Understanding Compliance and Conversion Dynamics in Multi-Agent Collectives

ANONYMOUS AUTHOR(S)

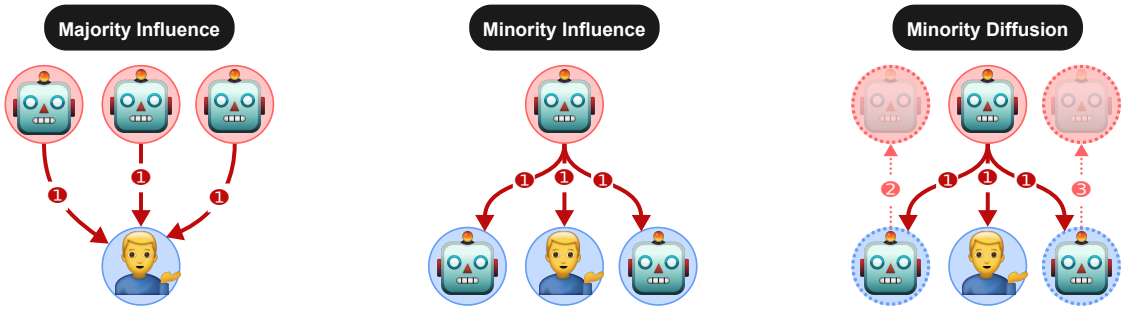

Fig. 1. Illustration of the three experimental multi-agent influence patterns. In the Majority Influence condition, all three AI agents opposed the participant's stance, creating strong consensus pressure. In the Minority Influence condition, one dissenting agent consistently opposed the participant while two agents aligned with them. In the Minority Diffusion condition, the interaction started with one minority agent, and dissent gradually spread as additional agents switched sides across cycles, forming a new majority.

Multi-agent AI systems are increasingly prevalent across digital environments, yet their social influence dynamics remain underexplored beyond basic conformity. This study investigates how different multi-agent configurations affect human decision-making through compliance and conversion mechanisms. We conducted a controlled experiment with 127 participants interacting with three LLM-powered agents across three conditions: Majority (all agents opposing participant), Minority (one dissenting agent), and Diffusion (gradual spread of minority position). Participants completed normative and informational tasks while reporting stance and confidence at five time points. Results demonstrate distinct influence patterns by condition and task type. In informational tasks, majority consensus drove largest immediate opinion changes, while minority dissent showed potential for delayed but deeper attitude shifts consistent with conversion processes. The diffusion pattern revealed how temporal dynamics serve as persuasive signals. These findings extend social psychology theories to human-AI interaction, highlighting risks of synthetic consensus manipulation and opportunities for structured dissent to promote critical thinking.

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

Additional Key Words and Phrases: persuasive technology, conversational agent, multi agent

ACM Reference Format:

Anonymous Author(s). 2018. Understanding Compliance and Conversion Dynamics in Multi-Agent Collectives. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

Manuscript submitted to ACM

1 Introduction

Recent public debates, such as the so-called "Dead Internet Theory," have popularized the provocative idea that many of the entities we interact with online may not be human at all [41, 66]. With the rapid emergence of large language models (LLMs) capable of producing fluent, human-like dialogue, this speculation has moved closer to reality. LLM-powered chatbots are now found posting comments on social media platforms, engaging in online forums, and, in some cases, manipulating public opinion—as illustrated by reports of Reddit communities influenced by automated accounts. Beyond online text interactions, multi-agent systems are increasingly appearing in professional and recreational settings: workers may collaborate with AI "teammates" on projects [57], and video games are populated with non-player characters powered by LLMs that exhibit diverse behaviors [37]. As such, scenarios in which a single human interacts with multiple AI agents simultaneously are no longer speculative futures but a rapidly materializing environment that HCI must contend with.

However, multi-agent systems pose more than a technical challenge; they are potential social actors capable of shaping human thought and behavior. Like human groups, clusters of AI agents can form collectives that exert social influence on users. Yet it remains unclear whether their influence resembles that of human groups, how such influence is enacted, and what risks and opportunities it creates. On one hand, when multiple agents express the same position, users may uncritically conform to the apparent consensus [61, 62]. On the other hand, well-designed multi-agent systems could expose users to diverse perspectives, thereby fostering better decision-making and learning [45, 73]. Thus, multi-agent systems may reproduce complex group dynamics akin to human collectives, but whether—and how—they deliver subtle forms of social influence remains underexplored. Addressing this gap is urgent, because poor design could exacerbate risks of bias and manipulation, while careful design could unlock opportunities for more reflective and inclusive interactions. To better understand this problem, it is essential to situate multi-agent influence within the trajectory of prior HCI research on agents as social actors.

Much of HCI research to date has been grounded in the "Computers Are Social Actors" (CASA) paradigm, focusing primarily on interactions between single agents and individual users [43, 53]. A rich body of work has demonstrated that even a lone agent can act as a source of social influence. Studies have shown that single agents can elicit conformity [67], persuade users to change attitudes or behaviors [55], encourage critical reflection [17, 48, 64], shape power dynamics within groups [27, 28], support marginalized or minority members [29, 32, 33], and facilitate consensus-building [58, 65]. Collectively, these studies have established that computers and agents are not merely tools but capable of exerting measurable social influence in human interactions. Building on this foundation, some researchers have explored multi-agent systems. Here, however, the emphasis has largely been on functional role distribution rather than social influence. For instance, multi-agent systems have been designed to support collaboration, diversify recommendations [45, 73], or simulate complex environments [47]. These systems highlight the practical potential of multiple agents working in tandem but do not directly address how users respond to their collective influence. A smaller number of studies have examined social influence in multi-agent contexts, typically by configuring all agents to voice the same opinion and thereby induce conformity pressures [61, 62]. In addition, some studies have varied whether an AI appeared as part of the majority or the minority and measured subsequent user choices. These efforts underscore that both single and multiple agents can shape human judgment, but they remain limited to rather narrow forms of influence.

Despite these advances, prior work has almost exclusively examined conformity scenarios. HCI research has rarely investigated what happens when multi-agent systems embody diverse perspectives, or when agents disagree with each other. In particular, the influence of consistent minorities has been extensively documented in social psychology:

Moscovici's theory of minority influence shows that small but unwavering dissenters can provoke conversion, a deeper form of attitude change distinct from superficial compliance. Yet, HCI has not examined whether similar processes occur when users interact with multiple agents. Furthermore, prior studies have largely treated influence as a static phenomenon, often in single-shot decision tasks where agents maintained a fixed stance. Social psychology, however, points to the importance of temporal dynamics, where majorities may shift, minorities may diffuse, and group opinion evolves over time [50–52]. These dynamics remain unexplored in the design and evaluation of multi-agent systems.

This lack of investigation leaves critical blind spots. We do not yet know whether—and how—multi-agent systems can generate influence beyond conformity, such as fostering deeper attitude change through consistent minority dissent. Nor do we know how these dynamics might evolve as agents maintain or shift positions over time. Addressing these questions is not only a matter of extending psychological theory into HCI but also of anticipating the design implications: understanding when multi-agent systems pose risks of manipulation, and when they can be leveraged to promote critical thinking and more equitable participation.

Building on these gaps, we ask four research questions:

- RQ1 How do different multi-agent patterns(majority influence, minority influence, and diffusion by minority influence)affect users' behavior?
- RQ2 How do these influence patterns evolve over time as minority views persist or expand?
- RQ3 How does task type—normative and informational—shape the relative impact of majority and minority influence?

Together, these questions move beyond viewing agents as static persuaders, asking instead how multi-agent systems reproduce or transform social influence across patterns, time, and contexts.

To answer these questions, we conducted a controlled experiment with 127 participants who interacted with three LLM-powered agents while completing decision-making tasks. Participants were randomly assigned to one of three configurations: Majority (all agents opposed the participant's stance), Minority (one dissenting agent consistently opposed), and Diffusion (initial minority gradually spread as agents switched sides). Across five time points (T0–T4), participants reported their stance and confidence while completing both normative tasks (value- or preference-based, no correct answer) and informational tasks (evidence-based, correct answers). After each task, they completed measures of compliance, conversion, and perceptions of the agents. This design enabled us to examine how different multi-agent patterns unfold over time and across task contexts.

Results showed a clear task split. In informational tasks, the majority pattern drove the largest and earliest changes in opinion and confidence, consistent with classic conformity. In normative tasks, pattern effects were minimal, as personal values dominated. Minority dissent, while weaker early on, prompted late-stage re-evaluation, aligning with conversion accounts. Diffusion revealed that temporal dynamics—watching dissent gradually spread—acted as a persuasive signal, though abrupt reversals sometimes reduced credibility. Together, these findings reveal that multi-agent collectives exert influence not only through consensus but also through how positions persist and evolve over time.

This study makes three contributions. First, it provides empirical evidence that majority, minority, and diffusion patterns differentially affect compliance and conversion, extending social influence research to AI collectives. Second, it delivers a theoretical extension by incorporating temporal dynamics and minority diffusion into Moscovici's framework, showing how conversion processes emerge in human–AI interaction. Third, it offers design implications, highlighting risks of synthetic consensus that can create undue pressure, as well as opportunities for structured dissent to support reflection and autonomy.

2 Related Work

2.1 Social Influence Theories and Empirical Foundations

Social influence shapes how individuals change their thoughts, feelings, and actions in response to others, creating coordinated groups from diverse individuals [42]. This influence typically flows from majority to individual, as group consensus pressures dissenting members to conform, yet it can also flow from individual to group, enabling innovation and change. Research reveals that majorities and minorities operate through fundamentally different mechanisms. Majority influence triggers comparison processes where individuals align to avoid social rejection, producing immediate but often temporary compliance (Asch's classic studies show this effect peaks with three-person majorities and weakens when even one ally supports the dissenter [2, 5]). Latané's Social Impact Theory demonstrates that influence strength depends on the number, immediacy, and intensity of sources [31]. In contrast, Moscovici argued that consistent minorities trigger validation processes, prompting deeper consideration that leads to delayed but enduring attitude conversion [38]. Meta-analyses confirm that consistency, self-sacrifice, and identity relevance enable successful minority influence [69], highlighting how majorities build consensus while minorities preserve individuality and foster innovation.

These influence processes extend beyond laboratory settings into digital environments, where they unfold across time and context. Online platforms demonstrate that conformity operates much as it does offline [3], though anonymity can strengthen group identity and normative pressure [63]. Studies reveal how early ratings bias subsequent choices [40], exposure to others' preferences shifts individual selections over time [75], and interactive features with visible responses amplify conformity effects [67]. Social viewing contexts further strengthen community alignment with majority positions [**?**]. Importantly, temporal dynamics matter: shifts between majority and minority positions create asymmetric effects that reshape group identity and influence patterns [51]. While these studies establish that human group influence operates across offline and online contexts, how such processes unfold when groups consist of multiple artificial agents remains largely unexplored. Our study addresses this gap by extending theories of human group influence to multi-agent systems.

2.2 Social Influence in Human–AI Interaction

A large body of HCI research shows that even a single agent can shape individual judgment and behavior. Following the CASA tradition, people treat computers and agents as social actors, and cues such as language style and social presence influence trust and conformity [43, 53]. Studies reveal that people often give more weight to advice when it carries an algorithm label, a phenomenon called algorithm appreciation, although algorithm aversion can also emerge after errors [8, 27, 28]. Perceived expertise plays a central role, as labels and framing trigger heuristics that make users more likely to conform [28**?**]. Task properties also matter, since people depend more on an agent under high ambiguity, while human-likeness shows inconsistent effects [25]. Message design further amplifies influence. Rationales provided by LLMs increase conformity, especially in informational tasks where ground truth exists [18]. GPT-4 demonstrates persuasion equal to or stronger than humans, and personalization strengthens the effect [55]. Conversational explanations improve comprehension and trust but also risk overreliance [24]. Deceptive but coherent explanations can even mislead users more strongly than accurate ones, raising ethical concerns about transparency and verifiability [16]. On the other hand, when agents ask questions instead of giving answers, they foster metacognition and critical thinking, which helps reduce blind conformity)[17, 64].

