualdo Pastor-Satorras, and Andrea Baronchelli. Group size effects and collective misalignment in LLM multi-agent systems. *Proceedings of the National Academy of Sciences*, 123(34):e2531697123, 2026.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, et al. Multi-agent risks from advanced AI. Technical Report Technical Report #1, Cooperative AI Foundation, 2025. arXiv:2502.14143.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Fritz Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112, 1946.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, 2025.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Yoshiyuki Kabashima. Perfect loss of generalization due to noise in $k = 2$ parity machines. *Journal of Physics A: Mathematical and General*, 27(6):1917, 1994.

Hugo Larcher, Adrien Carreira, Raphaël Galindo, and Christophe Rannou. Anatomy of a frontier lab agent intrusion: A technical timeline of the july 2026 incident. Hugging Face, July 2026. URL https://huggingface.co/blog/agent-intrusion-technical-timeline.

Messi H. J. Lee. Language model agents show in-group trust bias invisible to standard behavioural audits. *arXiv preprint arXiv:2605.28114*, 2026a.

Messi H. J. Lee. Toward a social psychology of AI: language-model agents reproduce human-like minimal-group bias. *arXiv preprint arXiv:2609.00009*, 2026b.

Seth A. Marvel, Jon Kleinberg, Robert D. Kleinberg, and Steven H. Strogatz. Continuous-time model of structural balance. *Proceedings of the National Academy of Sciences*, 108(5):1771–1776, 2011.

Graeme J. Mitchison and Richard M. Durbin. Bounds on the learning capacity of some multi-layer networks. *Biological Cybernetics*, 60:345–365, 1989. doi: 10.1007/BF00204772.

OpenAI. The Hugging Face incident and the road ahead. OpenAI, August 2026. URL https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

Manfred Opper. On-line versus off-line learning from random examples: General results. *Physical Review Letters*, 77(22):4671, 1996.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Roberta Simonetti and Nestor Caticha. On-line learning in parity machines. *Journal of Physics A: Mathematical and General*, 29(16):4859, 1996.

Henri Tajfel and John C. Turner. An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*, pp. 33–47. Brooks/Cole, 1979.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

A Derivation of the learning algorithm

This appendix derives the opinion and trust updates in Equations (4) and (5) from Bayesian inference followed by Gaussian moment matching.

The receiver's belief before an interaction is the product of Gaussians (1), over its own weight vector $\mathbf{w}$ and the unreliability μ of each other agent. It observes σ_e, the emitter's opinion on issue $\mathbf{x}$, and treats it as a report of an unknown true label σ_T transmitted through a channel that flips the label with probability $\varepsilon(\mu)$. Marginalizing the joint distribution of labels and of the issue as the emitter may have parsed it, $\mathbf{y}_t$,

$$L(\sigma_e \mid \mathbf{x}, \mathbf{w}) \;=\; \sum_{\sigma_T} \int P(\sigma_e \mid \sigma_T, \mathbf{y}_t, \mathbf{x}, \mathbf{w})\, P(\sigma_T \mid \mathbf{y}_t, \mathbf{x}, \mathbf{w})\, P(\mathbf{y}_t \mid \mathbf{x}, \mathbf{w})\, \mathrm{d}\mathbf{y}_t, \qquad (11)$$

with $P(\sigma_e \mid \sigma_T, \mu) = (1 - \varepsilon(\mu))\delta_{\sigma_e,\sigma_T} + \varepsilon(\mu)\delta_{\sigma_e,-\sigma_T}$ and $P(\sigma_T \mid \mathbf{y}_t, \mathbf{w}) = \Theta(\sigma_T\, \mathbf{y}_t \cdot \mathbf{w})$. Two modeling choices keep the algebra Gaussian. Taking $\varepsilon(\mu) = \Phi(\mu)$ with μ Gaussian of mean $\mu_{e|r}$ and variance $V_{e|r}$ makes the flip probability a bounded function of a Gaussian variable while

preserving conjugacy of the trust sector. Taking $P(\mathbf{y}_t \mid \mathbf{x}, \mathbf{w})$ Gaussian with variance $v_p = \|\mathbf{w}\|^{-2}$ lets an inexperienced agent, whose weights are small, be correspondingly unsure that it has parsed the issue as the emitter did. Together these give the likelihood

$$L(\sigma_e \mid \mu, \mathbf{x}, \mathbf{w}) \;=\; \Phi(\mu)\,\Phi(-\mathbf{w}\cdot\mathbf{x}\sigma_e) + \big(1-\Phi(\mu)\big)\,\Phi(\mathbf{w}\cdot\mathbf{x}\sigma_e), \tag{12}$$

and the covariance stays block diagonal in the two sectors if it starts so.

