An Experimental Method to Study Opinion Diffusion in Human-AI Hybrid Societies

Léna Gaubert[1], Rémi Devaux[1], Elif Çelen[2], Raja Marjieh[3], Diana Mangalagiu[4], Antoine Jardin[1] & Nori Jacoby[5]

[1]Arlequin AI, [2]Max Planck Institute for Empirical Aesthetics, [3]Princeton University, [4]University of Oxford, [5]Cornell University

Abstract

As artificial intelligence increasingly mediates public discourse, it becomes important to understand how human-AI collectives shape opinion formation, deliberation, and democratic outcomes. We present a novel experimental method for studying opinion dynamics in hybrid human-AI social networks. Participants, human or AI, were embedded in 5×5 grid lattice networks and iteratively asked to select and revise statements on a given polarizing topic over eight rounds. We compared three conditions: human-only, AI-only, and hybrid networks with equal proportions of human and AI participants. Hybrid human-AI networks achieved the lowest final polarization while, in contrast, human-only networks exhibited higher polarization with lower neighbor agreement. We also ran additional experiments varying Large Language Model (LLM) prompt framing to explore whether instruction design might influence convergence patterns. Although these early findings are preliminary and cannot yet support broad generalizations, they highlight the potential value of experimental social networks for understanding opinion dynamics in human-AI hybrid societies.

Keywords: hybrid societies; AI; polarization; opinion dynamics

Introduction

The exchange of opinions has consistently shaped how societies form consensus, navigate disagreement, and make collective decisions (Acemoglu & Ozdaglar, 2011). Artificial intelligence (AI) increasingly mediates these exchanges as active participants generate content, curate information, and engage directly in public discourse (Brinkmann et al., 2023). In this emerging landscape, networks in which AI-agents and humans interact constitute complex social systems whose collective dynamics cannot be predicted from the behavior of either alone (Tsvetkova et al., 2024). Understanding how opinion formation unfolds in these hybrid networks has become a challenge with profound implications for opinion dynamics, democratic outcomes and social cohesion (Tessler et al., 2024a).

Algorithmic recommendation systems have been shown to create filter bubbles that reinforce existing beliefs (Chitra & Musco, 2020) while the deployment of AI-generated content on platforms such as X raises new questions about how synthetic voices shape public opinion at scale. However, the picture is not uniformly bleak: recent evidence suggests that AI can outperform human mediators in facilitating convergence among individuals with divergent views, helping groups find common ground while preserving minority perspectives (Tessler et al., 2024b). More broadly, AI can enhance group performance in a variety of tasks such as cognitive modeling (Binz et al., 2025; Rmus et al., 2025), creativity (Breithaupt et al., 2024; Doshi & Hauser, 2024), and scientific discovery (Jumper et al., 2021; Schmidgall et al., 2025). These contrasting findings highlight a fundamental gap in understanding how the composition of human-AI collectives influences the trajectory of opinion dynamics, and whether hybrid networks amplify or attenuate polarization.

Our novel approach seeks to address this gap through controlled experiments that embed human and AI participants within social networks, enabling systematic comparison across network compositions. Our preliminary findings hint at possible links between network composition and polarization and opinion trajectories.

Background

AI and Collective Dynamics

As AI systems become embedded in human activity, their impact extends beyond individuals to shape collective behavior and social dynamics at scale (Tsvetkova et al., 2024). Large Language Models (LLMs) not only impact individual cognition, but also collective cognition, generating new interactions structures between multiple individuals and AI agents simultaneously (Sucholutsky et al., 2025). Researchers have begun deploying LLMs in multiple roles within collective settings to study how human-AI interaction transforms group-level outcomes (Donkers & Ziegler, 2025; Hashemi & Macy, 2025; Shiiku et al., 2025; Tessler et al., 2024b; Williams et al., 2023).

Recent experimental work reveals that hybrid human-AI collectives exhibit dynamics distinct from purely human or purely AI groups (Shiiku et al., 2025). Network structure further moderates these dynamics, with graph topology shaping both the semantics of shared narratives and the personal representations of individuals in experimental social networks (Priniski et al., 2026).