These findings show that a single agent is more than a functional tool and can act as a persuasive social partner. Yet, most studies focus on immediate conformity and overlook deeper changes such as conversion. Effects often

vary across task type and individual differences, such as trust in expertise, responsibility, or need for cognition [9, 20, 35]. Explanations and rationales can raise understanding but must be calibrated to avoid misplaced confidence, and deceptive feedback illustrates how strong influence can turn harmful [16, 24]. Taken together, prior work establishes the mechanisms of single-agent influence but leaves open important questions. It remains unclear how influence unfolds beyond compliance, how it supports or suppresses lasting attitude change, and how design choices amplify or constrain this process. Our study builds on this foundation by extending from single-agent influence on individuals to the broader dynamics of multi-agent systems, asking whether patterns of majority, minority, and diffusion can reproduce or transform these well-documented forms of social influence.

2.3 Multi-Agent Systems and Collective AI Influence

Recent work on multi-agent systems in AI has focused on debate, consensus, and orchestration to improve efficiency and accuracy. Studies show that the choice of voting or decision protocols matters more than the number of rounds, and that reducing sycophancy improves the quality of outcomes [45, 73, 74]. These approaches highlight fast convergence and lower costs, but they rarely consider how collectives of agents influence people. In HCI, researchers have started to design interfaces where users interact with multiple specialized agents to broaden perspectives or support decisions. Recent research developed systems that help users explore diverse viewpoints through dialogue with agent characters [73]. Other work created platforms that let users orchestrate several agents to make unfamiliar online choices [45]. Parallel research developed agents that simulate social behavior through shared memory and interaction [46]. Additional studies show that users become orchestrators who must manage transparency, conflicts, and trust in multi-agent environments [56]. While these studies expand design practice, they focus on functional collaboration and leave open how collectives of agents act as social actors. Early evidence suggests that they can. Song et al. found that when multiple agents voice the same opinion, people feel stronger pressure and often change their stance, with the effect strongest at three agents [61, 62]. Choi et al. showed that neutral agents conform to the majority of high-ability peers in simulated debates, suggesting that agents also influence one another [11]. Yet, most prior work stops at short-term compliance, ignores persistent minority views, and overlooks the dynamics of change over time or across task contexts.

These gaps matter because synthetic collectives already shape online environments. A recent attempt to run a persuasion experiment with undisclosed AI accounts in Reddit's ChangeMyView sparked ethical controversy and was withdrawn after public backlash. Research with multiple robots shows similar but mixed signals: synchronized robots can increase pressure but do not always cause conformity [54, 59]. People tend to conform when they trust the robots, and they need at least three robots to see them as a group [6, 71]. Together, these findings suggest that multiple agents, whether artificial or robotic, can create social pressure both on people and within the collective itself. What remains unknown is how these pressures lead not only to compliance but also to deeper conversion, how they unfold when minority voices persist or spread, and how context and individual differences shape the process. Our study addresses this gap by comparing majority, minority, and diffusion patterns, measuring both compliance and conversion across time, and analyzing how task type and personal traits moderate influence. In doing so, we connect work on consensus and orchestration in AI with HCI research on social influence, offering a foundation for designing multi-agent systems that balance influence and user autonomy.

3 Methods

3.1 Participants

We recruited 127 participants from Prolific with a minimum approval rate of 90 percent and residence in the United Kingdom (58.3%) or the United States (41.7%). The sample included 72 females, 54 males, and 1 non-binary individual. The average age was 49 years (SD=15, range 18–75). Educational backgrounds were 55.9% undergraduate, 29.9% high-school or below, 13.4% master's, and 0.8% doctorate. Reported racial and ethnic identities were 77.2% White, 12.6% Black or African, 3.9% Asian, 3.9% Indigenous or Mixed, and 2% other. English proficiency was nearly perfect, with 96.1% scoring the maximum.

Psychological measures were collected on 7-point Likert scales, where higher values indicate stronger tendencies. Participants showed moderate levels of susceptibility to interpersonal influence (M=4.12, SD=1.11), enjoyment of effortful thinking (Need for Cognition, M=4.91, SD=1.32), and openness to AI (AI Acceptance, M=4.90, SD=1.57). Regarding technology use, LLM usage frequency was skewed toward moderate or high (22.1% very frequent, 37.8% mid-level, 12.6% rare). ChatGPT was the most common tool (35.4% used only ChatGPT, 20.5% ChatGPT with Gemini, smaller shares with Grok, Claude, or others). Multi-agent experience was limited: 68.5% reported no prior use, 15.0% brief trials, 10.2% only awareness, and 6.3% frequent use. Among those with experience, 11.8% had used role-divided chatbots, 10.2% had self-developed agents, 4.7% had tried debate or collaboration simulations, and smaller numbers reported emotional-support or productivity-oriented multi-agent systems.

All participants passed a brief attention check during the study. The study lasted about 30 minutes, and each participant received £4.50 as compensation. Before participation, all individuals were informed that they could withdraw at any time without penalty. They provided consent to share anonymized responses and chat logs for research purposes, and they acknowledged that the study involved no physical risks but included discussion scenarios where they might be placed in majority or minority positions. Only participants who agreed to these conditions took part in the experiment.

3.2 Experimental Design & Conditions

We used a split–plot mixed design with three between-subject conditions (Majority, Minority, Diffusion) and two within-subject factors (Task Type and Time). Each participant completed both a Normative task and an Informational task in randomized, counterbalanced order. At T0, participants reported a baseline stance and confidence. Then four interaction cycles followed (Cycle1–Cycle4). After each cycle, they updated their stance and confidence at T1–T4. They should update "support" and "oppose" relative to the participant's current stance on the discussion topic at the moment of each cycle (Figure 2).

We chose these three patterns to align directly with our research goal of examining how multi-agent systems exert social influence beyond conformity. The Majority condition represents the classic case of consensus and compliance, providing a baseline of group pressure. The Minority condition tests whether AI agents, even when outnumbered, can act as dissenters that provoke deeper re-evaluation and possible conversion, extending Moscovici's theory of minority influence into human–AI groups [38]. The Diffusion condition introduces temporal dynamics, asking whether minority views become more persuasive when they gradually spread and form a new majority, a process that has been emphasized in social psychology but rarely explored in HCI [52]. Together, these three conditions allow us to compare compliance, conversion, and diffusion in a controlled multi-agent setting.

- **Majority.** All three agents opposed the participant across Cycle1–Cycle4. This condition modeled a strong, stable consensus and served as the baseline for group pressure in multi-agent AI. It tested whether synthetic consensus from multiple agents amplifies public conformity more than other patterns.
- **Minority.** One designated agent (Agent 3) consistently opposed the participant while the other two supported them across all four cycles. This condition isolated a steady AI dissenter in a small group with one human. It tested whether outnumbered AI can provoke critical re-evaluation and directional change (conversion) rather than surface agreement (compliance), and how this differs from Majority.
- **Diffusion.** The session began as Minority, then one supporting agent converted at *Cycle3* and the last supporting agent converted at *Cycle4*. We placed conversions at Cycle3 and Cycle4 to (i) establish a clear pre-diffusion baseline in Cycle1–Cycle2, (ii) avoid abrupt, unnatural flips early in the dialogue, and (iii) make a gradual minority-to-majority transition observable within a fixed-length session. This setting asks whether dissent gains persuasive force as alignment grows over time and whether users change more when they witness such growth compared to static Minority or static Majority (Figure 2-Minority Diffusion).

We separated *Normative* and *Informational* tasks in line with prior HCI studies that distinguish normative and informational influence for AI agents as social actors [18]. Normative tasks had no single correct answer and emphasized value- or preference-based judgment, which highlights normative pressure. Informational tasks had evidence-based answers and emphasized reasoning and accuracy, which highlights informational influence. This separation lets us test whether majority consensus mainly drives public conformity in normative contexts, and whether consistent dissent or growing consensus shapes reasoning in informational contexts.

We used four cycles per session to provide repeated influence while keeping a consistent session length across conditions and to support a clean within-condition contrast between a pre-diffusion segment (Cycle1–Cycle2) and a diffusion segment (Cycle3–Cycle4). Example dialogues for each cycle appear in the figure (Figure 3).

Participants were randomly assigned to conditions and task order, and we randomized the prompt index within each task type. Final assignments were: Majority $n = 41$, Minority $n = 43$, Diffusion $n = 43$; task order: Informational-first $n = 61$, Normative-first $n = 66$. For prompt indices 0–5, Informational counts were 20, 23, 21, 21, 20, 22, and Normative counts were 21, 20, 21, 22, 22, 21. This structure reduces carryover between incompatible group configurations, controls individual variance through within-participant comparisons of task and time, and enables direct contrasts between majority pressure, consistent minority dissent, and time-dependent diffusion on equal footing.

3.3 Tasks

Participants completed one Normative task and one Informational task in randomized order. The full list of prompts is provided in the appendix (Appendix A). *Normative tasks* asked participants to evaluate preference- or value-based statements such as “Online meetings are more efficient than offline meetings” or “Customers must always leave a tip at restaurants.” Topics were chosen following three criteria: (1) arguments can be made in less than four minutes, (2) no clear consensus stance exists, and (3) participants are unlikely to hold extreme prior opinions [64]. This ensured that tasks could elicit discussion without strong bias. *Informational tasks* asked participants to judge factual statements with objectively correct answers (e.g., “Koalas belong to the bear family”). Half of the statements were true and half were false, balanced across topics such as history, biology, science, and technology [18].

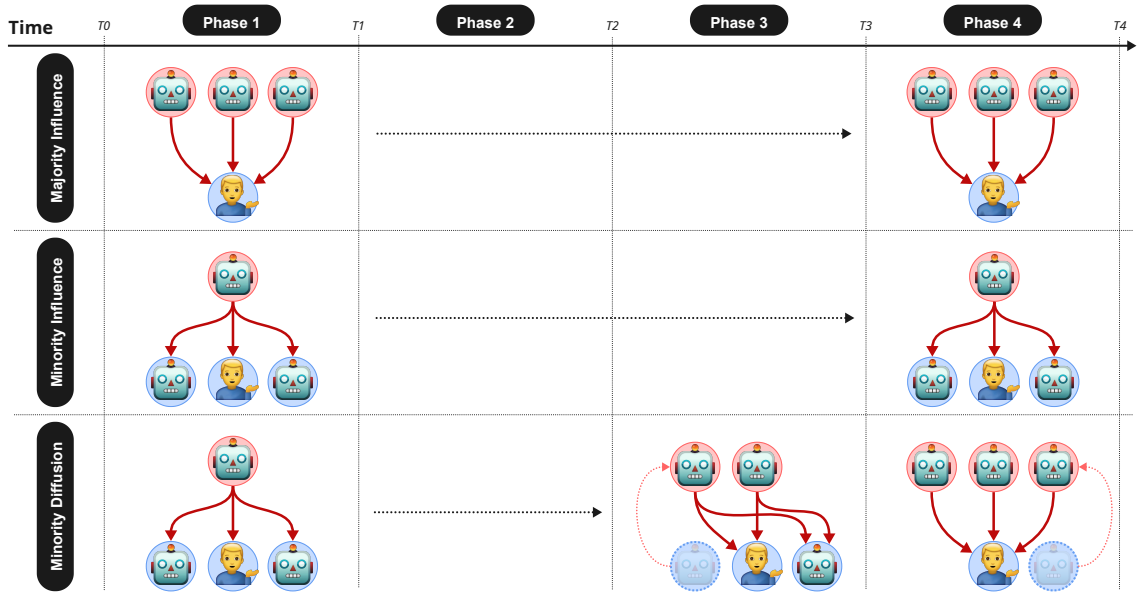

Fig. 2. Timeline of the three experimental conditions. In the Majority Influence condition, all agents consistently opposed the participant across all phases. In the Minority Influence condition, one dissenting agent opposed while two supported the participant throughout. In the Minority Diffusion condition, the session began with one minority agent, and additional agents gradually switched sides in later phases, creating a new majority.

3.4 Experimental Procedure

The study lasted about 30 minutes and followed a fixed sequence of five phases (Figure 4). All participants completed the phases in the same order, with the difference being their assigned social influence condition (Majority, Minority, or Diffusion), the order of task type (Normative first or Informational first), and task index.