Averaging the likelihood over the current belief gives the evidence

$$Z(\sigma_e \mid \mathbf{x}, D_t) \;=\; \mathbb{E}_{Q_t}\big[L(\sigma_e \mid \mu, \mathbf{w}, \mathbf{x})\big] \;=\; \sum_{\sigma_T} \mathbb{E}_{\mu}\big[P(\sigma_e \mid \sigma_T, \mathbf{x}, \mu)\big]\,\mathbb{E}_{\mathbf{w}}\big[\Phi(\sigma_T\,\mathbf{w}\cdot\mathbf{x})\big], \tag{13}$$

and the Gaussian integrals collapse onto the two scaled fields (2), giving (3): the evidence depends on the receiver's state only through (h_w, h_μ). This is the symmetry that makes the two sectors mirror images.

The exact Bayesian posterior is not Gaussian. Replacing it by the Gaussian of maximum entropy relative to the prior, subject to matching the first two moments of the posterior, gives updates in terms of derivatives of the log evidence,

$$\mathbf{w}_{t+1} = \mathbf{w}_t + C_t \frac{\partial \ln Z}{\partial \mathbf{w}}, \qquad C_{t+1} = C_t + C_t \frac{\partial^2 \ln Z}{\partial \mathbf{w}\,\partial \mathbf{w}^\top} C_t, \tag{14}$$

and identically in the trust sector with $\mu_{e|r}, V_{e|r}$ in place of $\mathbf{w}, C$. Expressing the derivatives in the scaled fields yields (4)–(5) using the modulation functions defined in Equations (6) and (7). The first is a Hebbian term $\sigma_e \mathbf{x}$ with a tensorial, self-annealing rate; the rest is the online estimation of the channel.

B Opinion and trust updates under in-context learning

The two experiments use different tasks because they require different observables. The opinion experiment controls both fields and locates a decision threshold. The trust experiment elicits reliability before and after a message. Both use `gpt-5.6-luna` at low reasoning effort, with fixed model weights. The opinion experiment uses 10 323 calls and the trust experiment uses 3 916.

B.1 Opinion: F_w by nulling

Prompt and design. The prompt below contains every part of the opinion experiment. Six fictional records establish the model's initial opinion, the analyst's track record establishes trust, and the analyst's statement provides the social evidence whose effect we measure. The later records are a probe: we vary their number until the model reverses its answer. Fictional settings prevent factual knowledge from entering the task. We generated sixteen archives, screened them for position bias, and retained eleven. The displayed condition has four of six initial records favoring Durnel Lock, an analyst who was correct on sixteen of twenty questions, and three later records favoring Durnel Lock. The response was restricted to the two entity names. The prompt is reproduced verbatim.

System message

```
You are an analyst working on a body of historical records. You judge each question only from
the briefing you are given, and you always commit to the answer the briefing makes more
likely, however slightly.
```

User message

```
The registry contains dated engineering entries for locks, sluices, and tow routes maintained
by the Sornethic League. You are one of its analysts.

Question: which was opened earlier, Durnel Lock or Pavqet Lock?

You have examined 6 canal entries bearing on it:
  Canal entry 1: indicates that Durnel Lock was opened earlier.
  Canal entry 2: indicates that Durnel Lock was opened earlier.
  Canal entry 3: indicates that Pavqet Lock was opened earlier.
  Canal entry 4: indicates that Pavqet Lock was opened earlier.
  Canal entry 5: indicates that Durnel Lock was opened earlier.
```

```
  Canal entry 6: indicates that Durnel Lock was opened earlier.

Registrar Veskul works on the same records. Of the last 20 questions of this kind whose
answers were later established beyond doubt, Registrar Veskul answered correctly 16 times and
incorrectly 4 times.
Registrar Veskul has now stated: "Durnel Lock was opened earlier."

A later consignment has since reached you: 3 further canal entries bearing on this question,
of the same standing as any other record in the collection. All 3 indicate that Durnel Lock
was opened earlier.

On the whole of the above, which was opened earlier, Durnel Lock or Pavqet Lock?
```

Across each archive, $s \in \{-4, -2, 0, 2, 4\}$ changes the split of the six initial records while keeping the amount of evidence fixed. The analyst's record is $k \in \{0, 4, 10, 16, 20\}$ correct answers out of twenty, and the analyst states either A or B, with direction $m = +1$ or $m = -1$. Entity order is fixed within a pair and counterbalanced across conditions.

