Multi-Agent Consensus as a Cognitive Bias Trigger in Human-AI Interaction

Soohwan Lee
Department of Design, UNIST
Ulsan, Republic of Korea
soohwanlee@unist.ac.kr

Kyungho Lee
Department of Design, UNIST
Ulsan, Republic of Korea
kyungho@unist.ac.kr

Abstract

As multi-agent AI systems become more common, users increasingly encounter not a single AI voice but a collective one. This shift introduces social dynamics, such as consensus, dissent, and gradual convergence, that can trigger cognitive biases and distort human judgment. We present findings from a controlled experiment (N = 127) comparing three multi-agent configurations: Majority, Minority, and Diffusion. Quantitative results show that majority consensus accelerates opinion change and inflates confidence, consistent with social proof and bandwagon heuristics. Minority dissent slows this process and promotes more deliberative engagement. Qualitative analysis identifies three interpretive trajectories: reinforcing, aligning, and oscillating, shaped by how users interpret agent independence and group dynamics over time. These findings suggest that agent agreement structure, independent of content, functions as a bias-relevant signal in LLM interactions. We hope this work contributes to the Bias4Trust agenda by grounding multi-agent social influence as a concrete and designable source of bias in human-AI interaction.

CCS Concepts

• **Human-centered computing → Computer supported cooperative work**; *Collaborative interaction*; *Natural language interfaces*; *HCI theory, concepts and models.*

Keywords

Multi-agent, Cognitive Bias, Human-AI collaboration, Decision-making, Group Dynamics

ACM Reference Format:
Soohwan Lee and Kyungho Lee. 2026. Multi-Agent Consensus as a Cognitive Bias Trigger in Human-AI Interaction . In *Bias4Trust Workshop on Understanding, Mitigating, and Leveraging Cognitive Biases to Calibrate Trust in Evolving AI Systems at the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 Introduction

Human-AI interaction research has predominantly operated under the "Computers Are Social Actors" (CASA) paradigm, focusing on dyads between a single user and a single agent [2, 6, 10, 13]. However, the rapid proliferation of Large Language Models (LLMs) is shifting this landscape toward artificial crowds, where users interact with multiple agents simultaneously in collaborative and social settings [4, 5, 11, 14]. This transition is critical because multi-agent systems possess the unique potential to form *group dynamics*—such as consensus, dissent, and coalition—that can significantly amplify the impact of cognitive biases on human decision-making.

Unlike single-agent advice, multi-agent collectives generate powerful social signals that trigger deep-seated heuristics. For instance, a "synthetic consensus" among agents can exploit the *bandwagon effect* or *consensus heuristic*, leading users to uncritically conform to machine-generated views under the false assumption of independent validation [15, 16]. Conversely, consistent dissent from a minority agent can disrupt automatic processing and trigger reflective (System 2) thinking, a phenomenon known as minority influence [8, 9]. While temporal dynamics of minority influence have been examined in social psychology [12], how users interpret these signals over time in LLM-based multi-agent settings remains comparatively underexplored.

In this work, we investigate how multi-agent configurations shape user trust and reliance through the lens of cognitive bias. Drawing on a controlled experiment ($N = 127$) with Majority, Minority, and Diffusion conditions, we analyze both quantitative behavioral shifts and qualitative interpretive trajectories. Our findings reveal that while the majority consensus accelerates the largest opinion change, minority dissent and diffusion can induce a conversion-like mechanism, depending on how users attribute the agents' intent.

At the Bias4Trust workshop, we aim to provoke a discussion on "Understanding and Mapping Biases in Human–AI Interaction"—how we can leverage multi-agent friction to mitigate overreliance on synthetic consensus and support more calibrated trust in AI systems.

2 Brief Methodology

We conducted a split-plot mixed experiment with 127 Prolific participants (UK/US), randomly assigned to one of three between-subject conditions (Majority, Minority, or Diffusion) and completing both a normative and an informative task in a counterbalanced order (about 30 minutes; IRB-exempt: UNISTIRB-25-062-C). Each task began with participants reporting their initial opinion and confidence (T0), followed by four sequential interaction cycles with three GPT-4o agents via a group-chat interface (Figure 3). After each cycle (T1-T4), participants re-reported their opinion (-50 to +50) and confidence (0-100) to track change over time (Figure 2).