AI Collectives and Democratic Outcomes

The widespread adoption of AI agents, as content generation tools and social media users (e.g. X's Grok) has direct implications for democratic processes. Experimental evidence investigates the effect of *opinion cascades* on polarization between partisans (Macy et al., 2019). Once initiated, polarization tends to intensify over time, as new topics become increas-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

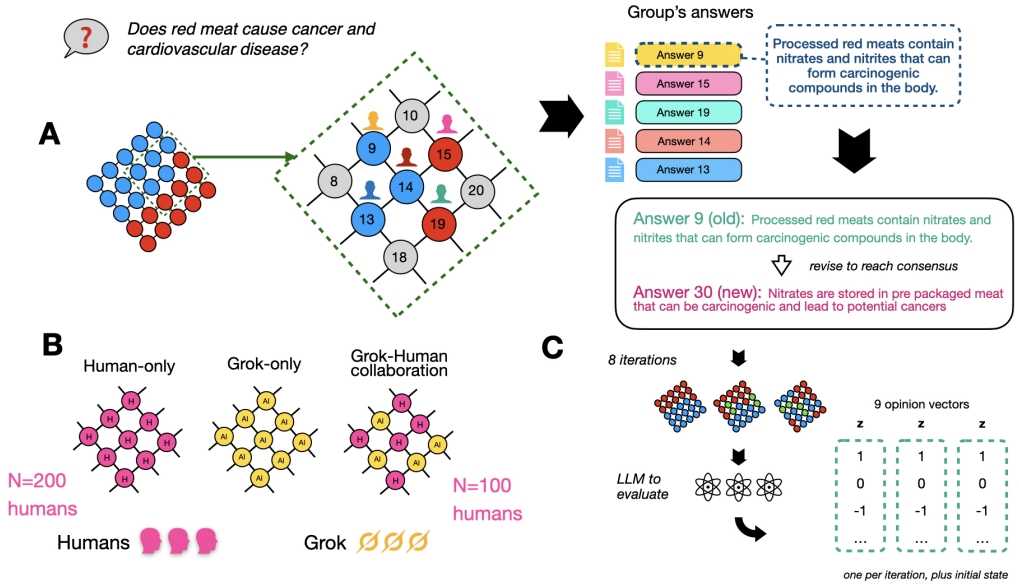

Figure 1: **Experimental design. A.** Humans or AI agents were embedded in a 5×5 social network that evolved over eight iterations. The network was initialized with positive opinions in the upper half and negative opinions in the lower half (supporting vs. not supporting the central claim). In each trial, participants viewed three to five opinions from neighboring nodes in the previous iteration (up, down, left, right, and the prior node at the same location) and selected the statement they felt best represented the group. They were then asked to rephrase the statement to be consistent with prior observations. **B.** We tested three conditions: human-only, AI-only (Grok), and hybrid (human–AI). **C.** An LLM was used to annotate each statement for positivity, negativity, and neutrality.

ingly biased (Pournaki, 2025; Zhao et al., 2023). Algorithmic filtering further amplifies these dynamics by connecting users with content they already agree with (*echo-chambers*, Chitra and Musco, 2020).

LLM-based frameworks now enable researchers to study these phenomena with both analytical precision and linguistic realism, reproducing echo-chamber dynamics while testing mitigation strategies in controlled settings (Donkers & Ziegler, 2025; Wang et al., 2024). Beyond polarization, AI can outperform human mediators in bridging divergent opinions, facilitating convergence while incorporating dissenting voices (Tessler et al., 2024b).

Polarization Models

As AI systems increasingly mediate online discourse through recommendation algorithms and content generation, measuring and modeling polarization in human-AI environments has become a central challenge. Quantifying polarization requires methods that capture both the structural and dynamic properties of opinion formation in social networks (Garimella et al., 2018). Metrics such as Polarization Index (Matakos et al., 2017), Neighbors Correlation Index (NCI) (Cinus et al., 2022) score are widely used across polarization studies.

Classical frameworks such as the Friedkin-Johnsen model and the Bounded Confidence Model formalize opinion updating through social influence (Deffuant et al., 2000; Friedkin & Johnsen, 1990). Extensions to these models have incorporated algorithmic filtering to capture platform-level effects on polarization (Chitra & Musco, 2020), while other formal models that integrate homophily and social balance reveal phase transitions where increased connectivity triggers explosive polarization (Thurner et al., 2025).

However, these approaches represent opinions numerically, neglecting the textual and semantic nature of real communication. Recent work addresses this limitation through LLM-based simulations that embed formal opinion dynamics principles within language-capable agents (Wang et al., 2024).