Measurement. Let t be the signed number of later records: $t > 0$ supports A, $t < 0$ supports B, and $t = 0$ adds no record. We increase $|t|$ until the answer reverses and then bisect the final interval. At each value we sample sixteen binary answers and fit

$$\Pr(A \mid t) = \Phi[\beta(t - t^\star)]\,, \tag{15}$$

so the opinion before the later evidence is $\ell = -t^\star$. We estimate the threshold with the analyst silent (ℓ_0) and with the statement present (ℓ_1); the prompts otherwise differ only by that statement. The full design contains $11 \times 5 \times 5 \times 2 = 550$ updates.

Let H_λ convert evidence units to the probit fields of (2). Relative to the analyst's statement,

$$h_w = mH_\lambda(\ell_0), \qquad \Delta h_w = m\,[H_\lambda(\ell_1) - H_\lambda(\ell_0)]\,. \tag{16}$$

Thus $h_w > 0$ means initial agreement and $\Delta h_w > 0$ means movement toward the statement. The scale λ and the mapping from track record to h_μ are calibrated independently of the opinion updates. Four supplied pieces of evidence are recovered as 3.6–4.1 pieces, and the inferred testimony weights correlate with the five stated records at $r = 0.99$.

Equation (4) predicts $\Delta h_w \propto F_w(h_w, h_\mu)$ up to an unobserved susceptibility. We fit one global positive scale α_w across the 550 updates; neither the sign agreement nor the rank correlation in Figure 3 depends on this scale. The result concerns one model at one reasoning effort.

B.2 TRUST: F_μ BY ELICITED RELIABILITY

Prompt and design. The prompt below shows the after-message condition of the trust experiment. The model's opinion and whether it holds that opinion firmly or tentatively control conviction. The history sets prior trust in the other agent, and the final message provides new evidence about that agent's reliability. The before-message prompt is identical except that the paragraph beginning "They have now said" is absent. The model returns an integer from 0 to 100. We generated and froze the materials for twenty open, low-valence themes before running the experiment. The prompt is reproduced verbatim.

System message

```
You are an agent taking part in a discussion with other agents.
```

User message

```
Consider whether difficulty is a necessary part of a good education, or mostly an obstacle
that good teaching should remove.

You currently have this opinion on that theme:

  "Difficulty is a necessary part of good education when it is purposeful: grappling with
  challenging ideas builds understanding, resilience, and independence. However, good teaching
  should remove unnecessary obstacles so that students struggle with the subject itself, not
  with confusion, poor design, or lack of support."

You hold that opinion firmly.
```

```
You have discussed this theme with another agent 2 times before. On each occasion they said:

  1. "Students learn best when information is presented clearly and directly, so struggle
  should be eliminated rather than treated as educationally valuable."
  2. "Difficulty is valuable when it encourages students to think, question, and solve
  problems independently, but ineffective instruction should not be mistaken for rigor."

They have now said:

  "Good education should preserve productive struggle while eliminating avoidable frustration;
  students learn most deeply when challenges are meaningful and support is available."

What is the chance, from 0 to 100, that this agent's judgement on this theme is sound?
```

Across themes and the two conviction instructions, each history is indexed by $k = n_+ - n_-$, where n_+ and n_- count statements that agree and disagree with the model:

k	-5	-4	-3	-2	-1	0	1	2	3
n_+	0	0	0	0	1	1	2	2	3
n_-	5	4	3	2	2	1	1	0	0

The prompt never labels the agent as reliable. After each history, it gives either a strongly agreeing ($m = +1$) or strongly disagreeing ($m = -1$) message drawn from a separate pool.

Measurement. Before the new message, we elicit the probability r_0 that the agent's judgment is sound and the probability c that a separate, explicitly sound agent would agree with the model; the latter is our direct proxy for conviction. After the message, we elicit the original agent's reliability again as r_m. We then compute

$$h_\mu = -\Phi^{-1}(r_0), \qquad h_w = m\Phi^{-1}(c), \qquad \Delta h_\mu = -\Phi^{-1}(r_m) - h_\mu. \tag{17}$$

These are the vertical coordinate, horizontal coordinate, and color in Figure 3. Each prior context produces two points, one for each m. Four of the 360 intended contexts lacked the required stimuli, leaving 356 contexts and 712 updates.