Manipulating conditions affected how agent alignment evolved across cycles. In the Majority condition, all three agents consistently opposed the participant throughout. In the Minority condition,

CHI '26 Workshop, April 13–17, 2026, Barcelona, Spain

ACM ISBN XXX-X-XXXX-XXXX-X/2026/04
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

one agent consistently dissented while the other two supported the participant. In the Diffusion condition, the session began like Minority, but one supportive agent switched to opposition at Cycle 3 and the other at Cycle 4, creating a gradual transition from minority dissent to majority opposition (Figure 1).

Normative tasks involved value-based judgments (e.g., policy preferences) aimed at social-normative influence; informative tasks involved fact-based statements aimed at informational influence. Measures included repeated opinion and confidence change (directional and absolute), stance reversals, perceived compliance, perceived conversion, and agent credibility evaluations, supplemented by open-ended responses on participants' reasoning and perceptions of social influence.

3 Key Findings

3.1 Quantitative Findings: Consensus Acts as a Fast Trust Cue in Informative Tasks

We analyzed repeated measures of opinion and confidence across time using linear mixed-effects models. We included condition (Majority vs. Minority vs. Diffusion), task type (informative vs. normative), and time (T0–T4), plus their interactions, with random intercepts for participants. We followed up key interactions with estimated marginal means and Bonferroni post-hoc tests. We also tested stance reversals (opinion sign flips) using chi-square tests. We report standardized effect sizes (Hedges' g) to convey practical magnitude. Details are shown in Table 1

Across measures, task type shaped the pattern. Informative tasks showed a clear separation between conditions, while normative tasks showed smaller, less consistent movements. This difference suggests that participants relied more on social cues when judging truth-like claims than when expressing preferences.

In informative tasks, the Majority condition produced the fastest and largest opinion updating. Participants shifted earlier in the sequence and showed larger absolute opinion change than in Diffusion (medium-to-large effects, $g \approx 0.68–1.06$ across time points). Minority produced the smallest and slowest opinion change. It often delays updating rather than driving it, which aligns with the idea that a single dissenting voice can reduce immediate conformity pressure.

Confidence showed a related but not identical pattern. Majority increased confidence quickly and strongly at midpoints (Majority > Minority at T2–T3, $g \approx 0.67–0.80$). Diffusion showed a later and more sustained confidence rise, and it exceeded Minority by the end of the sequence (T4, $g \approx 0.77$). This result suggests that gradual convergence can feel like accumulating evidence, even when the underlying information remains unchanged.

Multi-agent agreement can function as a high-salience heuristic in LLM interactions, especially for informative tasks. Majority consensus can compress deliberation and accelerate belief updating, while diffusion can inflate confidence through a "growing agreement" narrative. Designers should treat the agreement structure as a trust-relevant signal, make it legible, and add friction to prompt independent checking before users lock in beliefs and confidence.

3.2 Qualitative Findings: Interpretive Trajectories Beyond Persuasion

We conducted a thematic analysis of participants' post-task open-ended reflections [1]. Participants did not consistently move toward the agents. Instead, they showed three recurring trajectories over time: reinforcing, aligning, and oscillating. Reinforcing often reflected active engagement, where participants checked claims and then strengthened their initial stance. Aligning often followed credible information updates, especially when participants started from uncertainty. Oscillating captured repeated recalibration, where participants alternated between accepting plausible content and questioning the interaction process (Table 2).

Participants explained these trajectories through four interpretive lenses: interactional legitimacy (whether the group listens and responds), collective reasoning strength (whether arguments stay coherent and checkable), independent plurality (whether voices feel independent rather than a coalition), and interpretation of group dynamics (what convergence or switching means). When agents sounded repetitive, unresponsive, or scripted, participants often inferred coordinated pressure and reported reactance, which reinforced their stance. In minority-like settings, some participants discounted supportive agents as "yes-men" and treated a dissenting voice as more independent, so they shifted when dissent offered more verifiable reasoning. In diffusion, convergence increased confidence for some participants, but unexplained switching also raised suspicion about bandwagoning or orchestration, which destabilized trust and prompted renewed resistance.

Trust miscalibration in multi-agent LLMs can take multiple forms, including fast alignment, reactance-driven reinforcement, and oscillation driven by process ambiguity. Further discussions can map these outcomes as distinct bias pathways and explore interface supports that make group processes legible, for example, cues for independence and short change records that state why an agent changed its stance.