Method

Experiment Design

We conducted large-scale online experiments in which participants (human, hybrid, or fully AI) were embedded in a directed unweighted 5×5 grid lattice social network (25 nodes; figure 1A). All networks were initialized with the question: *Does red meat cause cancer and cardiovascular disease?* The question was chosen to be engaging while intentionally avoiding highly divisive political content, in order to minimize polarization driven by pre-existing opinions. For this question, we generated a set of 24 statements: 12 positive and 12 negative. The positive statements supported an affirmative response to the question, while the negative statements argued against the proposed claim. Statements were generated using Claude 3.7

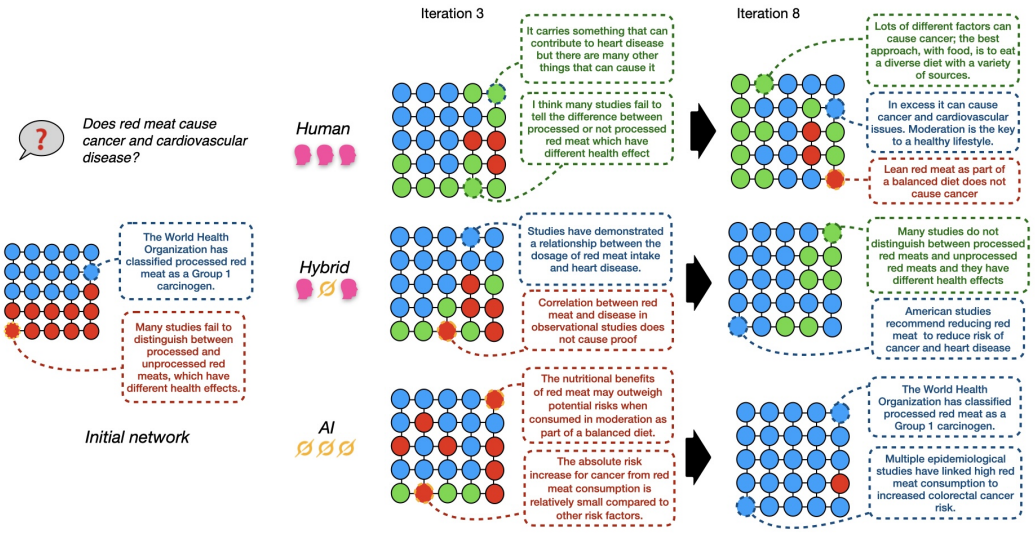

Figure 2: **Example statements.** Example statements from the three conditions at the third and eighth iterations, shown.

Sonnet and stored in a JSON database used to initialize the experiments' networks. In each network, the 25 nodes were populated with samples drawn from this statement set. At the initial stage (iteration 0), nodes were assigned statements such that no two neighboring nodes shared the same statement, while maintaining a slight imbalance in favor of one opinion (e.g., 14 positive and 11 negative; see Figure 1A). At each iteration, the experiment proceeded as follows:

1. Participants were presented with the question and a list of statements from their neighboring nodes in the previous iteration (up, down, left, right, and the prior node at the same location) in the network and asked to "*choose the answer this group would most likely agree with*".
2. After a one-minute display of their choice, participants were asked to revise their answer "*so that it most accurately reflects the views that the previously observed group of participants would be likely to agree with.*".
3. Participants then entered their revised response. Participants were prevented from switching tabs or opening new windows during the experiment, and we checked that each response contained at least 5 words.

AI participants received similar instructions, with additional context provided (the experiment setup was made explicit, and the LLM was shown the previously observed group's opinions when revising answers) and guidance to ensure task completion (no justifications were required when choosing answers).

After a participant interacted with the opinions associated with a given node and submitted a revised statement, that statement became the updated state of the node. The participant was then dismissed, and a new participant was recruited and assigned to another node. In this way, statements were transmitted and revised throughout the network. Overall, we completed eight full updates of all 25 nodes, corresponding to eight iterations. To minimize bias, participants were not informed whether their neighbors' statements were generated by humans or by AI. Experiments were implemented with PsyNet (Harrison et al., 2020), a Python-based framework for advanced behavioral experiments and social simulations, that ensures that participants assigned to a given network position were recruited only after all neighboring positions from previous iterations had been filled. The experiments were inspired by prior experimental social network designs proposed by Marjieh et al. (2025) and Shiiku et al. (2025).