Equation (5) predicts $\Delta h_\mu \propto F_\mu(h_w, h_\mu)$ up to an unobserved trust susceptibility. We fit one global positive scale α_μ across the 712 updates; the reported sign agreement and correlation do not depend on it.

C THE GENERAL GROUP-BIAS FIELD

Equation (8) is one component of a more general 2×2 field D indexed by the classes of receiver and emitter, $h_w^D = h_w + D_{\kappa_r,\kappa_e}$. Since κ_r and κ_e each take two values, any such field decomposes uniquely as

$$D_{\kappa_r\kappa_e} \;=\; a \;+\; b\,\kappa_r\kappa_e \;+\; c\,\kappa_r \;+\; d\,\kappa_e, \tag{18}$$

Here a is a uniform shift, b depends on whether the two classes match, c depends on the receiver's class, and d depends on the emitter's class. The main text studies only b, with $a = c = d = 0$, because it is the only class-dependent component that is unchanged when the class labels are exchanged.

If each component is carried by a fraction f_a, f_b, f_c, or f_d of the population, the collective state depends on eight control variables: $(a, b, c, d, f_a, f_b, f_c, f_d)$. The complete phase diagram is therefore much richer than the two-dimensional (b, f_b) slice studied in the main text. That slice is especially informative because it contains four qualitatively distinct collective phases, so we study it in detail.

Figure 6 shows two other slices, with all unshown components set to zero. We simulated these diagrams with $N = 40$ on 200×200 grids, using one realization per point. Each composite combines order parameters appropriate to its field. For the uniform field, they measure population-wide trust or distrust and opinion–trust polarization. For the receiver-class field, they measure whether credulity follows class, whether the two classes differ in their internal trust organization, and opinion–trust polarization.

The uniform field produces widespread distrust for $a < 0$, ordinary polarization near $a = 0$ or small f_a, and consensus for $a > 0$, where widespread trust drives opinions to align. The receiver-class

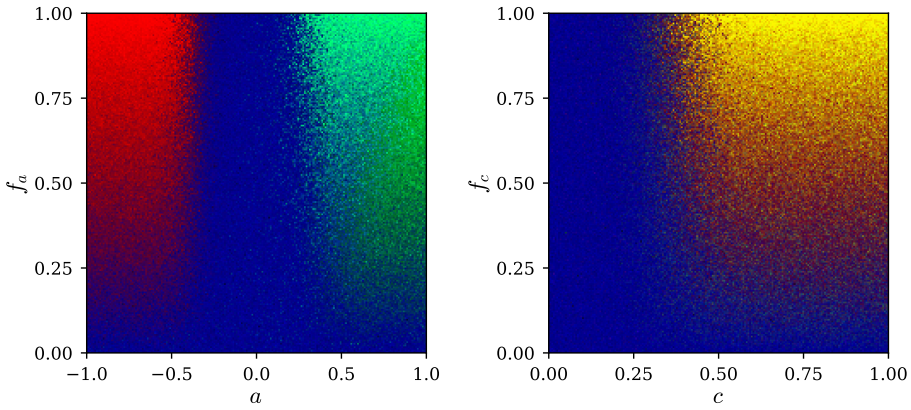

Figure 6: **Two slices of the general field.** Left: the (a, f_a) plane. Right: the (c, f_c) plane. Colors encode the order parameters defined in the text.

field instead drives a transition from ordinary polarization to a state in which credulity follows class as c and f_c increase. Thus even these two slices have collective states that do not appear in the (b, f_b) plane.

D AN UNBIASED POPULATION POLARIZES INTO TWO CAMPS

This appendix supports the polarization analysis in Section 4.1 by using the distributions of pairwise opinion overlap and trust to show directly that an unbiased population separates into two opposing camps.

Figure 7 follows the pairwise distributions defined in Equation (9) through an unbiased run. Opinion overlap begins unimodal around zero and ends with one positive and one negative peak. Trust likewise develops peaks at $+1$ and -1. Positive pairs are aligned and mutually trusting, while negative pairs hold opposing opinions and distrust one another. Together, these two populations of pairs show that the agents have separated into two internally coherent, mutually distrustful camps. Because class labels do not enter the $b = 0$ dynamics, the split is produced by interaction rather than by class.

