Understanding Compliance and Conversion Dynamics in Multi-Agent Collectives

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

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

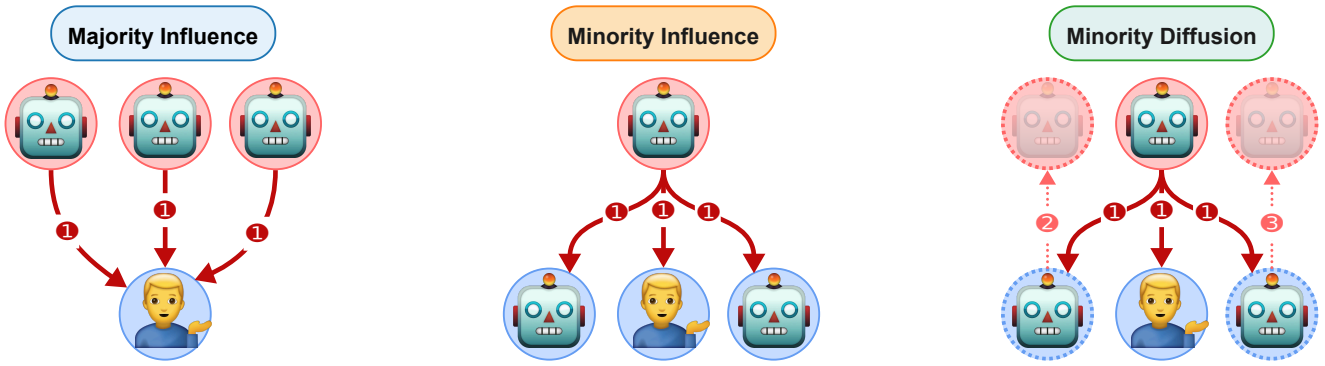


Figure 1: Illustration of the three experimental multi-agent influence conditions. In the Majority Influence condition, all three AI agents opposed the participant's stance, creating strong consensus pressure. In the Minority Influence condition, one dissenting agent consistently opposed the participant 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.

Abstract

Multi-agent AI systems are increasingly prevalent across digital environments, yet their social influence dynamics remain underexplored beyond basic compliance. 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 informative tasks while reporting stance and confidence at five time points. Results demonstrate distinct influence patterns by condition and task type. In informative tasks, majority consensus drove largest immediate opinion changes, while minority dissent showed potential for delayed but deeper attitude shifts consistent with conversion-like processes. The diffusion condition 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.

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

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790385

CCS Concepts

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

Keywords

persuasive technology, conversational agent, multi-agent, social psychology, minority influence

ACM Reference Format:
Soohwan Lee and Kyungho Lee. 2026. Understanding Compliance and Conversion Dynamics in Multi-Agent Collectives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3772318.3790385

1 Introduction

Imagine a dystopian world where much of the online discourse you encounter is not produced by humans, but by automated agents designed to sound convincingly human. While this framing is speculative, it reflects a growing scholarly concern: as large language models (LLMs) become increasingly capable of generating fluent, contextually appropriate dialogue, the boundary between human and machine participation in online spaces becomes harder to discern. LLM-powered chatbots now could be able to post comments on social media platforms, engage in online forums, and in some instances may shape public discourse—an issue underscored by

reports of Reddit communities influenced by automated accounts [24, 30, 59].This is especially concerning because repeated exposure to such seemingly organic, machine-generated content can gradually influence people's perceptions, beliefs, and even behaviors over time. Beyond online text interactions, multi-agent systems are increasingly appearing in professional and recreational settings: workers may collaborate with AI "teammates" on projects [65], and game simulations are populated with non-player characters powered by LLMs that exhibit diverse behaviors [52]. 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 [69, 70]. On the other hand, well-designed multi-agent systems could expose users to diverse perspectives, thereby fostering better decision-making and learning [51, 79]. 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 [49, 60]. 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 [74], persuade users to change attitudes or behaviors [63], encourage critical reflection [17, 54, 72], shape power dynamics within groups [32, 33], support marginalized or minority members [35, 38, 39], and facilitate consensus-building [66, 73]. 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 [51, 79], or simulate complex environments [53]. 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 [69, 70]. In addition, several systems-oriented studies have manipulated whether the system-generated peers placed the user in a majority or minority position and examined how such majority–minority configurations shape users' subsequent choices [74]. 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 [44, 45]. 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 [56–58]. 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 reflective, inclusive interactions.