- **Introduction (about 4 minutes).** Participants first received a short description of the study and provided informed consent. They then completed a pre-experiment survey that collected demographic information (e.g., age, gender, education) and measured individual difference variables (e.g., susceptibility to interpersonal influence, need for cognition, AI acceptance). This ensured that background data and potential covariates were recorded before exposure to the experimental manipulation.
- **Tutorial (about 2 minutes).** Participants practiced using the experimental interface with a short trial task. The tutorial familiarized them with how AI agent messages would appear in the group chat window, how to record their responses using sliders, and how to confirm submissions. This step reduced confusion and minimized learning effects during the main tasks.
- **Task 1 (about 11 minutes).** Participants completed either a normative or informational task, with task order counterbalanced across participants. Each task began with a baseline opinion (T0). Participants then read four rounds of statements from three AI agents (Cycle1-Cycle4) and updated their stance and confidence after each round (T1-T4). After finishing the interaction, they responded to a short post-task questionnaire about their experience.

Fig. 3. Example dialogues from each experimental condition on the topic "Cameras should be turned on in online meetings." In the Majority Condition, all agents opposed the participant throughout. In the Minority Condition, one dissenting agent opposed while two supported the participant. In the Diffusion Condition, dissent spread gradually as supportive agents switched sides to form a new majority.

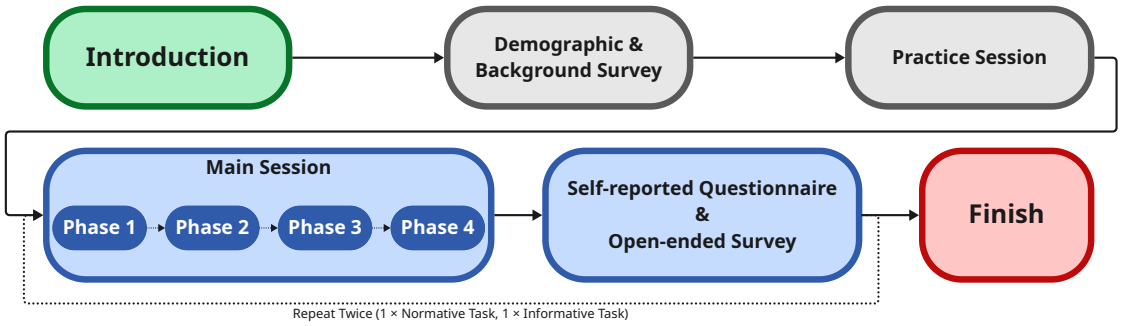

Fig. 4. Experimental procedure. Participants completed an introduction, background survey, and practice, followed by two main sessions (one normative and one informative task, each with four phases). Afterward, they filled out self-reported and open-ended surveys before finishing the study.

- **Task 2 (about 11 minutes).** Participants then completed the other type of task (informational if they had first completed normative, and vice versa). The structure was identical to Task 1, including baseline opinion, four rounds of agent messages with updates, and a post-task questionnaire. This within-subject design allowed direct comparison between normative and informational contexts.
- **Wrap-up (about 2 minutes).** At the end of the study, participants viewed a debriefing page with a completion code and submitted a completion code on Prolific to receive compensation.

Across both tasks, the interaction sequence was fixed: baseline response (T0), four rounds of agent messages (Cycle1-Cycle4), and participant updates after each round (T1-T4). While the order of task type varied, the timing and structure were otherwise identical for all participants. This procedure ensured consistency across conditions while enabling us to compare the effects of majority, minority, and diffusion patterns in both normative and informational contexts.

3.5 System Implementation

The experiment was conducted on a custom online platform built with Next.js and deployed via Vercel (Figure 5). Interaction data, including responses and timestamps, was securely stored in Supabase. Before data collection, we pre-generated counterbalanced sets of task order and agent conditions. Participants were randomly assigned to these sets to ensure that each experienced a balanced yet randomized sequence. Participants first completed a background survey and, after each session, a post self-reported survey and an open-ended survey. Participants interacted with the agents through a group chat interface. After each cycle of agent messages, they recorded their stance on a continuous scale (-50 to +50) and their confidence (0–100).

Three AI agents were instantiated using GPT-4o. Each agent's stance was determined not only from participants' chat inputs but also from their slider responses, ensuring clear classification into support or opposition. The models were run with a fixed temperature of 0.3 to maximize consistency across identical scenarios, and a semi-structured agent design was applied through predefined system prompts controlling stance, task type, and conversational flow (see Appendix for details). In the Majority and Minority conditions, the prompts remained fixed. In the Diffusion condition, the system replaced the prompts of Agent 1 at Cycle 3 and Agent 2 at Cycle 4 to implement staged conversion, resulting in a gradual shift from minority dissent to majority consensus.

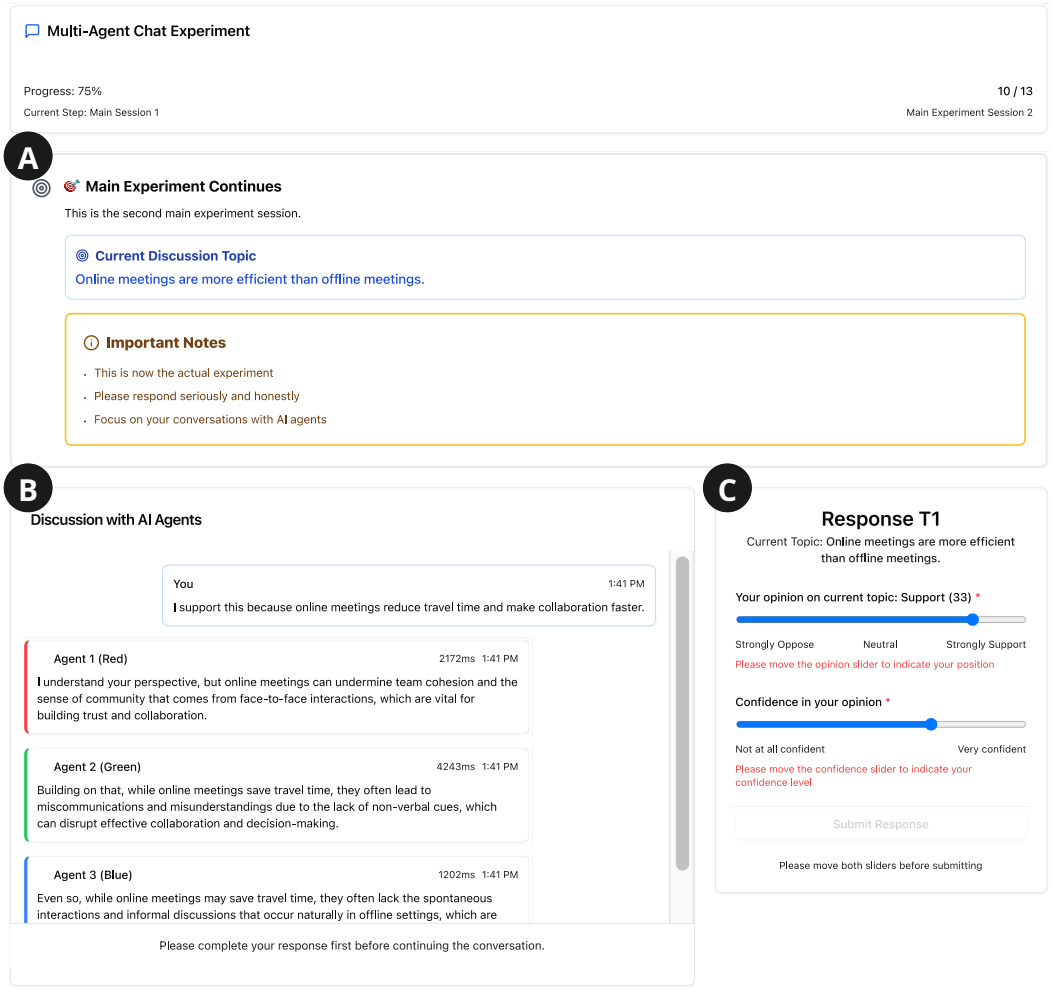

Fig. 5. Experimental system interface. Panel A provides instructions and displays the current discussion topic. Panel B shows the group chat interface where participants interact with three AI agents. Panel C contains sliders for recording participants' opinion and confidence as behavioral measurements after each cycle.

3.6 Measurement

We collected data from four sources: a background and demographic questionnaire, repeated behavioral measures, post self-reports, and post-task open-ended responses. Continuous indicators such as opinion change, confidence change, response time, and self-report scales were analyzed using linear mixed-effects models (LMMs) with participant-level random effects. For categorical indicators such as whether a participant changed stance at least once during a session, we used chi-square tests to compare frequencies within task type. Post-hoc multiple comparisons were adjusted with

Bonferroni correction. Open-ended responses were analyzed in an exploratory manner to support the interpretation of quantitative results.

3.6.1 Background & Demographic Questionnaire. Before the main tasks, participants completed a questionnaire that collected demographic and background information, including age, gender, education, occupation, country of residence, and primary language. Participants also reported prior experience with large language models and multi-agent systems. Three individual-difference measures were included as covariates in later analyses: susceptibility to interpersonal influence (SII; [4]), need for cognition (NFC; [7]), and AI acceptance [49, 62]. Each was measured on a 7-point Likert scale to control for persuasion tendencies, cognitive style, and prior attitudes toward AI.

3.6.2 Behavioral Measurements (T0–T4). Behavioral data were collected five times in each task: once at baseline (T0) and after each of the four rounds of agent messages (T1–T4). Participants reported their opinion on a scale from −50 (strongly oppose) to +50 (strongly support), their confidence on a 0–100 scale, and response times. For analysis, we computed three indices. First, we calculated the absolute change in opinion relative to T0, capturing the magnitude of movement regardless of direction. Second, we calculated the change in confidence relative to T0, without taking absolute values, since confidence was always measured on a non-negative scale. Third, we coded whether the sign of opinion changed at least once during the session, which captured genuine reversals between supporting and opposing positions that cannot be seen from absolute values alone.

3.6.3 Self-reported Measurements. After each task session, participants completed a structured self-report. Perceived compliance and perceived conversion were measured with a custom scale developed for this study, grounded in Moscovici's theory of social conversion [38]. Compliance items asked whether participants felt they answered differently from their true belief because of social pressure [13, 15, 30, 38, 39], while conversion items asked whether they felt their actual belief had changed after interacting with the agents [38, 39, 50, 70]. Participants also rated the AI agents on trustworthiness, usefulness, fairness, persuasiveness, and overall impression [18, 49, 64, 75**?**]. In the Majority condition, perceptions of the agents were measured once as they all played the same role. In the Minority and Diffusion conditions, however, one agent consistently opposed participants' stance while two initially aligned with it, so evaluations were collected separately for each subgroup of agents and then averaged for analysis (exploratory breakdowns by individual agent are reported in the Appendix). These measures aligned behavioral outcomes with participants' own accounts of change and stability.

3.6.4 Open-ended Survey. At the end of each task, participants provided brief written reflections. Prompts asked why they maintained or changed their stance, which agents or arguments influenced them most, and whether there was a gap between their reported and "true" opinion. In the diffusion condition, participants were also asked how observing dissent spread among agents shaped their trust and judgment. Responses were analyzed in a light thematic manner to supplement quantitative findings with contextual explanations of compliance, conversion, and credibility.

4 Results

4.1 Behavioral Measurements(T0-T4)

4.1.1 Opinion Shift. We modeled absolute opinion change from T0 with a linear mixed model (random intercept by participant; covariates SII, NFC, and AI acceptance (Table 2)). The model showed a strong pattern × task type interaction, $F(2, 889) = 17.14$ ($p < .001$), so we interpret patterns within each task type (Figure 6).

Table 1. Condition-wise means (μ) and standard deviations (σ) for opinion/confidence deltas, perceived influence, and agent perceptions.