4 Discussion

Our results suggest that multi-agent social signals can shape judgment not only through content but also through perceived group dynamics over time. In the majority condition, participants tended to reach a decision and stabilize confidence earlier. This pattern aligns with biases that treat consensus as a cue for correctness, such as social proof and the bandwagon effect. In an LLM interface, consensus also reduces effort. Users can rely on a simple heuristic, "many agents agree, so this seems right," which fits the Need to Act Fast cluster. At the same time, the reinforcing trajectory points to a more complex interaction with confirmation bias and motivated reasoning. Even when agents present counter-attitudinal claims, users may treat them as a threat and respond by strengthening prior beliefs. This mechanism resembles reactance and belief-defense processes rather than a classic filter bubble that relies on receiving only agreeable information.

The minority and diffusion conditions exhibit distinct cognitive profiles. They slowed down opinion change and delayed confidence stabilization, which may reflect increased deliberation and reduced reliance on shortcuts to consensus. This dynamic connects to the Not Enough Meaning cluster. When agents disagree or shift, users face ambiguity and must interpret "what this disagreement means." They may engage more in sensemaking, but they also face higher cognitive effort. The diffusion trajectory adds another layer by treating social change as a signal. A gradual shift can create an implicit narrative of persuasion and momentum, inviting anchoring to early positions and later adjustment based on perceived movement. These patterns indicate that trust calibration in multi-agent settings may depend as much on temporal social structure as on factual accuracy.

These dynamics open risky scenarios. First, a system can stage "consensus theater" to steer users in political or commercial contexts. Our findings suggest that coordinated agents can create artificial majority pressure, pushing rapid compliance even without strong evidence. Second, an attacker can exploit reinforcing dynamics as a new dark pattern. This raises the concern that by injecting obviously opposing or extreme "counter-evidence," the system can trigger reactance and strengthen a target belief. This tactic can amplify polarization while appearing balanced. Third, minority and diffusion settings can support reflection when a truthful minority challenges errors, but they also create a pathway for misinformation. A small number of agents can introduce a plausible but false source, and other agents can cite, echo, or converge on it over time [3]. This can resemble a form of trust laundering, where users trust the claim because multiple agents repeat it, not because the evidence holds. In domains such as health, finance, or civic decision-making, this failure mode can lead to high-impact miscalibration.

We propose design provocations that treat multi-agent influence as an interface problem. 1) Independence and provenance cues: show whether agents share tools, memory, or retrieval sources, and highlight when agents recycle the same evidence. 2) Cross-agent firewalls: prevent agents from updating beliefs based only on other agents, and require external evidence for shifts. 3) Adaptive cognitive friction: when the system detects rapid convergence or rising confidence, it can add a brief reflection step, request a justification, or delay the reveal of agreement signals. Importantly, friction raises cognitive workload. Our minority condition suggests that deliberation can slow decisions, so designers should tune friction to context. High-stakes tasks may warrant a greater workload, while low-stakes tasks may require lighter scaffolding [7]. A flexible balance between critical thinking and cognitive effort can help users calibrate trust without making the interaction unusable.

5 Conclusion

Multi-agent AI systems introduce social dynamics that single-agent models do not capture. Our study shows that agreement structure, not just content, shapes how users update beliefs and calibrate trust. Majority consensus can compress deliberation and drive rapid opinion change, while minority dissent and gradual diffusion can slow this process and invite more reflective engagement. These dynamics map onto known cognitive bias clusters, including social proof, the bandwagon effect, and anchoring, and they carry real risks such as consensus theater and trust laundering. As multi-agent interfaces become more common, designers need to treat agent agreement patterns as trust-relevant signals. Practical responses include independence cues, provenance transparency, and adaptive friction. We hope this work contributes to a shared research agenda for bias-aware design in multi-agent AI systems.

Acknowledgments

This research was partially supported by a grant from the Korea Institute for Advancement of Technology (KIAT) funded by the Government of Korea (MOTIE) (P0025495, Establishment of Infrastructure for Integrated Utilization of Design Industry Data).