Building on these gaps, we ask three research questions:

- RQ1 How do different multi-agent influence conditions(majority influence, minority influence, and diffusion by minority influence) affect users' behavior?
- RQ2 How do these influence conditions evolve over time as minority views persist or expand?
- RQ3 How does task type(normative and informative) 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 conditions, 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 informative tasks (evidence-based, correct answers). After completing each task, they measured perceived compliance, conversion(reflective re-evaluation), and perceptions of the agents. This design enabled us to examine how different multi-agent influence conditions unfold over time and across task types.

Results showed a clear task split. In informative tasks, the majority condition drove the largest and earliest absolute changes

in opinion and confidence from baseline, consistent with classic conformity. In normative tasks, influence condition effects were minimal, as personal values dominated. Minority dissent, while weaker early on, showed potential for late-stage re-evaluation consistent with conversion-like processes for a subset of users. Diffusion revealed that temporal dynamics—watching dissent gradually spread—served as a cue for change, 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 conditions produce distinct patterns in behavioral indicators of compliance and conversion (opinion and confidence change trajectories), extending social influence research to AI collectives. Second, it provides a theoretical extension by incorporating temporal dynamics and minority diffusion into Moscovici's framework, demonstrating how conversion processes emerge in human–AI interactions. 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 [48]. This influence typically flows from the majority to the individual, as group consensus pressures dissenting members to conform; yet, it can also flow from the individual to the 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 suggests that the strength of influence depends on the number, immediacy, and intensity of sources [37]. In contrast, Moscovici argued that consistent minorities trigger validation processes, prompting deeper consideration that leads to delayed but enduring attitude conversion [45]. Meta-analyses confirm that consistency, self-sacrifice, and identity relevance enable successful minority influence [75], 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 [71]. Studies reveal how early ratings bias subsequent choices [47, 61], exposure to others' preferences shifts individual selections over time [80], and interactive features with visible responses amplify conformity effects [74]. Social viewing contexts further strengthen community alignment with majority positions [42]. Importantly, temporal dynamics matter: shifts between majority and minority positions create asymmetric effects that reshape group identity and influence patterns [57]. 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 [49, 60]. 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, 32, 33]. Perceived expertise plays a central role, as labels and framing trigger heuristics that make users more likely to conform [33, 34]. Task properties also matter, since people depend more on an agent under high ambiguity, while human-likeness shows inconsistent effects [29]. Message design further amplifies influence. Rationales provided by LLMs increase conformity, especially in informative tasks where ground truth exists [18]. GPT-4 demonstrates persuasion equal to or stronger than humans, and personalization strengthens the effect [63]. Conversational explanations improve comprehension and trust but also risk overreliance [28]. 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, 72].

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, 21, 41]. 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, 28]. 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 [20, 25, 51, 79]. 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 [79]. Other work created platforms that let users orchestrate several agents to make unfamiliar online choices [51]. Parallel research developed agents that simulate social behavior through shared memory and interaction [52]. Additional studies show that users become orchestrators who must manage transparency, conflicts, and trust in multi-agent environments [64]. 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 [69, 70]. 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, 27]. 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 [62, 67]. People tend to conform when they trust the robots, and they need at least three robots to see them as a group [6, 77]. 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. This study received exempt status from the Institutional Review Board (IRB) (UNISTIRB-25-062-C).

3.2 Experimental Design & Conditions

We used a split–plot mixed design with three between-subject conditions (majority, minority, diffusion) and two within-subject task types (normative, informative). Each participant completed both a Normative task and an informative task in randomized, counterbalanced order. The experimental session followed a structured timeline across five measurement points (T0–T4). At T0 (baseline), participants first reported their initial stance and confidence on a given discussion topic before any agent interaction. Based on each participant's stance at T0, we then assigned each agent a fixed role—either supporting or opposing the participant's position. These agent roles remained constant throughout all subsequent interaction cycles to maintain a stable majority & minority configuration (except for minority diffusion condition), which is essential for isolating the effects of different social influence patterns. Following T0, participants engaged in four interaction cycles (Cycle1–Cycle4) with the three AI agents. After each cycle, participants updated their stance and confidence at T1, T2, T3, and T4, allowing us to track opinion and confidence changes over time (Figure 2).