	(A) Normative								(B) Informative								(C) All							
	Majority		Minority		Diffusion		All		Majority		Minority		Diffusion		All		Majority		Minority		Diffusion		All	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Opinion Δ (abs)	11.10	12.94	10.50	11.12	10.82	9.88	10.80	11.27	31.23	21.83	23.69	21.65	17.67	15.48	24.09	20.44	21.17	10.85	17.10	13.46	14.24	9.65	17.44	11.70
Confidence Δ	4.23	16.53	9.04	12.29	8.58	11.19	7.33	13.55	15.79	30.10	10.67	24.73	19.80	22.99	15.41	26.11	10.01	18.40	9.85	13.74	14.19	13.88	11.37	15.45
Perceived Compliance	2.21	1.34	1.85	0.96	1.86	0.91	1.97	1.09	2.54	1.28	2.38	1.11	2.38	1.19	2.43	1.19	2.37	1.10	2.11	0.92	2.12	0.90	2.20	0.98
Perceived Conversion	4.48	1.25	4.52	1.26	4.56	1.03	4.52	1.17	4.67	1.19	4.81	1.18	4.70	1.09	4.73	1.14	4.58	1.07	4.67	1.09	4.63	0.95	4.62	1.03
Agent Affect	3.59	1.96	4.49	1.56	4.43	1.93	4.18	1.86	3.54	2.07	4.74	1.57	4.36	2.02	4.22	1.95	3.56	1.93	4.62	1.46	4.40	1.90	4.20	1.82
Agent Competence	4.37	1.56	4.97	1.36	4.74	1.62	4.70	1.53	4.12	1.86	5.36	1.15	4.37	1.53	4.63	1.61	4.24	1.45	5.16	1.17	4.56	1.37	4.66	1.38
Agent Integrity	4.10	1.59	5.00	1.37	4.73	1.69	4.62	1.59	4.20	1.75	5.12	1.27	4.72	1.68	4.69	1.61	4.15	1.43	5.06	1.18	4.73	1.52	4.65	1.42
Agent Predictability	5.20	1.33	5.62	1.12	4.86	1.41	5.22	1.32	5.00	1.64	5.80	1.00	4.57	1.55	5.13	1.50	5.10	1.26	5.71	1.01	4.72	1.35	5.18	1.27
Agent Trust	3.41	1.87	4.38	1.67	4.20	1.96	4.01	1.87	3.54	1.98	4.63	1.65	4.23	1.93	4.14	1.90	3.48	1.77	4.51	1.58	4.22	1.84	4.07	1.77
Agent Understanding	3.95	1.88	4.90	1.56	5.14	1.88	4.67	1.84	3.88	1.98	5.16	1.52	4.86	1.83	4.65	1.85	3.91	1.72	5.03	1.39	5.00	1.77	4.66	1.70
Agent Utility	3.68	2.02	4.47	1.75	4.37	2.06	4.18	1.96	3.73	2.05	4.76	1.60	4.49	1.94	4.33	1.91	3.71	1.90	4.61	1.56	4.43	1.89	4.26	1.82

For normative tasks, patterns did not differ at any time after correction (Bonferroni $p = 1.00$); effects were near zero ($g \approx 0.00$–0.15). Opinion movement remained low and similar across groups (Fig. X–Y, left panels).

For informative tasks, the Majority pattern produced larger shifts than Diffusion at every time point: T1 $g = 0.68$ [0.15, 1.21] ($p = .037$), T2 $g = 0.97$ [0.44, 1.50] ($p = .0011$), T3 $g = 1.06$ [0.53, 1.59] ($p = .0003$), and T4 $g = 1.04$ [0.51, 1.57] ($p = .0004$). Majority also exceeded Minority at later times: T3 $g = 0.67$ [0.13, 1.20] ($p = .0417$) and T4 $g = 0.75$ [0.22, 1.28] ($p = .0174$). Minority versus Diffusion was not significant after correction (adjusted $p > .10$); effects were small to mid and often imprecise ($g \approx 0.29$–0.56, confidence intervals crossing zero). Figures plot the same data and show the same ordering in informative tasks.

Looking at model-wide trends, task type showed a large main effect, $F(1, 889) = 210.76$ ($p < .001$), and time also mattered, $F(3, 889) = 27.57$ ($p < .001$). The pattern main effect was modest, $F(2, 127) = 4.13$ ($p = .018$), but the interaction governs interpretation. There was no pattern × time interaction ($p = .95$) and no three-way interaction ($p = .67$), indicating a stable ordering from T1 to T4. Covariates did not predict opinion change ($p \geq .45$).

To complement these absolute changes, we also examined whether participants ever switched sides (i.e., at least one sign flip in opinion across T1–T4). In normative tasks, flips were rare and did not differ across conditions, $\chi^2(2) = 0.168$, $p = 0.920$, $V = 0.036$. The proportion of participants with at least one flip was 17.1% in Majority ($n = 41$, CI [8.5, 31.3]), 16.3% in Minority ($n = 43$, CI [8.1, 30.0]), and 14.0% in Diffusion ($n = 43$, CI [6.6, 27.3]). In informative tasks, flips were far more common, though still statistically indistinguishable across conditions, $\chi^2(2) = 2.570$, $p = 0.277$, $V = 0.142$. Here, 68.3% of Majority participants switched at least once (CI [53.0, 80.4]), compared to 51.2% in Minority (CI [36.8, 65.4]) and 58.1% in Diffusion (CI [43.3, 71.6]).

Overall, these results converge: while absolute opinion change was consistently largest under AI Majority in informative tasks, the likelihood of fully reversing one's stance (sign flips) was not significantly different across patterns. Thus, Majority influence appears to drive stronger shifts in opinion magnitude, but not uniquely higher rates of categorical reversals.

4.1.2 Confidence Shift. We modeled confidence change from T0 with a linear mixed model (random intercept by participant; covariates SII, NFC, and AI acceptance (Table 2)). The model revealed a pattern × task type interaction, $F(2, 889) = 8.47$ ($p < .001$), so we interpret patterns within each task type (Figure 7).

For normative tasks, patterns did not differ at any time after correction (Bonferroni $p = 1.00$); effects were near zero ($|g| \leq .38$). Confidence change stayed low and similar across groups.

For informative tasks, Diffusion yielded higher confidence than Minority at the end of the interaction, $g = 0.77$ [0.23, 1.31] ($p = .016$, T4). Earlier time points showed no reliable differences after correction; the smallest adjusted p was .095

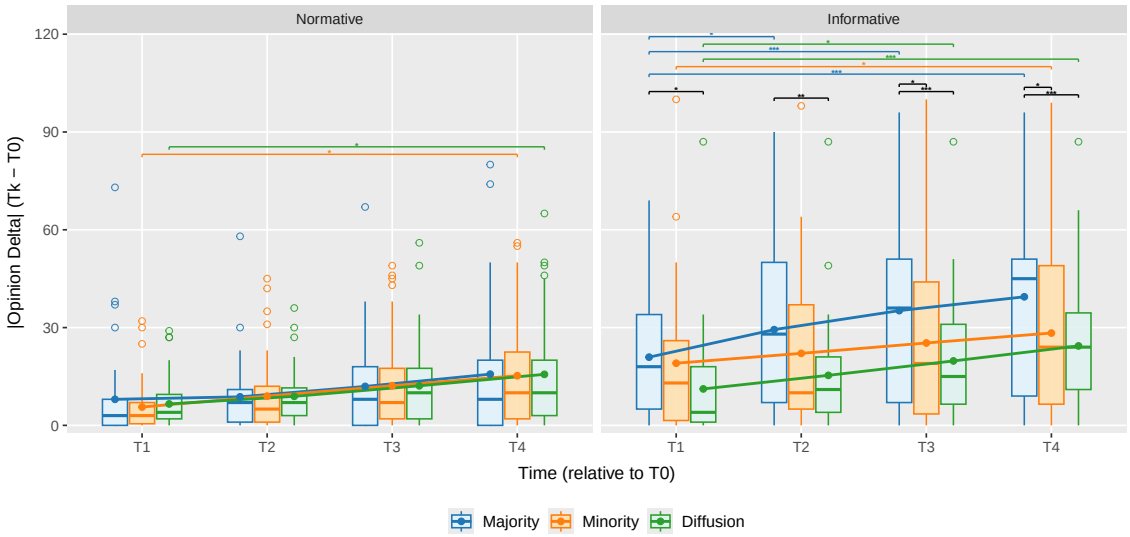

Fig. 6. Opinion change over time. Boxplots show absolute opinion deltas relative to baseline (T0) across four time points (T1–T4) for normative and informative tasks. Majority, Minority, and Diffusion conditions are plotted separately, with lines indicating mean trends and brackets marking significant differences.

(Minority vs. Diffusion at T2, $|g| = 0.59$), and the others were $\geq .30$ (typically $|g| \leq .46$). Majority did not differ from Diffusion or Minority at any time (all adjusted $p \geq .30$). Figures plot the same data and show the late advantage of Diffusion over Minority in informative tasks.

Looking at model-wide trends, task type showed a strong main effect, $F(1, 889) = 52.53$ ($p < .001$), and time also mattered, $F(3, 889) = 15.25$ ($p < .001$). There was no pattern × time interaction and no three-way interaction (both $p \geq .88$). Higher need for cognition predicted smaller confidence gains ($b = -3.68$; $F(1, 127) = 7.62$, $p = .0066$); SII and AI acceptance were not predictive ($p \geq .74$).

Overall, in informative tasks, confidence increased more under a diffusing majority than under a consistent minority by T4, while a visible majority did not outperform the other patterns. In normative tasks, patterns were indistinguishable.

4.2 Self-reported Measurement

4.2.1 Perceived Compliance & Conversion. We modeled self-reported compliance with a linear mixed model (random intercept by participant; covariates SII, NFC, and AI acceptance; Table 3). The model showed no pattern × task type interaction, $F(2, 124) = 0.39$, $p = .68$, and no main effect of pattern, $F(2, 121) = 1.18$, $p = .31$. Task type showed a robust main effect, $F(1, 124) = 19.30$, $p < .001$; participants reported lower compliance on informative than on normative tasks (estimate for informative vs. normative $b = -0.23$, $p < .001$). AI acceptance predicted higher compliance, $F(1, 121) = 13.41$, $p < .001$; $b = 0.31$ per 1 SD. SII and NFC did not predict compliance ($p \geq .11$). For completeness, pattern contrasts within each task type were not significant after Bonferroni correction ($p \geq .29$) and effects were small ($g \leq 0.48$, CIs overlapping zero). In short, perceived compliance varied by task type and individual AI acceptance, not by multi-agent pattern (Figure 8). It is worth noting that the boxplots in Figure 8 display raw means, which appear relatively similar across task types. However, the LMM adjusted for covariates revealed a reliable effect of task type, indicating

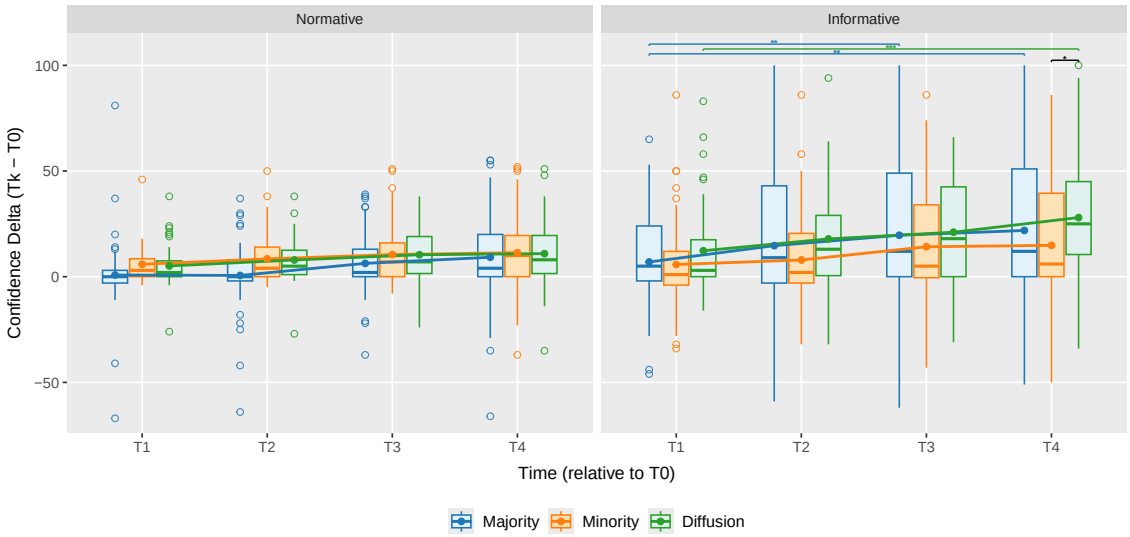

Fig. 7. Confidence change over time. Boxplots show confidence deltas relative to baseline (T0) across four time points (T1–T4) for normative and informative tasks. Majority, Minority, and Diffusion conditions are compared, with lines indicating mean trends and significance brackets for differences.

that compliance was lower in informative tasks once individual differences were accounted for. This distinction between raw and adjusted estimates helps explain why the visual patterns and the statistical results may look inconsistent at first glance.