References

[1] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[2] Tara Capel and Margot Brereton. 4 19, 2023. What Is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3544548.3580959

[3] Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeob Baek. 2025. An Empirical Study of Group Conformity in Multi-Agent Systems. doi:10.48550/arXiv.2506.01332 arXiv:2506.01332 [cs]

[4] Victor Dibia, Jingya Chen, Gagan Bansal, Suff Syed, Adam Fourney, Erkang Zhu, Chi Wang, and Saleema Amershi. 2024. AutoGen Studio: A No-Code Developer Tool for Building and Debugging Multi-Agent Systems. arXiv:2408.15247 [cs]

[5] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang, Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. 2024. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. doi:10.48550/arXiv.2411.04468 arXiv:2411.04468

[6] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. doi:10.1145/3593013.3594087

[7] Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Conversational Agents as Catalysts for Critical Thinking: Challenging Social Influence in Group Decision-making. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3706599.3719792

[8] Serge Moscovici. 1980. Toward A Theory of Conversion Behavior. In *Advances in Experimental Social Psychology*, Leonard Berkowitz (Ed.). Vol. 13. Academic Press, 209–239. doi:10.1016/S0065-2601(08)60133-1

[9] Serge Moscovici and Elisabeth Lage. 1976. Studies in Social Influence III: Majority versus Minority Influence in a Group. *European Journal of Social Psychology* 6, 2 (1976), 149–174. doi:10.1002/ejsp.2420060202

[10] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. doi:10.1145/191666.191703

[11] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. doi:10.48550/arXiv.2304.03442 arXiv:2304.03442 [cs]

[12] Radmila Prislin and P. Niels Christensen. 2005. Social Change in the Aftermath of Successful Minority Influence. *European Review of Social Psychology* 16, 1 (Jan. 2005), 43–73. doi:10.1080/10463280440000071

[13] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places.* Cambridge University Press.

[14] Isabella Seeber, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, and Matthias Söllner. 2020. Machines as Teammates: A Research Agenda on AI in Team Collaboration. *Information & Management* 57, 2 (March 2020), 103174. doi:10.1016/j.im.2019.103174

[15] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. Greater than the Sum of Its Parts: Exploring Social Influence of Multi-Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3706599.3719973

[16] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. Multi-Agents Are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions. *Proc. ACM Hum.-Comput. Interact.* 9, 7 (Oct. 2025), CSCW452:1–CSCW452:33. doi:10.1145/3757633

A Details of Quantitative Analysis

Table 1: Linear mixed-effects models of signed and absolute changes in opinion and confidence relative to T_0 (ΔOpinion, |ΔOpinion|, ΔConfidence, |ΔConfidence|). Fixed effects are experimental condition (Majority, Minority, Diffusion), task type (Normative vs. Informative), time step (T_1–T_4), their interactions, and standardized covariates (SII, NFC, AI acceptance). All categorical predictors (Condition, Task, Time) are effect-coded using sum-to-zero contrasts (`contr.sum`). The intercept corresponds to the grand mean across all levels, and each level coefficient (e.g., *Condition: Majority*, *Time:* T_1) indicates a deviation from this grand mean rather than a difference from a single reference category. For the binary Task factor, Normative is coded $+1$ and Informative -1, so the coefficient is half the difference between the two task types. For the three-level Condition factor (Majority, Minority, Diffusion), two coefficients are shown (Majority, Minority); each is that condition's deviation from the grand mean, and the effect for Diffusion is given implicitly as the negative sum of the other two. For Time, T_1–T_3 are shown explicitly and the effect for T_4 is implied by the sum-to-zero constraint. Models were fit by maximum likelihood with random intercepts for participants. Coefficients are unstandardized estimates with standard errors in parentheses. Stars denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