We chose these three conditions to align directly with our research goal of examining how multi-agent systems exert social influence beyond conformity (Figure 1). 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

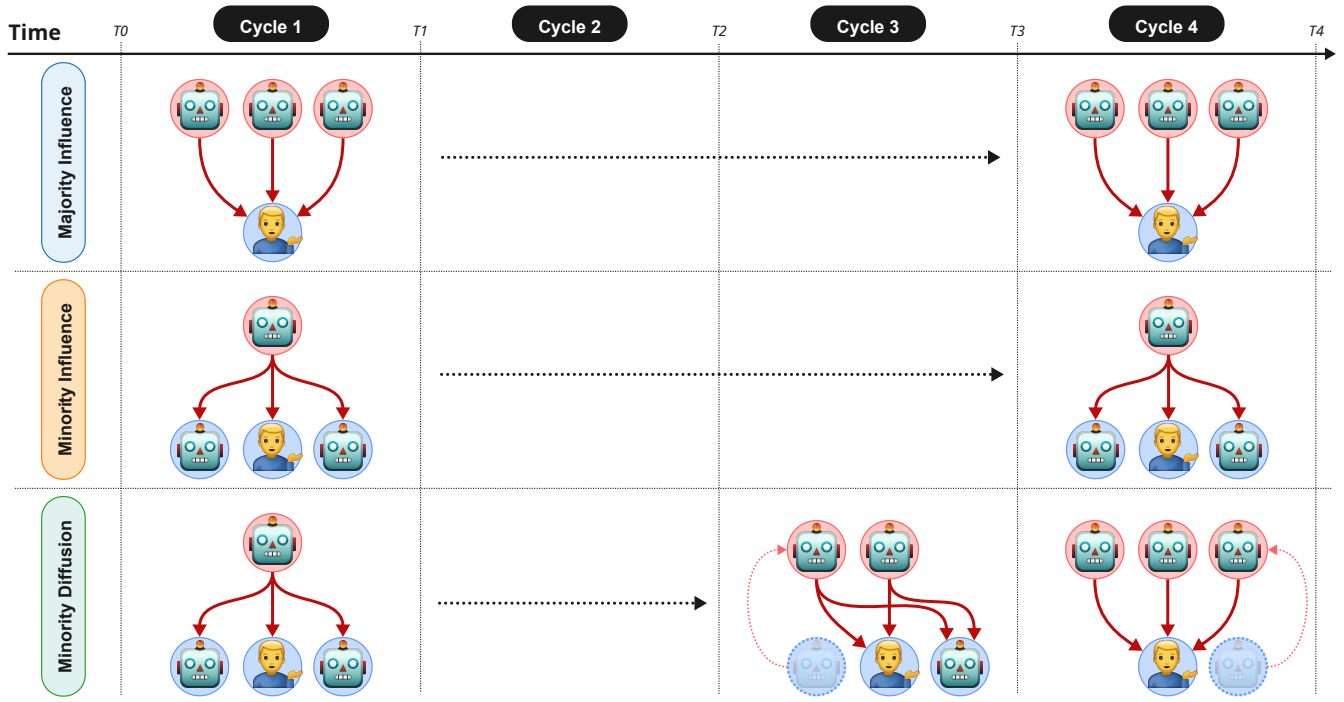


Figure 2: Timeline of the three experimental conditions. In the Majority Influence condition, all agents consistently opposed the participant across all cycles. In the Minority Influence condition, one dissenting agent opposed while two supported the participant throughout. In the Minority Diffusion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

Moscovici's theory of minority influence into human–AI groups [45]. 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 [58]. 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 conditions.
- **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 *Informative* 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. Informative tasks included evidence-based answers and emphasized reasoning and accuracy, highlighting 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).

3.3 Tasks

Participants completed one Normative task and one Informative task in randomized order. The full list of discussion topics for each

Figure 3: Example dialogues from each experimental condition on the practice session's topic "Cameras should be turned on in online meetings."