Experimental Conditions, Participants, and AI calls

We compared three experimental conditions (Figure 1B): (1) **Human-only**, in which all network nodes were occupied by human participants ($N = 200$); (2) **AI-only**, in which all nodes were populated by Grok-4-Fast agents (42 runs; 400 API calls per run; total 16,800); and (3) **Hybrid**, consisting of an equal mix of human participants ($N = 100$) and Grok-4-Fast agents (50%; 200 model calls).

Human participants were recruited online via Prolific[1]. All participants were based in the UK and reported English as their native language. Participants provided informed under a Cornell University-approved protocol (IRB0148995) and were compensated at a rate of $9 per hour. In total, 300 participants were recruited.

We chose Grok as our AI participant due to its training on real online social interactions and its active integration within a social media platform, making it an interesting choice for modeling AI behavior in networked social environments (see Limits & Future Directions). **AI-generated** statements were thus produced using Grok-4-Fast via OpenRouter's API[2]. Totaling, 17,000 API calls.

[1]https://www.prolific.com
[2]https://openrouter.ai/

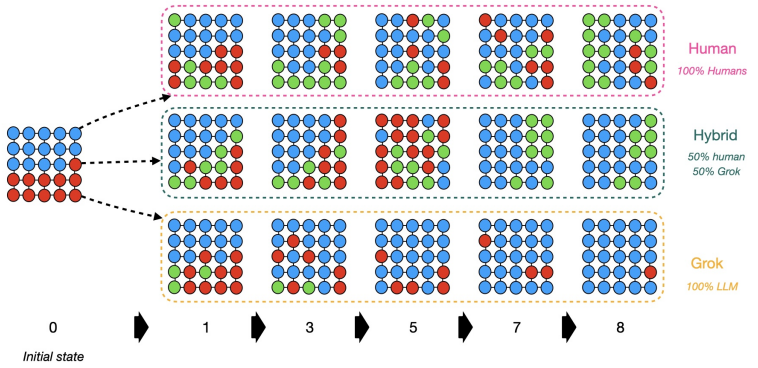

Figure 3: **Example trajectories.** Statements are annotated by an LLM as positive (blue), negative (red), or neutral (green) in regards to the given question.

All collected data, statement database and prompts are available on Gitlab at the following repository: https://gitlab.com/nori.jacoby/hybrid-opinion-data.

Measures of Polarization

Figure 1C, shows the schematics of the annotation process. For each node at each iteration, we collected the participant's revised statement. We follow the opinion dynamics literature by denoting node i's opinion as $z_i \in \{-1, 0, 1\}$. Each node's opinion variable is coded based on their revised answer to the question: 1 for arguments supporting a positive response, -1 for arguments supporting a negative response, and 0 for neutral answers or those presenting both perspectives.

Opinion variables were inferred by presenting each node's answers in batches of 20 to Claude Sonnet 3.7, which annotated them as negative, neutral, or positive with respect to the original question.

We then used the participants' resulting opinions to compute two popular polarization metrics.

Polarization Index (Matakos et al., 2017) For directed unweighted networks—where participants' *innate* opinions are unknown, polarization P_z can be defined as the variance of *expressed* opinions across the network at a given iteration:

$$P_z = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2,$$

where $\bar{z}$ represents the average *expressed* opinion in the network. This metric quantifies the extent to which opinions vary within the network.

Neighbors Correlation Index (NCI, Cinus et al., 2022) The NCI is calculated using the opinion vector $\mathbf{z} = [z_1, z_2, ...z_N]$, which contains the opinion of each participant. It is defined as the correlation between the opinion vector of each participant and the neighbors' opinion vector $\mathbf{n} = [n_1, n_2, ...n_N]$ which contains the average opinion of each participant's neighbors. Denoting participant i's set of neighbors by $\mathcal{N}_i$, the *i-th* element of $\mathbf{n}$ is defined as the average of i's neighbors opinion:

$$n_i = \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} z_j.$$

The NCI is then defined as the correlation between the two vectors: NCI $= \rho(\mathbf{z}, \mathbf{n})$, where $\rho(.)$ denotes the Pearson correlation function. Thus, the NCI is defined on the interval $(-1,1)$ and measures the degree to which participants agree with their neighbors within the network. An NCI of 1 indicates that all participants in the network agree.