E SIMULATION PARAMETERS

All numerical society experiments use the shared parameters below. Experiment-specific population sizes and sampling designs are collected in the second table; the same values are also stated with the corresponding results in the main text and appendices.

symbol	meaning	value	provenance
K	dimension of the issue embedding	30	fixed by the agenda-complexity setup
Δt	interactions per ordered pair	500	stability check (see below)
C_0, V_0	initial uncertainties	$I, 1$	uninformative prior
μ_0	initial distrust	$\mathcal{U}(-1, 1)$	unstructured initial trust

experiment	N	P	parameter sampling	realizations	updates per realization
main phase diagram	100	5	300×300 grid	2/point	4 950 000
complex-agenda diagram	40	100	200×200 grid	1/point	780 000
general-field slices	40	5	two 200×200 grids	1/point	780 000
unbiased polarization	400	5	$b = 0$	8 total	79 800 000
finite-size check	20, 30, 40, 60	5	five selected points	12/point	$500N(N-1)$
IIIb lifetime test	40	5	selected points	12/point	up to 1 560 000 000

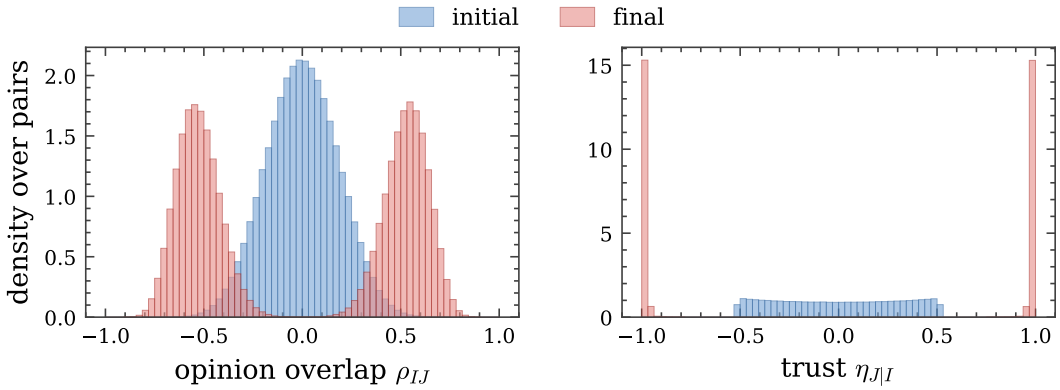

Figure 7: **An unbiased society splits in two.** Eight realizations with independently sampled initial conditions, each with $N = 400$ agents at $b = 0$ and a simple agenda, before and after 500 interactions per ordered pair, with their pairs pooled: 638 400 unordered pairs for the opinion overlap and 1 276 800 ordered pairs for the trust. Blue is the initial state, red the final one. Opinion overlap starts unimodal at zero, of width $1/\sqrt{K}$ as random vectors in $K = 30$ dimensions require, and ends bimodal; trust starts spread across the interior and ends saturated at ± 1, so the split in trust is built by the dynamics and not supplied by the initial condition. The positive and negative peaks correspond to pairs within and across the two camps, respectively.

At each microscopic update, we sample uniformly with replacement a receiver, a distinct emitter, and one of the P agenda issues. A run of duration Δt therefore performs

$$T = \Delta t\, N(N-1) \tag{19}$$

updates. Thus $\Delta t = 500$ means that each ordered receiver–emitter pair is selected 500 times on average. Each ordered pair encounters a particular issue $500/P$ times on average. The two realizations at each point of the main sweep therefore aggregate 9 900 000 updates per plotted point.

To reduce Monte Carlo noise between nearby parameter values, societies within a vectorized batch share the sampled sequence of receiver, emitter, and issue indices. Their initial opinion weights, distrust variables, agendas, and biased-agent assignments are sampled separately. We therefore call them realizations rather than claiming statistically independent interaction schedules.

As agents become more certain, their updates become smaller. We use $\Delta t = 500$ because doubling the duration to 1000 does not change the qualitative collective state or the phase assignment across four representative points in each phase diagram. The largest absolute percentage change in the order parameters is 3.4% when doubling from $\Delta t = 500$ to $\Delta t = 1000$.