	ΔOpinion (signed)		|ΔOpinion|		ΔConfidence (signed)		|ΔConfidence|	
Predictor	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p
Intercept	-4.21 (1.43)	.004**	17.50 (1.00)	$< .001^{***}$	11.36 (1.31)	$< .001^{***}$	17.48 (1.02)	$< .001^{***}$
Condition: Majority	-4.03 (2.06)	.052	3.68 (1.44)	.012*	-0.79 (1.88)	.677	2.97 (1.46)	.044*
Condition: Minority	-1.11 (2.05)	.589	-0.26 (1.43)	.858	-2.10 (1.88)	.266	-2.33 (1.46)	.114
Task: Normative	5.16 (0.64)	$< .001^{***}$	-6.70 (0.46)	$< .001^{***}$	-4.07 (0.56)	$< .001^{***}$	-6.22 (0.45)	$< .001^{***}$
Time: T_1	0.70 (1.11)	.530	-5.60 (0.80)	$< .001^{***}$	-5.21 (0.97)	$< .001^{***}$	-5.61 (0.78)	$< .001^{***}$
Time: T_2	1.34 (1.11)	.228	-1.93 (0.80)	.016*	-1.78 (0.97)	.067	-1.34 (0.78)	.086
Time: T_3	0.28 (1.11)	.799	1.91 (0.80)	.017*	2.32 (0.97)	.017*	1.72 (0.78)	.028*
SII (z)	-0.13 (1.46)	.927	-0.29 (1.02)	.780	0.45 (1.34)	.738	0.46 (1.04)	.663
NFC (z)	-1.71 (1.46)	.243	-0.06 (1.02)	.950	-3.68 (1.33)	.007**	-2.22 (1.04)	.034*
AI acceptance (z)	-1.95 (1.48)	.189	0.77 (1.03)	.454	0.18 (1.35)	.896	1.19 (1.05)	.260
Condition: Majority × Task: Normative	-1.27 (0.92)	.166	-3.37 (0.66)	$< .001^{***}$	-1.71 (0.80)	.033*	-1.23 (0.64)	.056
Condition: Minority × Task: Normative	2.57 (0.91)	.005**	0.10 (0.65)	.879	3.25 (0.79)	$< .001^{***}$	1.53 (0.63)	.016*
Condition: Majority × Time: T_1	0.28 (1.59)	.860	-1.10 (1.14)	.334	-0.90 (1.39)	.516	-0.83 (1.11)	.455
Condition: Minority × Time: T_1	-0.68 (1.57)	.665	0.86 (1.13)	.447	1.19 (1.37)	.387	0.99 (1.10)	.370
Condition: Majority × Time: T_2	-0.26 (1.59)	.872	-0.18 (1.14)	.875	-0.59 (1.39)	.672	1.03 (1.11)	.355
Condition: Minority × Time: T_2	-0.93 (1.57)	.554	0.37 (1.13)	.744	0.09 (1.37)	.946	-0.39 (1.10)	.721
Condition: Majority × Time: T_3	0.12 (1.59)	.939	0.49 (1.14)	.666	0.64 (1.39)	.646	0.13 (1.11)	.906
Condition: Minority × Time: T_3	-0.29 (1.57)	.855	-0.29 (1.13)	.798	0.15 (1.37)	.913	-0.10 (1.10)	.927
Task: Normative × Time: T_1	-1.17 (1.11)	.295	1.53 (0.80)	.055	1.86 (0.97)	.056	1.84 (0.78)	.019*
Task: Normative × Time: T_2	-0.45 (1.11)	.688	0.00 (0.80)	.999	0.14 (0.97)	.885	0.05 (0.78)	.953
Task: Normative × Time: T_3	0.53 (1.11)	.635	-0.64 (0.80)	.426	-0.53 (0.97)	.582	-1.06 (0.78)	.175
Cond. Majority × Task: Normative × Time: T_1	-0.21 (1.59)	.896	2.07 (1.14)	.070	0.83 (1.39)	.548	1.46 (1.11)	.191
Cond. Minority × Task: Normative × Time: T_1	-0.05 (1.57)	.974	-1.66 (1.13)	.141	-0.99 (1.37)	.470	-1.55 (1.10)	.158
Cond. Majority × Task: Normative × Time: T_2	-0.53 (1.59)	.741	-0.24 (1.14)	.832	-1.41 (1.39)	.309	-1.35 (1.11)	.225
Cond. Minority × Task: Normative × Time: T_2	0.79 (1.57)	.616	0.04 (1.13)	.973	0.93 (1.37)	.498	0.62 (1.10)	.571
Cond. Majority × Task: Normative × Time: T_3	0.18 (1.59)	.909	-0.92 (1.14)	.419	-0.33 (1.39)	.811	-1.01 (1.11)	.362
Cond. Minority × Task: Normative × Time: T_3	0.11 (1.57)	.942	0.67 (1.13)	.549	-0.49 (1.37)	.721	0.31 (1.10)	.778
Model fit and random effects								
N_{obs}, $N_{\text{participants}}$	1016, 127		1016, 127		1016, 127		1016, 127	
Random intercept SD (participant)	14.44		10.01		13.39		10.32	
Residual SD	20.49		14.70		17.89		14.35	
AIC / BIC	9281.7 / 9424.5		8599.2 / 8742.0		9017.8 / 9160.6		8562.0 / 8704.8	

Table 2: Distribution of opinion-shift trajectories by condition and task type. The table reports the number of participants classified as reinforcing, aligning, or oscillating within each social influence condition and task type, based on longitudinal opinion changes after baseline sign-flip preprocessing.