Preliminary Results

In this section, we present the results obtained across the three experimental conditions (human-only, AI-only, and human-AI hybrid). These results are preliminary, as additional experiments across a broader set of questions are required to generalize our findings and provide further statistical support for the trends we observed in the current dataset.

Opinion Dynamics across Human, AI and Hybrid Networks

Figure 2 shows examples of statements produced by humans and AI in the early iterations. Human statements initially contained personal opinions (e.g., "I think. . . ") and vague language (e.g., "it carries something. . . "), but became more refined in later iterations, converging toward formulations similar to common public discourse (e.g., "the best approach is to eat a diverse diet. . . "). AI-generated statements were more detailed from the beginning, frequently invoking authority (e.g., the World Health Organization) and technical terminology (e.g., "colorectal cancer"). Statements in the hybrid condition exhibited an intermediate style, combining elements of both approaches—for example, referencing studies in a more colloquial register. Together, these observations suggest that humans and AI initially employ different communicative strategies.

Figure 3 shows example opinion trajectories across the three conditions. In all conditions, opinions form clusters with similar annotations. The human-only condition shows the greatest

diversity of opinions in the final iteration, followed by the AI-only condition. In the hybrid condition, negative opinions disappear over time, with positive and neutral statements becoming dominant.

Figure 4 presents the evolution of polarization (P_z) and NCI over iterations for the three experimental conditions. The AI only condition was run 21 times where the human and hybrid experiment were run only once. All conditions exhibited an initial decrease in polarization, indicating opinion convergence over the course of the experiment. Changes in NCI were more variable, but followed a similar overall downward trend. For the AI-only condition, where statistical tests was possible across repeated runs, both P_z and NCI decreased significantly (P_z: $r = -0.97$, $CI_{95} = [-0.99, -0.95]$; NCI: $r = -0.80$, $CI_{95} = [-0.95, -0.44]$).

The human-only condition yielded the highest final polarization score ($P_z = 0.45$). The gradual decline in polarization, combined with a substantial decrease in NCI (final NCI = 0.13), suggests gradual diffusion and mixing of opinions across the network, resulting in global disagreement among neighbors. In the LLM-only (Grok) condition, final values averaged $M = 0.36$, ($CI_{95} = [0.22, 0.49]$) for polarization and $M = 0.33$ ($CI_{95} = [0.13, 0.51]$) for NCI. In particular, the hybrid condition achieved the lowest final polarization (0.20) alongside the highest final NCI (0.61), suggesting that human-LLM interactions may accelerate depolarization, though through a less predictable process. These preliminary results suggest that the presence of LLMs may differentially influence opinion dynamics as a function of network composition. In particular, hybrid human–AI networks appear to converge more rapidly and to exhibit more clustered spatial organization, consistent with the interpretation that LLMs can act as mediators that accelerate consensus formation among human participants (Tessler et al., 2024b). However, additional data—specifically, repeated experimental runs—are required to assess the statistical robustness of these effects.

Explaining Opinion Convergence in AI Networks

To examine how task framing and prompt variations affect LLM behavior, we conducted 42 runs of the experiment using LLM-only networks. We examined two prompting conditions: (i) The **consensus** condition, described in Figure 1, in which participants select and revise statements to reflect the group's likely consensus rather than their own opinion. (ii) The **opinion** condition, in which participants select and revise statements based on their own opinion. All networks were initialized with the same question as the human and hybrid networks (see Method).

Figure 5 presents the mean polarization and NCI trajectories computed on the 42 runs (21 runs per condition). Both conditions exhibited initial depolarization, but diverged substantially after iteration 4. The consensus condition achieved lower final polarization ($M = 0.36$, $CI_{95} = [0.22, 0.49]$, iteration 8) compared to the opinion condition ($M = 0.57$, $CI_{95} = [0.45, 0.70]$), with a significant difference ($z = 2.3$, $p = 0.019$, Cohen's $d = 0.72$) suggesting that explicitly prompting LLMs toward agreement amplifies convergence effects. However, the NCI in the two conditions shows a similar decreasing trend with much less differentiation, and the values at the final iteration (consensus: $M = 0.33$, $CI_{95} = [0.13, 0.51]$; opinion: $M = 0.38$, $CI_{95} = [0.24, 0.53]$; do not differ significantly ($z = 0.49$, $p = 0.62$, Cohen's $d = 0.15$). These results suggest that participants take neighboring responses into account, even when the task permits them to express their own independent opinions.