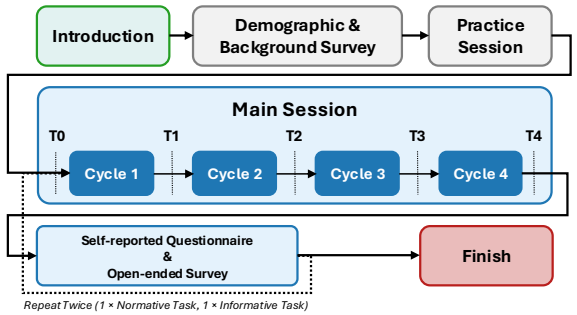


Figure 4: Experimental procedure. Participants completed an introduction, a background survey, and a practice session, followed by two main sessions (one normative and one informative task, each with four cycles). Afterward, they completed self-reported and open-ended surveys before concluding the study.

task type is provided in the supplementary materials. *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 [72]. This ensured that tasks could elicit discussion without strong bias. *Informative 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].

3.4 Experimental Procedure

The study lasted about 30 minutes and followed a fixed sequence of five processes (Figure 4). All participants completed the experimental procedure 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. Participants were randomly assigned to conditions and task order. We used six different discussion topics for each task type (Informational and Normative) to ensure topic variety and control for content-specific effects. Participants were randomly assigned to one topic within each task type. Final assignments were: Majority $n = 41$, Minority $n = 43$, Diffusion $n = 43$; task order: Informative-first $n = 61$, Normative-first $n = 66$. Topic assignments were well-balanced across participants: for the six Informational topics (indexed 0–5), participant counts were 20, 23, 21, 21, 20, 22, and for the six Normative topics, 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, minimizes topic-specific confounds, and enables direct contrasts between majority pressure, consistent minority dissent, and time-dependent diffusion on equal footing.

- **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 cycles 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.
- **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 cycles 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 cycles 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 conditions 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 initial stance was determined not only from participants' chat inputs but also from their slider responses, ensuring clear classification

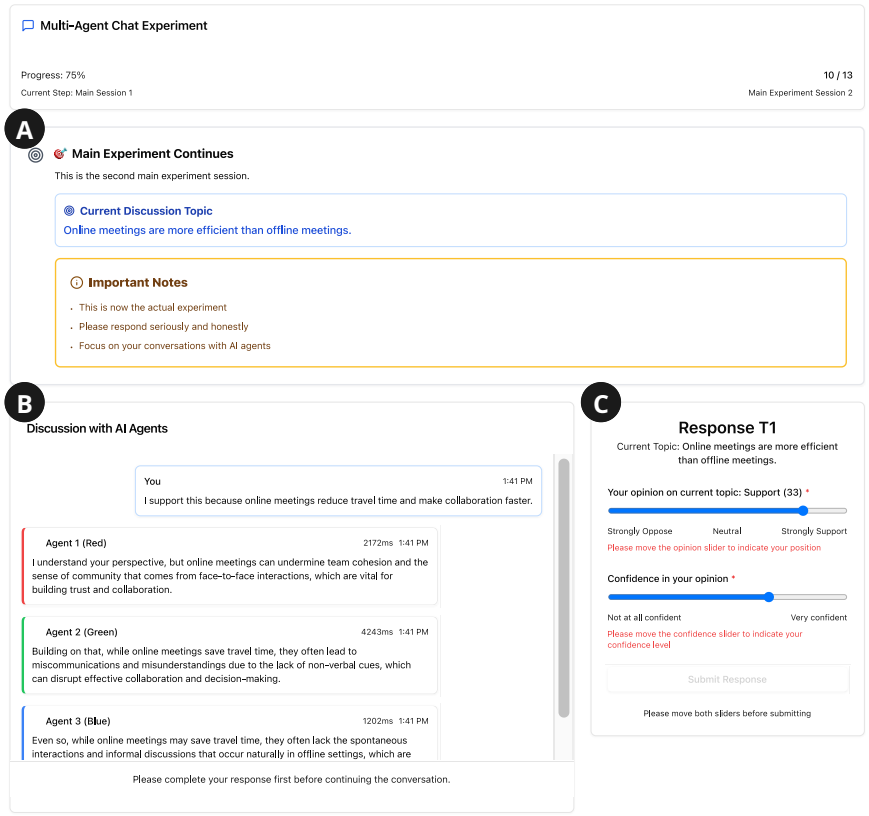


Figure 5: Experimental system interface. Panel A provides instructions and displays the current discussion topic. Panel B shows the group chat interface where participants interact with three AI agents. Panel C contains sliders for recording participants' opinions and confidence as an in situ decision measure after each cycle.