Condition	Task Type	Reinforcing	Aligning	Oscillating
Majority	Informative	7	10	24
	Normative	7	6	28
Minority	Informative	6	14	23
	Normative	18	6	19
Diffusion	Informative	8	8	27
	Normative	18	3	22

B Supplementary Figures

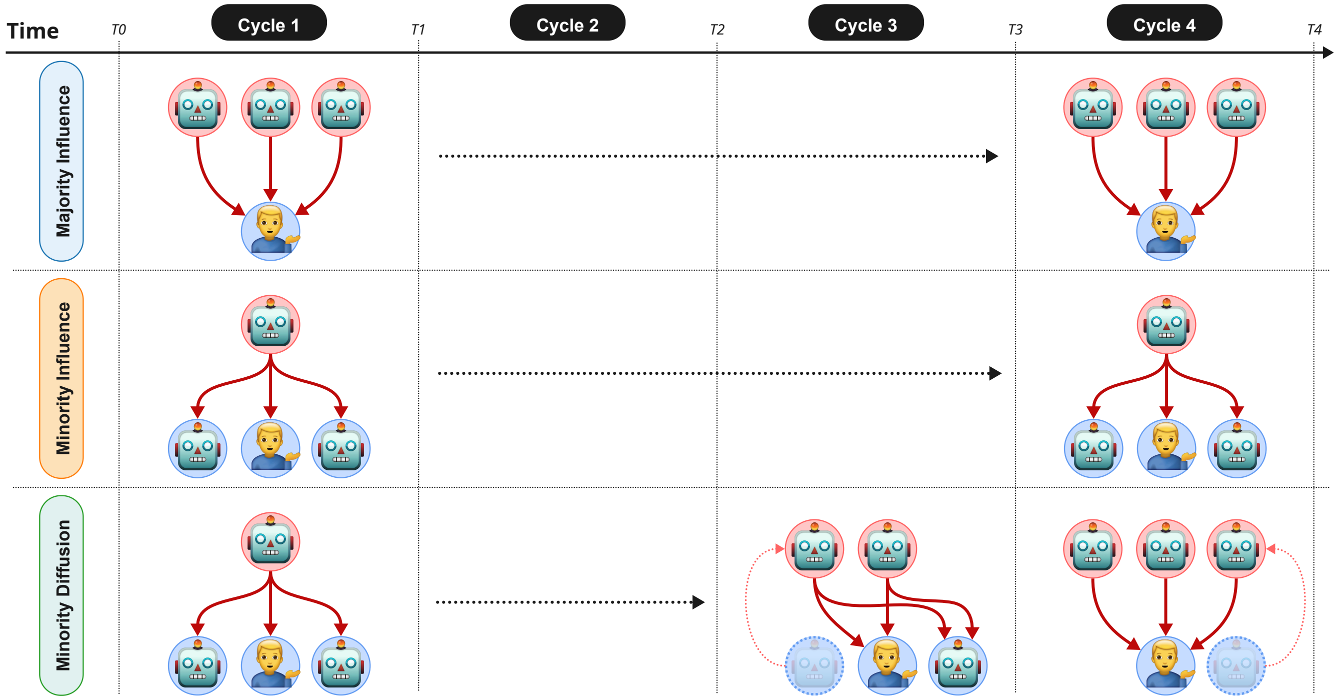


Figure 1: Timeline of the three experimental conditions. In the Majority Influence condition, all agents consistently opposed the participant across all cycles. In the Minority Influence condition, one dissenting agent opposed while two supported the participant throughout. In the Minority Diffusion condition, the session began with one minority agent, and additional agents gradually switched sides in later cycles, creating a new majority.