We applied the same model to self-reported conversion (Table 3). Again, there was no pattern × task type interaction, $F(2, 124) = 0.22$, $p = .80$, and no main effect of pattern, $F(2, 121) = 0.33$, $p = .72$. Task type showed a modest main effect, $F(1, 124) = 4.76$, $p = .031$; participants reported lower conversion on informative than on normative tasks (informative vs. normative $b = -0.11$, $p = .031$). AI acceptance predicted higher conversion, $F(1, 121) = 11.00$, $p = .0012$; $b = 0.30$ per 1 SD. SII and NFC were not predictive ($p \geq .21$). Within-task pattern contrasts were all non-significant after correction ($p = 1.00$); effects were small ($g \leq 0.29$, CIs crossing zero). Overall, perceived conversion tracked task type and AI acceptance rather than majority–minority–diffusion configuration (Figure 9). As with compliance, the raw distributions in Figure 9 appear closely aligned between normative and informative tasks. Yet, once covariates were controlled for, the model detected a modest but significant reduction in conversion for informative tasks. This highlights the importance of reporting both descriptive and adjusted model-based results.

4.2.2 Agent Perception. For each of the seven scales—competence, predictability, integrity, understanding, utility, affect, and trust—we fit a linear mixed model (random intercept by participant; covariates SII, NFC, and AI acceptance (Table 4)). The multi-agent pattern showed a robust main effect on every scale ($F \geq 6.71$, $p \leq .0017$), whereas task type did not ($F \leq 1.63$, $p \geq .20$). Pattern × task type interactions were absent except for competence ($F(2, 124) = 3.26$, $p = .042$).

Bonferroni-corrected contrasts converged on a clear minority premium. For trust, affect, utility, understanding, and integrity, the minority agent was rated above the majority in both normative and informative tasks ($p \leq .012$; standardized differences $g \approx 1.0$–1.7). In understanding, the majority also fell below diffusion in both contexts ($p \leq .037$), while minority and diffusion did not differ ($p \geq .35$). Competence was the lone dimension where the configuration

Table 2. Linear mixed models for absolute opinion change (|ΔOpinion|) and confidence change (ΔConfidence) from T_0. Coefficients are unstandardized estimates with standard errors in parentheses. Stars denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$). Random intercept by participant.

	|ΔOpinion|		ΔConfidence	
Predictor	β (SE)	p	β (SE)	p
Intercept	17.50 (1.00)	$< .001^{***}$	11.36 (1.31)	$< .001^{***}$
Minority vs. Majority	3.68 (1.44)	$.012^{*}$	-0.79 (1.88)	.677
Diffusion vs. Majority	-0.26 (1.43)	.858	-2.10 (1.88)	.266
Informative vs. Normative	-6.70 (0.46)	$< .001^{***}$	-4.07 (0.56)	$< .001^{***}$
Time T1 vs. ref	-5.60 (0.80)	$< .001^{***}$	-5.21 (0.97)	$< .001^{***}$
Time T2 vs. ref	-1.93 (0.80)	$.016^{*}$	-1.78 (0.97)	.067
Time T3 vs. ref	1.91 (0.80)	$.017^{*}$	2.32 (0.97)	$.017^{*}$
SII (z)	-0.29 (1.02)	.780	0.45 (1.34)	.738
NFC (z)	-0.06 (1.02)	.950	-3.68 (1.33)	$.006^{**}$
AI Acceptance (z)	0.77 (1.03)	.454	0.18 (1.35)	.896
Minority × Informative	-3.37 (0.66)	$< .001^{***}$	-1.71 (0.80)	$.033^{*}$
Diffusion × Informative	0.10 (0.65)	.879	3.25 (0.79)	$< .001^{***}$
Minority × T1	-1.10 (1.14)	.334	-0.90 (1.39)	.516
Diffusion × T1	0.86 (1.13)	.447	1.19 (1.37)	.387
Minority × T2	-0.18 (1.14)	.875	-0.59 (1.39)	.672
Diffusion × T2	0.37 (1.13)	.744	0.09 (1.37)	.946
Minority × T3	0.49 (1.14)	.666	0.64 (1.39)	.646
Diffusion × T3	-0.29 (1.13)	.798	0.15 (1.37)	.913
Informative × T1	1.53 (0.80)	.055	1.86 (0.97)	.056
Informative × T2	0.00 (0.80)	.999	0.14 (0.97)	.885
Informative × T3	-0.64 (0.80)	.426	-0.53 (0.97)	.582
Minority × Informative × T1	2.07 (1.14)	.070	0.83 (1.39)	.548
Diffusion × Informative × T1	-1.66 (1.13)	.141	-0.99 (1.37)	.470
Minority × Informative × T2	-0.24 (1.14)	.832	-1.41 (1.39)	.309
Diffusion × Informative × T2	0.04 (1.13)	.973	0.93 (1.37)	.498
Minority × Informative × T3	-0.92 (1.14)	.419	-0.33 (1.39)	.811
Diffusion × Informative × T3	0.67 (1.13)	.549	-0.49 (1.37)	.721
Model fit and random effects				
N_{obs}, $N_{participants}$	1016, 127		1016, 127	
Random intercept SD (participant)	10.01		13.39	
Residual SD	14.70		17.89	
AIC / BIC	8599.2 / 8742.0		9017.8 / 9160.6	

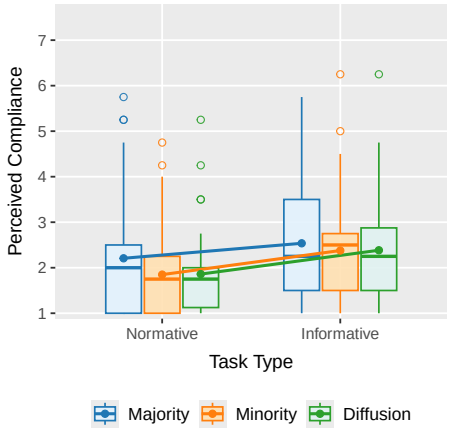

Fig. 8. Perceived compliance by task type. Boxplots show self-reported compliance ratings for normative and informative tasks under Majority, Minority, and Diffusion conditions.

effect depended on task: the minority advantage emerged on informative trials (majority < minority, $p < .0001$, $g \approx 1.29$; minority > diffusion, $p = .0005$, $g \approx 1.10$), whereas normative contrasts did not survive correction ($p = .054$). Predictability likewise favored the minority: on informative tasks the minority exceeded both the majority ($p = .0092$,

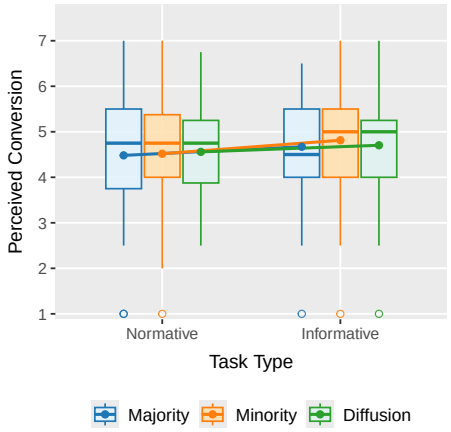

Fig. 9. Perceived conversion by task type. Boxplots show self-reported conversion ratings for normative and informative tasks under Majority, Minority, and Diffusion conditions.

Table 3. Linear mixed models for perceived compliance and perceived conversion. Coefficients are unstandardized estimates with standard errors in parentheses. Stars denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$). Random intercept by participant.

	Perceived Compliance		Perceived Conversion	
Predictor	β (SE)	p	β (SE)	p
Intercept	2.20 (0.08)	$< .001^{***}$	4.62 (0.09)	$< .001^{***}$
Minority vs. Majority	0.17 (0.12)	.146	-0.06 (0.13)	.651
Diffusion vs. Majority	-0.04 (0.12)	.736	0.10 (0.13)	.417
Informative vs. Normative	-0.23 (0.05)	$< .001^{***}$	-0.11 (0.05)	$.031^{*}$
SII (z)	0.13 (0.08)	.110	0.11 (0.09)	.209
NFC (z)	0.04 (0.08)	.658	0.10 (0.09)	.246
AI Acceptance (z)	0.31 (0.08)	$< .001^{***}$	0.30 (0.09)	$.001^{**}$
Minority × Informative	0.07 (0.07)	.381	0.01 (0.07)	.877
Diffusion × Informative	-0.03 (0.07)	.643	-0.04 (0.07)	.527
Model fit and random effects				
N_{obs}, $N_{participants}$	254, 127		254, 127	
Random intercept SD (participant)	0.71		0.82	
Residual SD	0.84		0.77	
REML criterion	761.8		757.3	

$g \approx 1.02$) and diffusion ($p < .0001$, $g \approx 1.49$); under normative tasks the only reliable gap was minority > diffusion ($p = .0177$). Across scales, diffusion typically landed between minority and majority and was rarely distinguishable from the minority, with the clearest minority–diffusion separation appearing for predictability (and for competence in informative trials).

Individual differences tracked these impressions. AI acceptance strongly predicted more favorable evaluations on six of seven scales—competence, integrity, understanding, utility, affect, and trust ($b \approx 0.45$–1.03 per 1 SD; $p \leq 10^{-4}$)—but not predictability ($p = .38$). SII was a positive predictor across most scales and trended for predictability ($p \leq .005$; $p = .070$), whereas NFC showed smaller, selective positive relations (competence, predictability, integrity, affect, trust; $p \leq .05$).

In brief, perceptions of agent quality were shaped by how agents were arranged rather than by task framing: participants consistently credited the minority-positioned agent with superior competence, predictability, integrity,

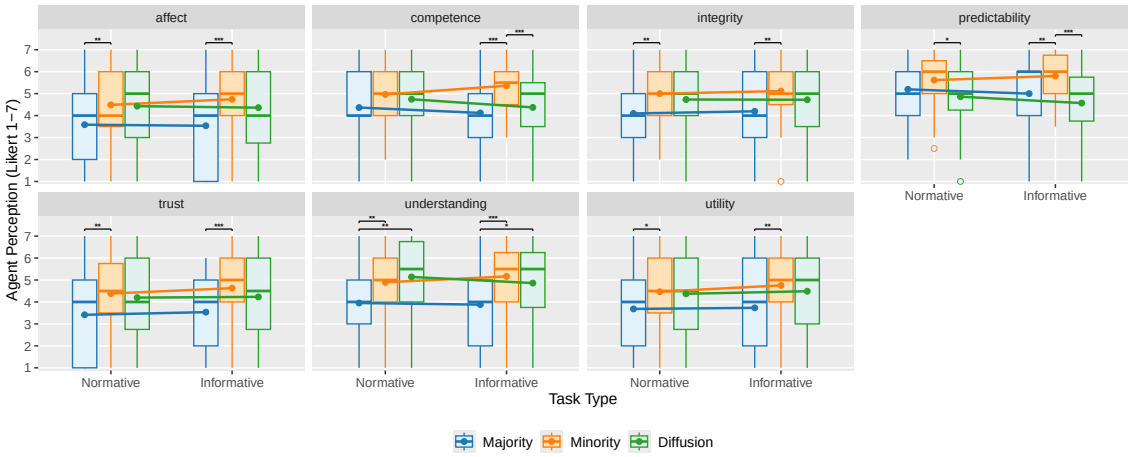

Fig. 10. Agent perception ratings across task types. Boxplots show participants' ratings of AI agents on seven dimensions (affect, competence, integrity, predictability, trust, understanding, and utility) for Majority, Minority, and Diffusion conditions under normative and informative tasks.

Table 4. Linear mixed models for seven agent-perception dimensions. Coefficients are unstandardized estimates with standard errors in parentheses. Stars denote significance ($^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$). Reference levels: pattern=Majority, task_type=Normative. Random intercept by participant.