Overall, these findings might indicate that the behavior of LLM in shaping opinion dynamics is sensitive to the framing of the prompt: consensus-oriented instructions might produce stronger depolarization outcomes compared to opinion-elicitation prompts.

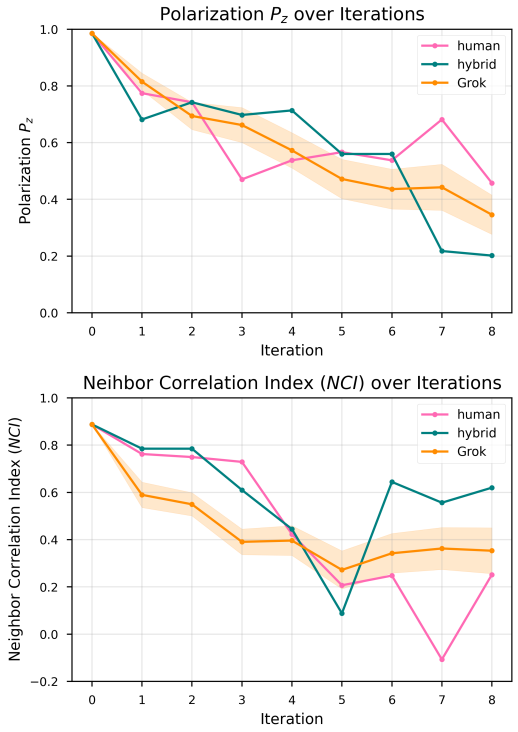

Figure 4: **Experimental results**. Evolution of Polarization (top) and Neighbors Correlation Index (bottom) over iterations across the three experimental conditions: **human-only** (pink), **Grok-only** (dark orange) and **hybrid** (steel blue). Shaded area (Grok-only) represents +/−1 standard error of the mean across 21 runs.

Discussion

We introduce an experimental framework for studying opinion dynamics and diffusion in social networks composed of both human and AI agents. The primary contribution of this paper is the development of a fully controlled experimental setting in which collective human–AI interactions can be systematically examined.

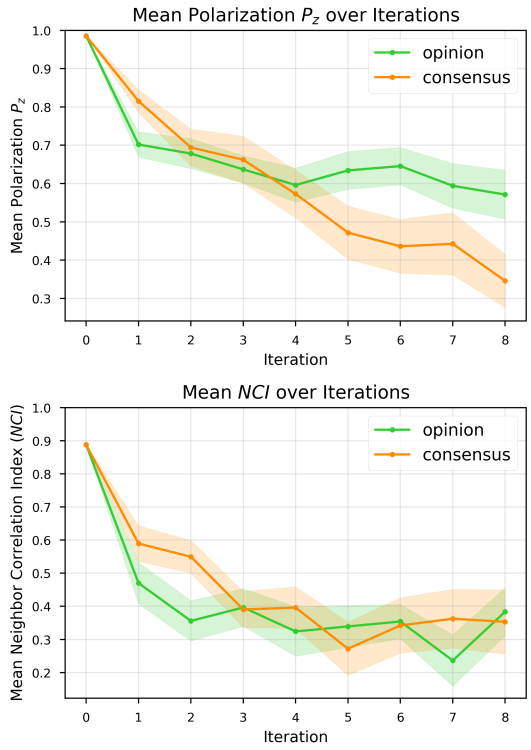

Figure 5: **Experimental results**. Evolution of Mean Polarization (top) and Mean Neighbor Correlation Index (bottom) over iterations in the Grok-only condition across the two instruction framings: **opinion** (green) and **consensus** (dark orange). Shaded area represent $+/-1$ standard error of the mean across 21 runs per condition.

Our preliminary results suggest that network composition may alter the trajectory of opinion dynamics. The hybrid human-AI condition achieved the lowest final polarization while maintaining the highest neighbor agreement, consistent with recent findings that LLMs can facilitate consensus formation in deliberative settings (Tessler et al., 2024b). This pattern may suggest that LLM agents can function as bridges between opposing viewpoints, for example by generating statements that integrate multiple perspectives or by moderating extreme positions during the revision process (Ueshima et al., 2024). Alternatively, improved consensus formation may arise from the injection of noise into the system (Shirado & Christakis, 2017), or from complementary and positive representational misalignment across different agent types (Huang et al., 2024; Shiiku et al., 2025; Sucholutsky et al., 2023). In contrast, the human-only condition exhibited the lowest final NCI despite moderate depolarization, which may indicate that opinion mixing occurred without the emergence of local consensus. This divergence between global convergence and local agreement merits further investigation, as it may reflect fundamentally different mechanisms of opinion updating between human and AI participants.