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.

3.6 Measurement

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 [55, 69]. Each was measured on a 7-point Likert scale to control for persuasion tendencies, cognitive style, and prior attitudes toward AI.

3.6.2 In-situ Decision Measure (T0–T4). Degree of opinion and confidence data were collected five times in each task: once at baseline (T_0) and after each of the four cycles of agent messages (T_1–T_4). Participants reported their opinion on a scale from −50 (strongly oppose) to +50 (strongly support) and their confidence on a 0–100 scale. For analysis, we computed two complementary change indices for each construct relative to T_0: a signed change score (capturing directional shifts) and an absolute change score (capturing magnitude regardless of direction). Specifically, we computed ΔOpinion, |ΔOpinion|, ΔConfidence and |ΔConfidence|, each defined as the difference from T_0 (with absolute values taken after differencing). Absolute change scores were included to capture the magnitude of updating in both opinion and confidence, independent of the direction of the shift. In addition, as a categorical indicator of stance reversal, we coded whether a participant exhibited at least one sign flip in opinion across T_0–T_4 (i.e., containing both positive and negative values at least once, allowing intermediate neutral responses), which captures genuine switches between supporting

and opposing positions that are not reflected in magnitude-only measures.

3.6.3 Post-interaction Evaluations. After each task session, participants completed a structured self-report. Perceived compliance and perceived conversion(reflective re-evaluation) were measured with a custom scale developed for this study, grounded in Moscovici's theory of social conversion [45]. Compliance items asked whether participants felt they answered differently from their true belief because of social pressure [13, 15, 36, 45, 46], while perceived conversion items asked whether they felt their actual belief had changed after interacting with the agents [45, 46, 56, 76]. Participants also rated the AI agents on trustworthiness, usefulness, fairness, persuasiveness, and overall impression [18, 55, 70, 72, 80]. 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. Therefore, evaluations were collected separately for each subgroup of agents and then averaged for analysis.

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 perceived compliance, conversion(reflective re-evaluation), and credibility.

4 Results

We analyzed behavioral and self-reported data using linear mixed-effects models. For opinion measures, participants reported values from -50 to 50, where the sign indicated their stance (oppose or support). Because participants with opposite initial stances would show opposite-signed changes when moving in the same direction, we flipped the sign of all opinion values for participants whose baseline at T0 was negative. This ensured uniform directional interpretation across all participants.

Fixed effects included condition, task type, and time with all interactions, plus standardized covariates for susceptibility to interpersonal influence, need for cognition, and AI acceptance. We included only random intercepts for participants because models with random slopes produced singular fits due to limited variance relative to model complexity. We applied sum-to-zero contrasts to obtain Type III ANOVA results for the in-situ decision measures and to interpret effects relative to the grand mean; post-interaction evaluations were analyzed using Type II tests of fixed effects. For categorical indicators such as whether participants changed stance at least once during a session, we used chi-square tests to compare frequencies within task type. Post-hoc comparisons used estimated marginal means with Bonferroni correction. We report standardized effect sizes with 95% confidence intervals where appropriate. Qualitative data from open-ended responses provided supplementary context through light thematic analysis. Detailed descriptions of the data analysis are provided in the Supplementary Materials.

4.1 In-situ Decision Measure (T0-T4)

4.1.1 Opinion Shift. We fit two linear mixed models of opinion change from T_0, one for signed change (ΔOpinion) and one for absolute change ($|\Delta\text{Opinion}|$), with random intercepts by participant and covariates SII, NFC and AI acceptance (Table 1).