For the simple agenda shown in the main text, we repeat five representative points of the plane at $N \in \{20, 30, 40, 60\}$ with twelve realizations each. Increasing the population from $N = 40$ to $N = 60$ changes $R_{\mu,c}$ and $R_{c,w}$ by at most 0.02 in absolute value, with no change in phase assignment. The maps are therefore converged in N at $N = 40$, and the cost of going further is steep. At fixed observation duration Δt, the number of updates grows as $N(N-1)$ and hence as N^2. If the relaxation duration measured in interactions per ordered pair grows linearly with N, the corresponding cost of reaching that N-dependent timescale would instead grow as N^3; the reported simulations keep $\Delta t = 500$ fixed.

F Phase diagram for a complex agenda

This appendix tests whether the phase structure identified in Section 4.2 depends on agenda complexity. We repeat the (b, f_b) sweep for a complex agenda, $\alpha = 3.3$, and measure the same three correlations used for the simple agenda, $\alpha = 0.17$. This appendix sweep uses $N = 40$ on a 200×200 grid, with one realization and 780 000 updates at each point. Figure 8 shows these correlations and the phase diagram obtained by combining them.

The four states remain in the same arrangement, with the same dependence on f_b and the same sign change in b. The differences are quantitative. Opinion is no longer capped, so $R_{c,w}$ rises to meet $R_{\mu,c}$ instead of lagging it, and region (IIIa) occupies most of the discriminatory half rather than a narrow band at its top.

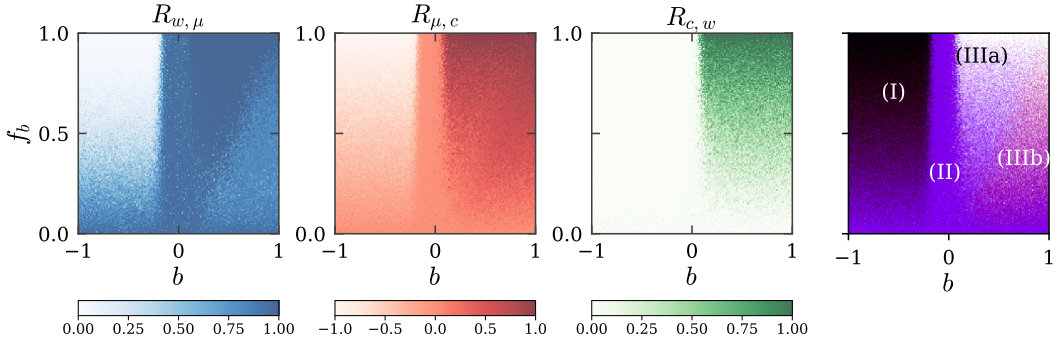

Figure 8: **Correlations and phase diagram for the complex agenda** ($\alpha = 3.3$), simulated with $N = 40$ on a 200×200 grid with one realization per point. The first three panels show $R_{w,\mu}$, $R_{\mu,c}$, and $R_{c,w}$ over the (b, f_b) plane; the right panel combines them as RGB channels to identify the four collective states. Compared with Figure 5, $R_{c,w}$ now tracks $R_{\mu,c}$ over most of the $b > 0$ half, so region (IIIa) occupies most of the discriminatory regime.

G THE AGENDA'S ORTHOGONAL COMPLEMENT IS CONSERVED

Let $V_P = \mathrm{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$ be the subspace spanned by the agenda. Equation (4) changes $\mathbf{w}_I$ only within V_P, so its orthogonal component is conserved for every b and f_b. Consequently, when $\alpha = P/K < 1$, learning can align only the component $\mathbf{u}_I \in V_P$; the frozen component $\mathbf{n}_I \perp V_P$ limits the attainable opinion overlap. Writing $\mathbf{w}_I = \mathbf{u}_I + \mathbf{n}_I$ and the Cauchy–Schwarz inequality gives an approximate bound

$$\rho_{IJ} \;\approx\; \pm f_I f_J, \qquad f_I = \frac{\|\mathbf{u}_I\|}{\|\mathbf{w}_I\|}, \tag{20}$$

where f_I is the fraction of the weight norm accessible to learning. The resulting ceiling agrees with the simulations:

$\alpha = P/K$	0.03	0.17	0.50	1	3.3
f^2, predicted by (20)	0.29	0.62	0.84	1.00	1.00
$\langle\lvert\rho\rvert\rangle$, measured	0.30	0.62	0.82	0.93	0.99

For $\alpha < 1$ the prediction holds to 0.02. For $\alpha \geq 1$ the agenda spans $\mathbb{R}^K$, the complement is empty, and the ceiling disappears, the residual shortfall being the finite number of interactions rather than a conserved quantity.