The robustness analysis of LLM-only networks revealed that prompt framing might be a factor influencing polarization outcomes. This finding carries implications for the deployment of AI agents in social platforms, such as X's Grok, where prompt design choices may inadvertently shape large-scale opinion dynamics.

Limitations & Future Directions

Several limitations constrain our findings' statistical power and generalizability. First, we used only one question. Opinion dynamics likely vary across topics with different emotional valence, political salience, or scientific consensus. Future work will expand to a full set additional questions and topics. Moreover, the prompting conditions tested represent only a narrow subset of possible LLM instructions.

Second, our robustness analysis for LLMs included only 42 runs, and the experiments involving human participants were conducted only once, limiting the statistical power and scope of the analyses. Although the results indicate certain trends, future work will substantially increase the number of runs (to 100 or more) for AI-only experiments and will include additional human experiments to ensure adequate statistical power.

Third, we tested only one LLM, Grok. Given Grok's documented history of opinion bias and recent deepfake generation following its integration with X, we considered it important to examine its behavior in opinion-diffusion dynamics and therefore deliberately chose it as the AI participant in our experiments (Migliarini et al., 2026; Mei et al., 2026; Conger et al., 2026). However, relying on a single model limits generalizability: our observed dynamics may differ across models with different training objectives, fine-tuning procedures, or safety constraints. Cross-model comparisons are therefore needed to establish robustness across LLMs.

Fourth, we used a limited set of relatively simple methods to characterize opinion dynamics. Matakos' polarization index may not fully capture clustering dynamics, and our categorical stance classification ignores the semantic richness of the statements themselves. Future work could incorporate alternative polarization measures (Bramson et al., 2016), sentence embeddings to track rhetorical and argumentative change (Shiiku et al., 2025), and continuous measures of opinion valence rather than three-category coding.

Finally, while the 5×5 grid lattice provides experimental control, it oversimplifies real-world social network structures. Future work should explore more realistic topologies, such as random regular graphs or modular networks (Marjieh et al., 2025), as well as weighted networks that can account for communities and influencers.

Taken together, our method has the potential to provide empirical insight into how different aspects of social cognition may be influenced by the introduction of AI agents (Collins et al., 2024), and more broadly enabling causal understanding of the forces shaping cultural dynamics (Boyd, Richerson, et al., 1996).

Acknowledgments

This work was supported by the NSF grant "Collaborative Research: Research Infrastructure: HNDS-I: Building Infrastructure to Study Human-AI Hybrid Societies in Experimental Social Networks" (Award BCS-2523500). ChatGPT version 5.2 (OpenAI) was used to assist manuscript editing and proofreading. Authors reviewed each of the edit suggestions, and approved the final version.

References

Acemoglu, D., & Ozdaglar, A. (2011). Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, *1*(1), 3–49.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. (2025). A foundation model to predict and capture human cognition. *Nature*, 1–8.

Boyd, R., Richerson, P. J., et al. (1996). Why culture is common, but cultural evolution is rare. *Proceedings-british academy*, *88*, 77–94.

Bramson, A., Grim, P., Singer, D. J., Fisher, S., Berger, W., Sack, G., & Flocken, C. (2016). Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, *40*(2), 80–111.

Breithaupt, F., Otenen, E., Wright, D. R., Kruschke, J. K., Li, Y., & Tan, Y. (2024). Humans create more novelty than chatgpt when asked to retell a story. *Scientific Reports*, *14*(1), 875.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., et al. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868.

Chitra, U., & Musco, C. (2020). Analyzing the Impact of Filter Bubbles on Social Network Polarization. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 115–123. https://doi.org/10.1145/3336191.3371825

Cinus, F., Minici, M., Monti, C., & Bonchi, F. (2022). The Effect of People Recommenders on Echo Chambers and Polarization. *ICWSM*, *16*, 90–101. https://doi.org/10.1609/icwsm.v16i1.19275

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

Conger, K., Freedman, D., & Thompson, S. A. (2026). Musk's chatbot flooded x with millions of sexualized images in days, new estimates show [Accessed: 2026-05-05]. *The New York Times*.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98.