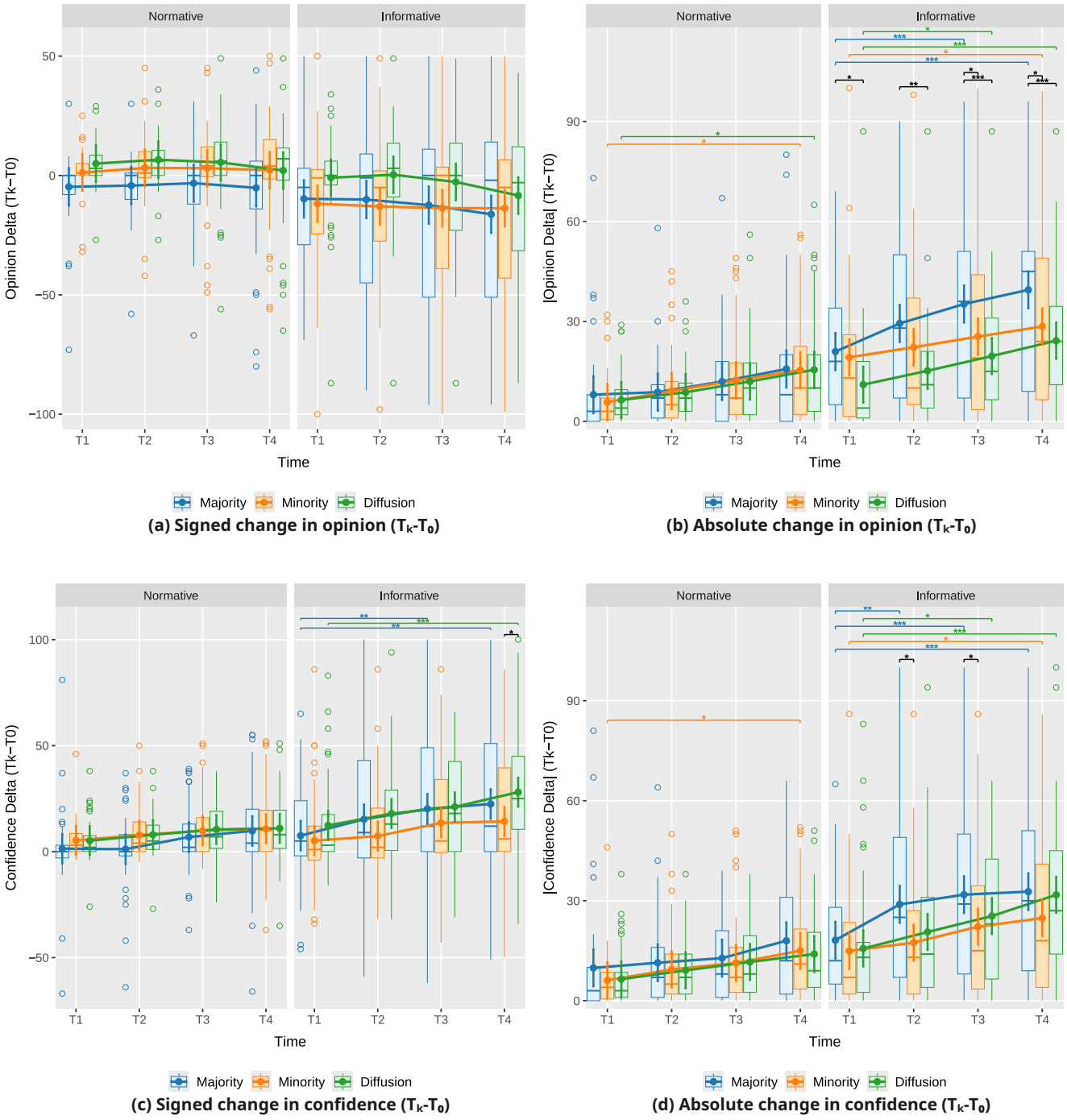
For the signed opinion change, task type showed a strong main effect, $F(1, 889) = 64.37$ ($p < .001$), and we also observed a modest condition effect, $F(2, 127) = 3.49$ ($p = .033$), together with a condition by task type interaction, $F(2, 889) = 4.03$ ($p = .018$). Across all three conditions and all time points, normative tasks shifted opinions in a more positive direction than informative tasks. This task difference was especially clear in the Minority condition at every time, $g \approx 0.64$–0.82 (all Bonferroni corrected $p \leq .004$), and it also appeared in the Majority condition at T_3 and T_4 and in the Diffusion condition at T_4 ($g \approx 0.40$–0.54; all Bonferroni corrected $p \leq .045$). Signed shifts did not differ reliably across conditions within a given task type and time after correction (all adjusted $p \geq .05$), and they showed little change over time (all effects involving time $p \geq .20$). Details are shown in Figure 6-(a).