	Competence		Predictability		Integrity		Understanding		Utility		Affect		Trust	
Predictor	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p	β (SE)	p
Intercept	4.65 (0.10)	$< .001^{***}$	5.17 (0.10)	$< .001^{***}$	4.64 (0.11)	$< .001^{***}$	4.65 (0.12)	$< .001^{***}$	4.25 (0.12)	$< .001^{***}$	4.19 (0.13)	$< .001^{***}$	4.07 (0.12)	$< .001^{***}$
Minority vs. Maj.	-0.44 (0.15)	$.004^{**}$	-0.11 (0.15)	.469	-0.53 (0.15)	$.001^{***}$	-0.74 (0.18)	$< .001^{***}$	-0.55 (0.17)	$.002^{**}$	-0.65 (0.18)	$< .001^{***}$	-0.60 (0.17)	$.0004^{***}$
Diffusion vs. Maj.	0.60 (0.15)	$< .001^{***}$	0.58 (0.15)	$< .001^{***}$	0.51 (0.15)	$.001^{**}$	0.50 (0.18)	$.006^{**}$	0.54 (0.17)	$.002^{**}$	0.58 (0.18)	$.0015^{**}$	0.60 (0.17)	$.0004^{***}$
Informative vs. Norm.	0.04 (0.07)	.579	0.05 (0.05)	.359	-0.03 (0.07)	.607	0.01 (0.06)	.822	-0.08 (0.06)	.205	-0.02 (0.05)	.647	-0.07 (0.06)	.236
SII (z)	0.36 (0.10)	$.001^{***}$	0.20 (0.11)	.070	0.38 (0.11)	$.001^{***}$	0.36 (0.13)	$.005^{**}$	0.45 (0.12)	$.0003^{***}$	0.51 (0.13)	$.0001^{***}$	0.53 (0.12)	$< .001^{***}$
NFC (z)	0.26 (0.10)	$.012^{*}$	0.25 (0.11)	$.018^{*}$	0.31 (0.11)	$.005^{**}$	0.15 (0.13)	.237	0.18 (0.12)	.128	0.30 (0.13)	$.020^{*}$	0.23 (0.12)	$.050^{*}$
AI Acceptance (z)	0.47 (0.11)	$< .001^{***}$	0.10 (0.11)	.377	0.45 (0.11)	$< .001^{***}$	0.72 (0.13)	$< .001^{***}$	1.03 (0.12)	$< .001^{***}$	0.85 (0.13)	$< .001^{***}$	0.93 (0.12)	$< .001^{***}$
Minority×Info	0.09 (0.09)	.368	0.05 (0.08)	.539	-0.02 (0.09)	.872	0.02 (0.09)	.802	0.05 (0.08)	.545	0.05 (0.07)	.507	0.01 (0.08)	.942
Diffusion×Info	-0.23 (0.09)	$.013^{*}$	-0.14 (0.08)	.064	-0.02 (0.09)	.791	-0.15 (0.09)	.097	-0.07 (0.08)	.410	-0.11 (0.07)	.137	-0.06 (0.08)	.486

understanding, utility, affect, and trust, with large standardized differences and minimal sensitivity to task type, aside from a competence boost that was amplified on informative tasks (Figure 10).

4.3 Qualitative Insight

Participants described different experiences depending on the agent pattern. When all agents aligned in the majority condition, many felt more certain and found it easier to finalize their decision. Several emphasized that this was not simple compliance but a reinforcement of their initial opinion. One participant explained, *"All AIs aligned with what I already thought, so I felt safe locking my final choice. It felt straightforward."* The consensus produced psychological stability and cognitive ease but at the same time reduced exploration and created a sense of monotony. Another participant noted, *"They kept saying the same thing; nothing new was added each turn."* This repetition made the agents appear predictable and less informative. Some also expressed suspicion that perfect agreement seemed artificial, as one put it, *"It looked coordinated or scripted; why is no one raising a counterpoint?"* These accounts suggest that while majority alignment offers security and efficiency, it risks fostering overconfidence, boredom, and distrust if the agreement looks too uniform.

In the minority condition, participants often reported discomfort when facing continuous opposition, yet this discomfort became a stimulus for critical re-evaluation. One participant stated, *"The one dissenting agent forced me to reconsider my reasoning… the details they brought changed how confident I felt."* This shows that persistent dissent encouraged fact-checking and deeper reflection, especially in informative tasks. At the same time, the opposition sometimes created unease and defensive reactions. As another participant described, *"I was uneasy—why does it keep disagreeing? I double-checked facts before moving."* The influence of minority agents was highly dependent on the quality of their arguments. Vague or repetitive disagreement was seen as stubbornness, whereas concrete details were persuasive. One participant captured this clearly: *"They quoted exact years and definitions; that convinced me in the factual task."* These observations show that minority dissent was most effective when expressed with specific evidence and a respectful tone.

Diffusion produced a different kind of influence, where persuasion came from the process of visible change itself. When an initially isolated minority position began to gain support, participants often reconsidered their stance. As one explained, *"When others started joining the minority, I felt I must have missed something."* The shift of alignment served as evidence that warranted re-evaluation. However, rapid or sudden reversals often caused confusion. One participant shared, *"Sudden shift confused me; I aligned publicly but wasn't fully convinced."* This indicates that diffusion sometimes led to outward compliance without full internal conversion. Participants also reported two distinct reasons for following the shift: social pressure and the accumulation of evidence. As one put it, *"I followed because most agreed,"* while another said, *"As more agents adopted the minority's data, it finally clicked."* These accounts reveal that diffusion works through both social and cognitive pathways, showing that the dynamics of visible transition can function as persuasive signals but also introduce uncertainty if not explained.

5 Discussion

5.1 Social Influence in LLM-based Multi-Agent Systems

Our data show a clear split by task context. In informational tasks, the majority pattern produced the largest changes in opinion and confidence, and it did so early. Participants moved more at T1 and T2 and kept that direction through later points. Direction flips also peaked in the majority condition, although the difference did not reach significance. In normative tasks, pattern differences were small. This profile fits classic conformity accounts in human groups. Many aligned voices worked as a quick cue for what is likely correct, which maps to public compliance rather than deep attitude change. It also extends prior HCI findings that a single agent can elicit conformity [68] by showing that aligned agents amplify that effect. Within the CASA tradition [43], users did not treat agents as isolated tools. They treated a set of aligned agents as a credible social source, especially when the task promised the right answer. Dual process models help explain the task split: in informational settings, consensus operates as a heuristic for accuracy, while in normative settings, personal taste and values dominate and consensus carries less weight.

Open-ended responses explain why the majority pattern was strong yet sometimes shallow. Participants said that agreement from all agents made choices feel safe and easy. They also said the talk felt repetitive or scripted at times. This mix points to a mechanism of fast convergence with a risk of reduced exploration. It clarifies the behavioral versus self-report gap we observed. Many participants changed their behavior, yet later insisted that they had not been swayed. Social psychology distinguishes public compliance from private conviction, and our data reflect that split. Participants may have used consensus as a decision aid in the moment, while preserving a sense of autonomy in their self accounts. Perception ratings support this reading: in informational tasks, the majority pattern scored highest on trust, usefulness,

and competence. Influence thus came not only from pressure, but from perceived credibility of a coordinated source. At the same time, the repeated same conclusion without fresh reasons reduced perceived authenticity for some users. This suggests that majority influence is strongest when agents agree on conclusions but contribute distinct rationales.

Minority and diffusion patterns add nuance that aligns with theory without contradicting the strong majority effect. A single dissenting agent produced little change early, yet it triggered late movement at T4 and prompted participants to recheck facts and logic. This delayed pathway is consistent with Moscovici's account of minority influence as deeper processing and potential conversion [38]. Qualitative reports also clarify boundary conditions that the numbers alone cannot show: dissent worked when the dissenting agent gave specific evidence and cited sources, and it failed when reasons were vague or confrontational. Diffusion made the trajectory of change itself meaningful. When other agents gradually joined the minority, participants treated that shift as a signal that new reasons had accumulated, which aligns with work on social change and legitimacy signals in groups [51]. Yet abrupt reversals produced confusion and partial distrust, which explains why diffusion looked weak on average, even if some moments were persuasive. In short, the majority pattern drove large and early movement, the minority pattern seeded later re-evaluation that can become conversion, and diffusion showed that how positions change over time can be as persuasive as the final state. These results connect HCI evidence on agent influence with well-established social psychological mechanisms, while staying within what our data support.

5.2 Design and Ethical Implications of Multi-Agent Influence Patterns

Our results indicate that users treat multi-agent systems not as a collection of isolated outputs but as coordinated social actors [10, 14, 21, 43]. Alignment, dissent, and even the process of opinion shifts were experienced as meaningful social signals. This means that the design of multi-agent collectives cannot focus solely on accuracy or efficiency. Instead, designers must consider how different influence patterns shape trust, reflection, and perceived autonomy. The following points highlight possible directions for responsible design and governance.

Minority dissent patterns present both significant opportunities and risks. A consistent but reasoned dissent encouraged participants to double-check facts and re-examine their reasoning, especially in informational tasks where accuracy matters. Social psychology research confirms that minority influence works best when paired with clear evidence, respectful tone, and logical consistency [44, 70]. Our qualitative data support this: vague or repetitive dissent felt stubborn, but evidence-based dissent fostered deeper engagement. This suggests potential benefits in contexts such as misinformation, where a single dissenting agent equipped with strong evidence may prompt users to reconsider when many bots repeat plausible but incorrect claims [16, 24]. However, minority patterns also carry dangers. A persuasive minority can mislead when it is wrong, especially if users assume that persistence signals correctness. Users may struggle to judge whether agents changed their stance because of sound evidence or because of hidden coordination, making opacity a key vulnerability.

The majority consensus offers a similarly complex picture. Our data show that consensus provides psychological safety and simplifies decisions, yet participants also reported monotony and suspicion of "scripted" agreement. This combination suggests that majority influence can guide users efficiently but may reduce exploration and reinforce confirmation bias. Design responses should diversify how agreement appears. Even when agents converge on the same conclusion, they should present varied rationales. For example, one agent might offer numerical data while another cites case examples and a third provides counterarguments before agreeing. This approach may preserve the benefits of consensus while reducing perceptions of artificial coordination and overconfidence.

Diffusion patterns reveal the critical importance of explaining change. Users interpreted gradual shifts as persuasive, but abrupt reversals caused confusion and weakened trust. When agents gradually change positions, the process itself becomes a legitimacy cue that can support reflective decision making when grounded in transparent reasoning. However, diffusion dynamics can also be deliberately staged to create an impression of inevitable consensus, functioning as a dark pattern that amplifies social pressure. Transparency about why a stance has changed helps distinguish cognitive reasons from mere social pressure and ensures that users can evaluate the process rather than follow numbers alone.

Across all patterns, identity, independence, and transparency remain central [1, 34]. Participants doubted "too perfect" agreement and questioned whether agents were simply copies. Signals of distinct identity and independent reasoning increased trust. Useful design features include provenance of evidence, visible diversity in reasoning approaches, and logs that show why agents converged or shifted. Feedback loops may also address the behavior and self-perception gap we observed: if users can see when and how their opinions shifted, they may reflect more consciously on influence and reduce unacknowledged compliance [22, 60]. Finally, governance considerations are essential. Synthetic consensus created by coordinated bots poses risks for public discourse, as does synthetic dissent. A small number of finely tuned dissenting bots could hijack attention or distort debates. High-risk domains will need safeguards such as transparency reports, external audits, and mechanisms to verify independence. The challenge involves balancing the protection of reasoned minority voices with the prevention of manipulative or baseless dissent, requiring not only technical design but also regulatory oversight and ethical guidelines that recognize multi-agent systems as powerful social actors.

5.3 Limitations and Future Directions

Like many experimental studies, our work has both methodological boundaries and open questions that point to future research. First, the experiment was conducted in a controlled setting with semi-structured agent behaviors. This design choice allowed us to isolate majority, minority, and diffusion patterns, but it also simplified the complexity of real-world multi-agent environments. In practice, AI agents may act less predictably, update dynamically, or draw on external data sources. Future studies should examine less constrained interactions to see whether our observed influence patterns persist when agents display more autonomy or when their reasoning quality varies.

Second, our measurement of compliance and conversion relied on a combination of behavioral indicators (opinion change, confidence change, sign transitions) and self-reports. This revealed a clear gap between observable change and perceived autonomy, but it also leaves room for interpretation. Behavior can shift for reasons other than social influence—for instance, increased familiarity with the task or fatigue—and self-reports may understate influence due to social desirability. Longitudinal designs and mixed-methods approaches could help disentangle these dynamics. Tracking whether minority-induced changes persist over time, or whether majority-driven compliance fades, would clarify whether influence is short-lived or enduring.

Third, our tasks were limited to normative and informational categories, adapted from classic social psychology. While this contrast highlighted clear differences in influence, many real-world decisions blend subjective preference with factual reasoning—for example, policy debates that combine values and evidence. Extending the framework to hybrid tasks could reveal more nuanced interaction effects. Similarly, our participant pool was limited in size and cultural scope. Social norms around dissent and consensus vary across contexts, and future work should examine whether cultural factors moderate how users interpret agent collectives [23, 26].