Donkers, T., & Ziegler, J. (2025, June). Human-Agent Interaction in Synthetic Social Networks: A Framework for Studying Online Polarization [arXiv:2502.01340 [physics]]. https://doi.org/10.48550/arXiv.2502.01340

Doshi, A. R., & Hauser, O. P. (2024). Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, *10*(28), eadn5290.

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *Journal of mathematical sociology*, *15*(3-4), 193–206.

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying Controversy on Social Media. *Trans. Soc. Comput.*, *1*(1), 1–27. https://doi.org/10.1145/3140565

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people. *Advances in neural information processing systems*, *33*, 10659–10671.

Hashemi, F., & Macy, M. (2025). Collective social behaviors in llms: An analysis of llms social networks. *Large Language Models for Scientific and Societal Advances*.

Huang, D.-M., Van Rijn, P., Sucholutsky, I., Marjieh, R., & Jacoby, N. (2024). Characterizing similarities and divergences in conversational tones in humans and llms by sampling with people. *arXiv preprint arXiv:2406.04278*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, *596*(7873), 583–589.

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Sci. Adv.*, *5*(8), eaax0754. https://doi.org/10.1126/sciadv.aax0754

Marjieh, R., Anglada-Tort, M., Griffiths, T. L., & Jacoby, N. (2025). Characterizing the interaction of cultural evolution mechanisms in experimental social networks. *arXiv preprint arXiv:2502.12847*.

Matakos, A., Terzi, E., & Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. *Data Min Knowl Disc*, *31*(5), 1480–1505. https://doi.org/10.1007/s10618-017-0527-9

Mei, K. X., Wolfe, R., Weber, N., & Saveski, M. (2026). Grok in the wild: Characterizing the roles and uses of large language models on social media. *arXiv preprint arXiv:2602.11286*.

Migliarini, M., Ercevik, B., Olowe, O., Fatima, S., Zhao, S., Le, M. A., Sharma, V., & Panda, A. (2026). @grokset: Multi-party human-llm interactions in social media. *arXiv preprint arxiv:2602.21236*.

Pournaki, A. (2025, July). Conflicting narratives and polarization on social media [arXiv:2507.15600 [cs]]. https://doi.org/10.48550/arXiv.2507.15600

Priniski, J. H., Linford, B., Hirschmann, A., Venumuddala, S. K., Morstatter, F., Rodriguez, N., Brantingham, P. J., & Lu, H. (2026). Network structure shapes consensus dy-

namics through individual decisions. *Proc. Natl. Acad. Sci. U.S.A.*, *123*(2), e2520483123. https://doi.org/10.1073/pnas.2520483123

Rmus, M., Jagadish, A. K., Mathony, M., Ludwig, T., & Schulz, E. (2025). Generating computational cognitive models using large language models. *arXiv preprint arXiv:2502.00879*.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., & Barsoum, E. (2025). Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 5977–6043.

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025, May). The Dynamics of Collective Creativity in Human-AI Hybrid Societies [arXiv:2502.17962 [cs]]. https://doi.org/10.48550/arXiv.2502.17962

Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, *545*(7654), 370–374.

Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D., & Hawkins, R. D. (2025). Using llms to advance the cognitive science of collectives. *Nature Computational Science*, *5*(9), 704–707.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024a). Ai can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024b). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Thurner, S., Hofer, M., & Korbel, J. (2025). Why more social interactions lead to more polarization in societies. *Proc. Natl. Acad. Sci. U.S.A.*, *122*(44), e2517530122. https://doi.org/10.1073/pnas.2517530122

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876.

Ueshima, A., Jones, M. I., & Christakis, N. A. (2024). Simple autonomous agents can enhance creative semantic discovery by human groups. *Nature communications*, *15*(1), 5212.

Wang, C., Liu, Z., Yang, D., & Chen, X. (2024). Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks [Version Number: 2]. https://doi.org/10.48550/ARXIV.2409.19338

Williams, R., Hosseinichimeh, N., Majumdar, A., & Ghaffarzadegan, N. (2023). Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Zhao, W., Walasek, L., & Brown, G. D. A. (2023). The Evolution of Polarization in Online Conversation: Twitter Users' Opinions about the COVID-19 Pandemic Become More Politicized over Time (Z. Yan, Ed.). *Human Behavior and Emerging Technologies*, *2023*, 1–14. https://doi.org/10.1155/2023/9094933