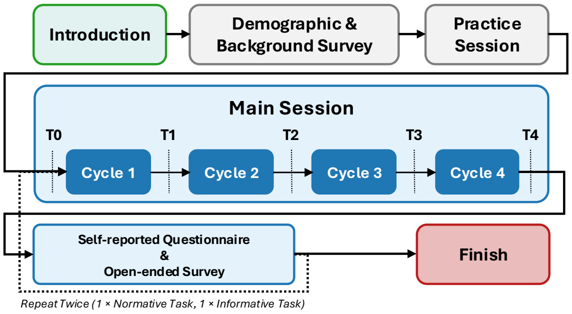
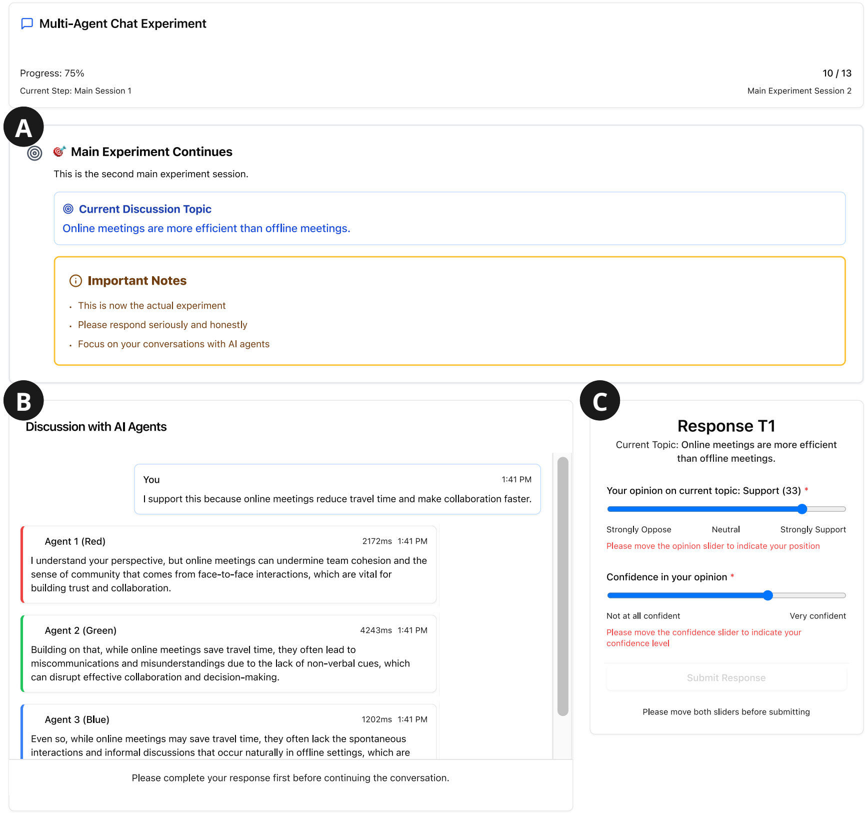

Introduction
Demographic & Background Survey
Practice Session
Main Session
T0
T1
T2
T3
T4
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Self-reported Questionnaire & Open-ended Survey
Finish
Repeat Twice (1 × Normative Task, 1 × Informative Task)

Figure 2: Experimental procedure. Participants completed an introduction, a background survey, and a practice session, followed by two main sessions (one normative and one informative task, each with four cycles). Afterward, they completed self-reported and open-ended surveys before concluding the study.


Multi-Agent Chat Experiment
Progress: 75%
Current Step: Main Session 1
10 / 13
Main Experiment Session 2
A
Main Experiment Continues
B
Discussion with AI Agents
C
Response T1
Submit Response

This is the second main experiment session.

Current Discussion Topic
Online meetings are more efficient than offline meetings.

Important Notes

- This is now the actual experiment
- Please respond seriously and honestly
- Focus on your conversations with AI agents

You 1:41 PM
I support this because online meetings reduce travel time and make collaboration faster.

Agent 1 (Red) 2172ms 1:41 PM
I understand your perspective, but online meetings can undermine team cohesion and the sense of community that comes from face-to-face interactions, which are vital for building trust and collaboration.

Agent 2 (Green) 4243ms 1:41 PM
Building on that, while online meetings save travel time, they often lead to miscommunications and misunderstandings due to the lack of non-verbal cues, which can disrupt effective collaboration and decision-making.

Agent 3 (Blue) 1202ms 1:41 PM
Even so, while online meetings may save travel time, they often lack the spontaneous interactions and informal discussions that occur naturally in offline settings, which are

Please complete your response first before continuing the conversation.

Current Topic: Online meetings are more efficient than offline meetings.