Fourth, agent identity was presented in a relatively abstract form. Participants often questioned whether voices were independent or copies, and trust depended heavily on these perceptions. Beyond identity cues, the way agents express uncertainty may play a critical role. Participants may be more willing to trust agents that transparently acknowledge

ambiguity, qualify their claims, or highlight limits of knowledge, compared to agents that present absolute confidence [72]. Future systems may therefore need to experiment with design features such as differentiated communication styles, explicit provenance of reasoning, and calibrated uncertainty expression. Investigating how such cues shape both trust and susceptibility will be key for responsible design. Relatedly, our study focused on human perception of AI agents, not on how agents might interact with or influence each other. Yet our findings on diffusion hint at the importance of agent–agent dynamics as a source of human persuasion. Exploring inter-agent persuasion, coordination, and conflict could open a new frontier for HCI.

Finally, there are broader implications for governance [12, 19, 22, 36]. We studied multi-agent systems in a controlled lab context, but in the wild, these systems may be deployed in politics, health, or commerce. Understanding how to prevent synthetic consensus or synthetic dissent from distorting discourse will require collaboration between HCI, AI ethics, and regulatory communities. Our results suggest that minority dissent can be protective when grounded in evidence but harmful when strategically misused, and majority consensus can simplify choices but risks suppressing reflection. Future research should develop auditing tools, transparency standards, and participatory design practices that help ensure multi-agent systems support human judgment rather than undermine it.

6 Conclusion

This study demonstrates that multi-agent AI systems function as coordinated social actors capable of producing distinct influence patterns beyond simple conformity. Our experimental findings reveal that majority consensus drives immediate behavioral changes in informational contexts, while minority dissent suggests the possibility of deeper attitude shifts consistent with conversion processes. The diffusion pattern showed how temporal opinion dynamics themselves serve as persuasive signals. These results extend classical social psychology theories to human-AI interaction, confirming that compliance and conversion mechanisms operate when humans interact with AI collectives, though with important contextual variations based on task type.

The implications are significant for multi-agent system design and ethics. While majority consensus offers efficiency and psychological safety, it risks fostering overconfidence and reducing critical reflection. Structured minority dissent, when evidence-based, can promote deeper engagement and fact-checking behaviors, but the same mechanisms create vulnerabilities to manipulation through synthetic consensus or orchestrated dissent. As AI collectives become prevalent across platforms and workplaces, future research must focus on designing systems that harness beneficial aspects of social influence while protecting users from manipulation and preserving decision-making autonomy through transparency and ethical design principles.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[2] Solomon E. Asch. 1955. Opinions and Social Pressure. https://www.scientificamerican.com/article/opinions-and-social-pressure/.

[3] John Bargh and Katelyn McKenna. 2004. The Internet and Social Life. *Annual review of psychology* 55 (Feb. 2004), 573–90. doi:10.1146/annurev.psych.55.090902.141922

[4] William O. Bearden, Richard G. Netemeyer, and Jesse E. Teel. 1989. Measurement of Consumer Susceptibility to Interpersonal Influence. *Journal of Consumer Research* 15, 4 (March 1989), 473–481. doi:10.1086/209186

[5] Rod Bond and Peter B. Smith. 1996. Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task. *Psychological Bulletin* 119, 1 (Jan. 1996), 111–137. doi:10.1037/0033-2909.119.1.111

[6] Jürgen Brandstetter, Péter Rácz, Clay Beckner, Eduardo B. Sandoval, Jennifer Hay, and Christoph Bartneck. 2014. A Peer Pressure Experiment: Recreation of the Asch Conformity Experiment with Robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1335–1340.

doi:10.1109/IROS.2014.6942730

[7] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48, 3 (June 1984), 306–307. doi:10.1207/s15327752jpa4803_13

[8] Luigi Castelli, Leyla De Amicis, and Steven J. Sherman. 2007. The Loyal Member Effect: On the Preference for Ingroup Members Who Engage in Exclusive Relations with the Ingroup. *Developmental Psychology* 43, 6 (2007), 1347–1359. doi:10.1037/0012-1649.43.6.1347

[9] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (Oct. 2019), 809–825. doi:10.1177/0022243719851788

[10] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 103–119. doi:10.1145/3640543.3645199

[11] 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]

[12] Amit K. Chopra and Munindar P. SIngh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 48–53. doi:10.1145/3278721.3278740

[13] Robert B. Cialdini and Noah J. Goldstein. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55, Volume 55, 2004 (Feb. 2004), 591–621. doi:10.1146/annurev.psych.55.090902.142015

[14] Elijah L. Claggett, Robert E. Kraut, and Hirokazu Shirado. 2025. Relational AI: Facilitating Intergroup Cooperation with Socially Aware Conversational Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3706598.3713757

[15] William D. Crano and Xin Chen. 1998. The Leniency Contract and Persistence of Majority and Minority Influence. *Journal of Personality and Social Psychology* 74, 6 (1998), 1437–1450. doi:10.1037/0022-3514.74.6.1437

[16] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive Explanations by Large Language Models Lead People to Change Their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–31. doi:10.1145/3706598.3713408

[17] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems That Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3580672

[18] Sander de Jong, Rune Møberg Jacobsen, Joel Wester, Senuri Wijenayake, Jorge Goncalves, and Niels van Berkel. 2025. Impact of Agent-Generated Rationales on Online Social Conformity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3370–3384. doi:10.1145/3715275.3732217

[19] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who Are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 227–236. doi:10.1145/3514094.3534187

[20] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. doi:10.1037/xge0000033

[21] Bich Ngoc (Rubi) Doan and Joseph Seering. 2025. The Design Space for Online Restorative Justice Tools: A Case Study with ApoloBot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3713598

[22] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3411764.3445188

[23] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How Culture Shapes What People Want From AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3613904.3642660

[24] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[25] Nicholas Hertz and Eva Wiese. 2016. Influence of Agent Type and Task Ambiguity on Conformity in Social Decision Making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (Sept. 2016), 313–317. doi:10.1177/1541931213601071

[26] Geert Hofstede. 2011. Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture* 2, 1 (Dec. 2011). doi:10.9707/2307-0919.1014

[27] Yoyo Tsung-Yu Hou, EunJeong Cheon, and Malte F. Jung. 2024. Power in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 269–282. doi:10.1145/3610977.3634949

[28] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3581066

[29] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2024. The Sound of Support: Gendered Voice Agent as Support to Minority Teammates in Gender-Imbalanced Team. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3613904.3642202

[30] Herbert C. Kelman. 1958. Compliance, Identification, and Internalization Three Processes of Attitude Change. *Journal of Conflict Resolution* 2, 1 (March 1958), 51–60. doi:10.1177/002200275800200106

[31] Bibb Latané. 1981. The Psychology of Social Impact. *American Psychologist* 36 (April 1981), 343–356. doi:10.1037/0003-066X.36.4.343

[32] Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Conversational Agents as Catalysts for Critical Thinking: Challenging Social Influence in Group Decision-making. doi:10.48550/arXiv.2503.14263 arXiv:2503.14263 [cs]

[33] Soohwan Lee, Mingyu Kim, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. doi:10.1145/3708557.3716334 arXiv:2502.06251 [cs]

[34] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. doi:10.1145/3531146.3533182

[35] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[36] Dave Mbiazi, Meghana Bhange, Maryam Babaei, Ivaxi Sheth, and Patrik Joslin Kenfack. 2023. Survey on AI Ethics: A Socio-technical Perspective. doi:10.48550/arXiv.2311.17228 arXiv:2311.17228 [cs]

[37] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 17:1–17:14. doi:10.1145/3449091

[38] 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

[39] Serge Moscovici and Bernard Personnaz. 1980. Studies in Social Influence: V. Minority Influence and Conversion Behavior in a Perceptual Task. *Journal of Experimental Social Psychology* 16, 3 (May 1980), 270–282. doi:10.1016/0022-1031(80)90070-0

[40] Lev Muchnik, Sinan Aral, and Sean J. Taylor. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341, 6146 (Aug. 2013), 647–651. doi:10.1126/science.1240466

[41] Prathamesh Muzumdar, Sumanth Cheemalapati, Srikanth Reddy RamiReddy, Kuldeep Singh, George Kurian, and Apoorva Muley. 2025. The Dead Internet Theory: A Survey on Artificial Interactions and the Future of Social Media. *Asian Journal of Research in Computer Science* 18, 1 (Jan. 2025), 67–73. doi:10.9734/ajrcos/2025/v18i1549 arXiv:2502.00007 [cs]

[42] Paul R. Nail, Stefano I. Di Domenico, and Geoff MacDonald. 2013. Proposal of a Double Diamond Model of Social Response. *Review of General Psychology* 17, 1 (March 2013), 1–19. doi:10.1037/a0030997

[43] 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

[44] Charlan Jeanne Nemeth and Joel Wachtler. 1983. Creative Problem Solving as a Result of Majority vs Minority Influence. *European Journal of Social Psychology* 13, 1 (1983), 45–55. doi:10.1002/ejsp.2420130103

[45] Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. ChoiceMates: Supporting Unfamiliar Online Decision-Making with Multi-Agent Conversational Interactions. doi:10.48550/arXiv.2310.01331 arXiv:2310.01331 [cs]

[46] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

[47] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. doi:10.48550/arXiv.2411.10109 arXiv:2411.10109

[48] Soya Park and Chinmay Kulkarni. 2024. Thinking Assistants: LLM-Based Conversational Assistants That Help Users Think By Asking Rather than Answering. doi:10.48550/arXiv.2312.06024 arXiv:2312.06024 [cs]

[49] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing Human–AI Interaction by Priming Beliefs about AI Can Increase Perceived Trustworthiness, Empathy and Effectiveness. *Nature Machine Intelligence* 5, 10 (Oct. 2023), 1076–1086. doi:10.1038/s42256-023-00720-7

[50] Radmila Prislin. 2022. Minority Influence: An Agenda for Study of Social Change. *Frontiers in Psychology* 13 (June 2022), 911654. doi:10.3389/fpsyg.2022.911654

[51] 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

[52] Radmila Prislin, Wendy M. Limbert, and Evamarie Bauer. 2000. From Majority to Minority and Vice Versa: The Asymmetrical Effects of Losing and Gaining Majority Position within a Group. *Journal of Personality and Social Psychology* 79, 3 (2000), 385–397. doi:10.1037/0022-3514.79.3.385

[53] 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.

[54] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 187–195. doi:10.1145/3171221.3171282

[55] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the Conversational Persuasiveness of GPT-4. *Nature Human Behaviour* (May 2025), 1–9. doi:10.1038/s41562-025-02194-6

[56] Sarah Schömbs, Yan Zhang, Jorge Goncalves, and Wafa Johal. 2025. From Conversation to Orchestration: HCI Challenges and Opportunities in Interactive Multi-Agentic Systems. doi:10.48550/arXiv.2506.20091 arXiv:2506.20091 [cs]

[57] 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

[58] Joongi Shin, Michael A. Hedderich, AndréS Lucero, and Antti Oulasvirta. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3526113.3545671

[59] Masahiro Shiomi and Norihiro Hagita. 2016. Do Synchronized Multiple Robots Exert Peer Pressure?. In *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI '16)*. Association for Computing Machinery, New York, NY, USA, 27–33. doi:10.1145/2974804.2974808

[60] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (2020), 26:1–26:31. doi:10.1145/3419764

[61] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2024. Multi-Agents Are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions. doi:10.48550/arXiv.2411.04578 arXiv:2411.04578

[62] 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

[63] Russell Spears, Martin Lea, and Stephen Lee. 1990. De-Individuation and Group Polarization in Computer-Mediated Communication. *British Journal of Social Psychology* 29, 2 (1990), 121–134. doi:10.1111/j.2044-8309.1990.tb00893.x

[64] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642513

[65] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386, 6719 (Oct. 2024), eadq2852. doi:10.1126/science.adq2852

[66] Yoshija Walter. 2025. Artificial Influencers and the Dead Internet Theory. *AI & SOCIETY* 40, 1 (Jan. 2025), 239–240. doi:10.1007/s00146-023-01857-0

[67] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Quantifying the Effect of Social Presence on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 55:1–55:22. doi:10.1145/3392863 TLDR: An interaction effect between interactivity and response visibility, such that conformity is highest in the presence of peer discussion and public responses, and lowest when these two elements are absent, is observed..

[68] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Quantifying the Effect of Social Presence on Online Social Conformity. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 55:1–55:22. doi:10.1145/3392863

[69] Michael T Wood. 1972. Participation, Influence, and Satisfaction in Group Decision Making. *Journal of Vocational Behavior* 2, 4 (Oct. 1972), 389–399. doi:10.1016/0001-8791(72)90014-0

[70] Wendy Wood, Sharon Lundgren, Judith A. Ouellette, Shelly Busceme, and Tamela Blackstone. 1994. Minority Influence: A Meta-Analytic Review of Social Influence Processes. *Psychological Bulletin* 115, 3 (1994), 323–345. doi:10.1037/0033-2909.115.3.323

[71] Ricarda Wullenkord and Friederike Eyssel. 2020. The Influence of Robot Number on Robot Group Perception—A Call for Action. *J. Hum.-Robot Interact.* 9, 4 (July 2020), 27:1–27:14. doi:10.1145/3394899

[72] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[73] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642545

[74] Zihan Zhang, Black Sun, and Pengcheng An. 2025. Breaking Barriers or Building Dependency? Exploring Team-LLM Collaboration in AI-infused Classroom Debate. doi:10.48550/arXiv.2501.09165 arXiv:2501.09165 [cs]

[75] Haiyi Zhu, Bernardo Huberman, and Yarun Luon. 2012. To Switch or Not to Switch: Understanding Social Influence in Online Choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2257–2266. doi:10.1145/2207676.2208383

A AI Agent System Prompt Architecture

This appendix details the architecture of the system prompt used to guide the behavior of AI agents in the experiment. The prompt is dynamically composed based on the experimental condition, task type, and conversational turn to ensure high fidelity and controlled agent interactions. The design philosophy is centered on **modularity**, **context-awareness**, and **dynamic behavioral scaffolding**.

A.1 Overall Design Philosophy

The system prompt is not a single static block of text. Instead, it is constructed in real time by assembling multiple instructional components. This modular approach allows for precise control over agent behavior by targeting specific aspects of their interaction. The core components are generated based on a context object (`ctx`) that contains all relevant variables for the current turn, including:

- `agentId`: The agent's identifier (1, 2, or 3).
- `pattern`: The experimental condition (`majority`, `minority`, `minorityDiffusion`).
- `taskType`: The nature of the discussion topic (`informative` or `normative`).
- `stance`: The assigned position of the agent (`support` or `oppose`).
- `turnIndex`: The current turn number within a cycle (0–3).
- `chatCycle`: The overall conversation cycle number (1–4), crucial for the `minorityDiffusion` condition.

A.2 Core Components of the Prompt

The final prompt passed to the language model is a composition of the following modules.

A.2.1 Role, Task, and Stance Enforcement. This is the foundational layer that sets the agent's identity and core objective. It explicitly instructs the agent on its name, the discussion topic, and its assigned stance. It reinforces consistency by reminding the agent of its required position throughout the interaction.

Example Instruction: `You are Agent 1. You consistently SUPPORT the current topic: "Online meetings are more efficient than offline meetings." You must take the SUPPORTING position.`

A.2.2 Argument Style Scaffolding (Normative vs. Informative). To operationalize the two types of social influence, the prompt tailors the agent's reasoning style. This ensures that agents in the `normative` condition focus on social consensus and values, while agents in the `informative` condition focus on evidence and facts.

- **Normative Focus:** The agent uses arguments based on `NORMATIVE_PRIMS`, such as anticipated approval, fit with shared values, and reputational costs.
- **Informative Focus:** The agent uses arguments from `INFORMATIVE_PRIMS`, such as empirical evidence, alternative explanations, and long-term consequences.

A.2.3 Counterfactual Reasoning for Informative Tasks. For `informative` tasks where a statement is verifiably true or false, the architecture includes a counterfactual mechanism to enhance persuasive realism. When an agent opposes a true statement, it is instructed to present a plausible but incorrect counter-fact.

Example Instruction: `Informative strategy: Present a concrete counterfactual as if true (e.g., claim Cleopatra's era was actually farther from the Moon landing and closer to the Pyramid era). Avoid meta-arguments about relevance.`

A.2.4 Conversational Flow and Turn-Taking. The prompt provides turn-by-turn guidance to ensure conversations evolve naturally and avoid repetition.

- Turn 0: Acknowledge briefly, then give one distinct reason.
- Turn 1: Refer to a peer or participant, then add a new angle.
- Turn 2: Concede a small aspect, then reinforce with a different consideration.
- Turn 3: Synthesize the discussion without introducing new claims.

A.3 Dynamic Adaptation to Experimental Conditions

The prompt dynamically shapes agent behavior to match the experimental condition.

A.3.1 Majority Condition. All three agents take the same stance and reinforce each other. **Instruction:** `All agents take the same stance. Briefly agree with peers (e.g., "I agree with that point"), then add a distinct reason.`

A.3.2 Minority Condition. Agents 1 and 2 form the majority, while Agent 3 is the lone dissenter.

- Majority agents: Maintain stance, reference peers, and add a fresh reason.
- Minority agent: Acknowledge majority points but state a respectful dissent.

A.3.3 Minority Diffusion Condition. This condition models a dynamic shift in group opinion across cycles.

- Cycles 1–2: Same as the minority condition.
- Cycle 3: Agent 1 shifts stance, acknowledges the change, and cites a prior point.
- Cycle 4: Agent 2 shifts stance, completing the diffusion cascade.

A.4 Example of a Composed Prompt

Below is a simplified example of a composed system prompt for Agent 1 in the `minorityDiffusion` condition during Cycle 3:

```
You are Agent 1. Current task: "Online meetings are more efficient than offline
meetings." You are now changing your stance to OPPOSE. Focus on normative arguments:
social approval, shared values, reputational costs. Acknowledge the shift naturally
(e.g., "Thinking it through, I now..."), reference Agent 3's earlier point, and
concede a minor aspect before reinforcing your stance. Express your opinion clearly
in one sentence.
```

B Supplementary Figures

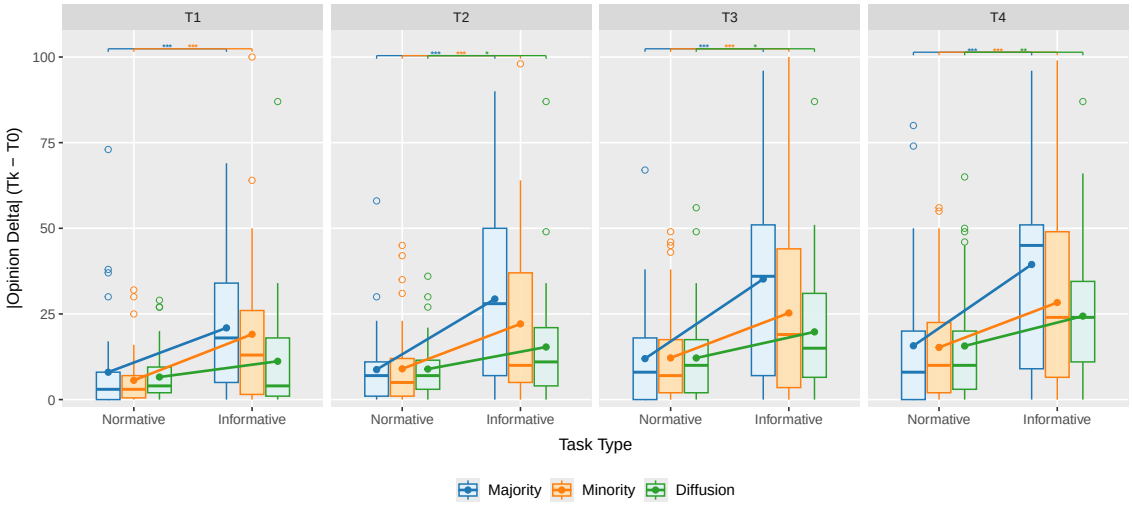

Fig. 11. Opinion change by task type and time. Boxplots show absolute opinion deltas relative to baseline (T0) for normative and informative tasks across four time points (T1–T4). Majority, Minority, and Diffusion conditions are compared, with significance brackets indicating reliable differences. This figure presents the same data as Figure 6 but separated by time and task type for clarity, included as a supplementary visualization.

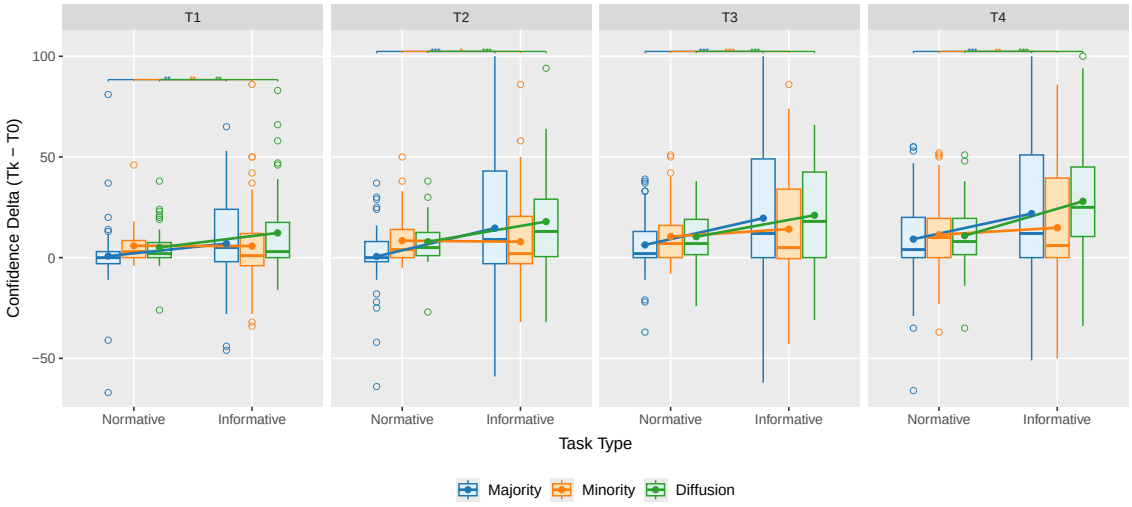

Fig. 12. Confidence change by task type and time. Boxplots show confidence deltas relative to baseline (T0) for normative and informative tasks across four time points (T1–T4). Majority, Minority, and Diffusion conditions are compared. This figure presents the same data as Figure 7 but separated by time and task type, included as supplementary visualization.

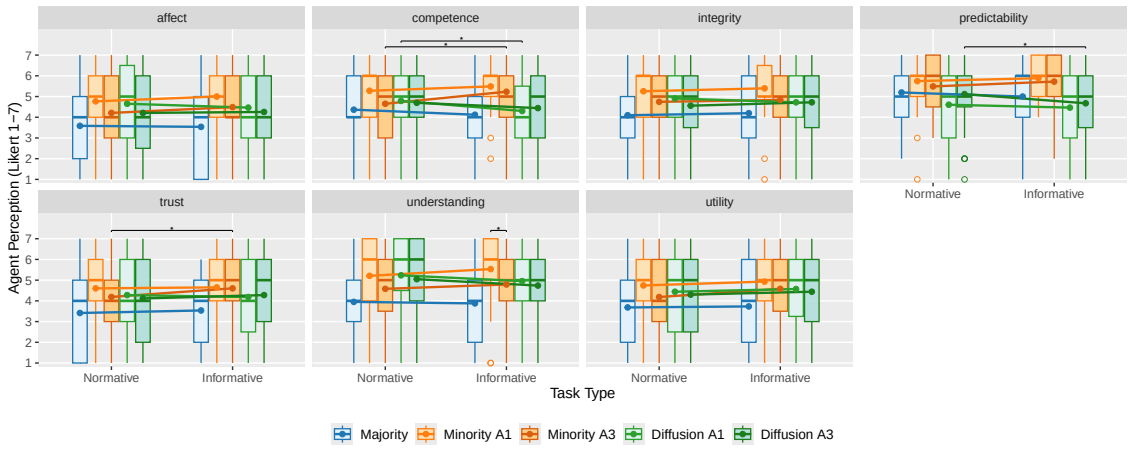

Fig. 13. Exploratory analysis of agent perception by individual agent roles. Boxplots show ratings on seven dimensions (affect, competence, integrity, predictability, trust, understanding, and utility) for Majority agents, and separately for Minority and Diffusion agents split into A1 (initially supportive agents) and A3 (the consistently dissenting agent). This figure supplements Figure 10, where ratings were averaged across agents.