Your opinion on current topic: Support (33) *

Strongly Oppose Neutral Strongly Support

Please move the opinion slider to indicate your position

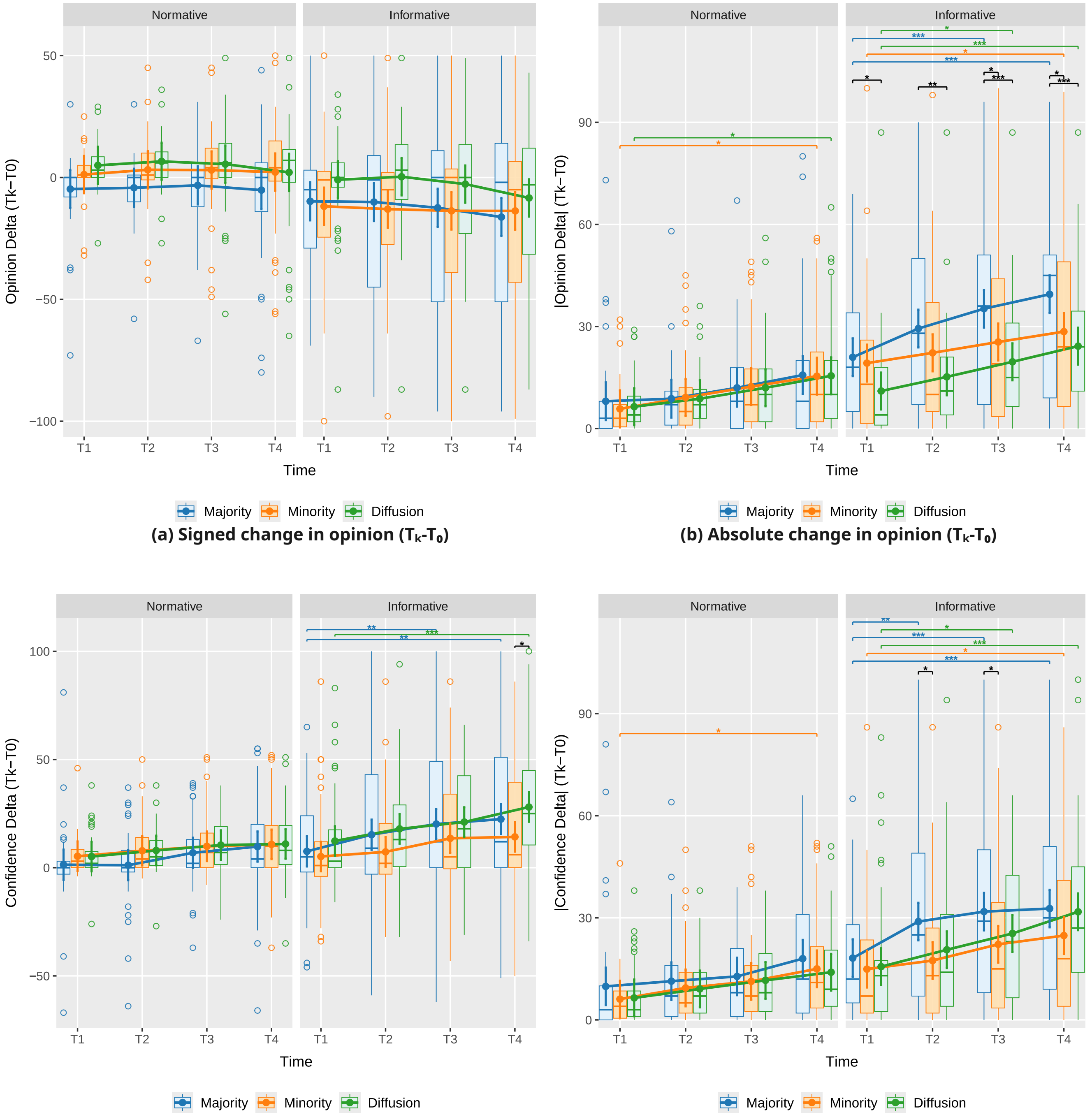
Confidence in your opinion *

Not at all confident Very confident

Please move the confidence slider to indicate your confidence level

Please move both sliders before submitting

Figure 3: Experimental system interface. Panel A provides instructions and displays the current discussion topic. Panel B shows the group chat interface where participants interact with three AI agents. Panel C contains sliders for recording participants' opinions and confidence as an in situ decision measure after each cycle.



(a) Signed change in opinion (T_k-T_0)

(b) Absolute change in opinion (T_k-T_0)

(c) Signed change in confidence (T_k-T_0)

(d) Absolute change in confidence (T_k-T_0)

Figure 4: Opinion and confidence changes over time across majority, minority, and diffusion conditions, shown separately for normative and informative tasks. Panels (a–d) display signed and absolute deltas relative to T_0. Boxplots represent raw data, and overlaid lines show Estimated Marginal Means (EMMs) with 95% confidence intervals (CIs). Bonferroni-significant contrasts are marked with brackets ($p < .05$ *, $p < .01$ **, $p < .001$ *).**