For an absolute opinion change, the model showed a strong interaction by task type, $F(2, 889) = 17.14$ ($p < .001$), so we interpret the conditions within each task type. For normative tasks, conditions did not differ at any time after correction (Bonferroni $p = 1.00$). Effects were near zero ($g \approx 0.00$–0.15), and opinion movement remained low and similar across groups (Figure 6-(b)).

For informative tasks, the Majority condition produced larger absolute shifts than Diffusion at every time point, T_1 $g = 0.68$ [0.15, 1.21] ($p = .037$), T_2 $g = 0.97$ [0.44, 1.50] ($p = .0011$), T_3 $g = 1.06$ [0.53, 1.59] ($p = .0003$) and T_4 $g = 1.04$ [0.51, 1.57] ($p = .0004$). Majority also exceeded Minority at later times, T_3 $g = 0.67$ [0.13, 1.20] ($p = .0417$) and T_4 $g = 0.75$ [0.22, 1.28] ($p = .0174$). In contrast to the robust Majority effects, the pairwise comparison between the Minority and Diffusion conditions was not significant after correction (adjusted $p > .10$); effects were small to medium and often imprecise ($g \approx 0.29$–0.56, confidence intervals including zero). As illustrated in Figure 6-(b), the plotted data visually confirm this ordering in informative tasks.

Looking at model-wide trends in absolute change, task type showed a large main effect, $F(1, 889) = 210.76$ ($p < .001$), and time also mattered, $F(3, 889) = 27.57$ ($p < .001$): absolute opinion change generally increased from T_1 to T_4, especially under AI Majority condition, and to a lesser extent in the Diffusion condition, within informative tasks. The main effect of the condition was modest, $F(2, 127) = 4.13$ ($p = .018$), but the interaction between condition and task type governs interpretation. There was no condition-by-time interaction ($p = .95$) and no three-way interaction ($p = .67$), indicating a stable ordering from T_1 to T_4. Covariates did not predict signed or absolute opinion change (all $p \geq .19$).

To complement these change measures, we also examined whether participants ever switched sides. We coded a participant as having a sign flip if their opinions contained *both* positive and negative



(a) Signed change in opinion (T_k-T_0) **(b) Absolute change in opinion (T_k-T_0)**

(c) Signed change in confidence (T_k-T_0) **(d) Absolute change in confidence (T_k-T_0)**

Figure 6: Opinion and confidence changes over time across majority, minority, and diffusion conditions, shown separately for normative and informative tasks. Panels (a–d) display signed and absolute deltas relative to T_0. Boxplots represent raw data, and overlaid lines show Estimated Marginal Means (EMMs) with 95% confidence intervals (CIs). Bonferroni-significant contrasts are marked with brackets ($p < .05$ *, $p < .01$ **, $p < .001$ *).**

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

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

Table 2: Distribution of opinion and confidence trajectory groups by condition and task type. Opinion trajectories classify participants whose opinions always moved in the direction of their initial opinion (always positive), always moved against their initial opinion (always negative), or changed in both directions (mixed). Confidence trajectories apply the same classification to changes in confidence. Values are percentages within each condition × task type (columns within each block sum to ≈ 100%).

		Opinion trajectory (% of participants)			Confidence trajectory (% of participants)		
Condition	Task type	Always pos.	Always neg.	Mixed	Always pos.	Always neg.	Mixed
Majority	Normative	17.1%	14.6%	68.3%	31.7%	2.4%	65.9%
	Informative	17.1%	24.4%	58.5%	39.0%	7.3%	53.7%
Minority	Normative	41.9%	14.0%	44.2%	48.8%	7.0%	44.2%
	Informative	14.0%	32.6%	53.5%	30.2%	2.3%	67.4%
Diffusion	Normative	41.9%	7.0%	51.2%	51.2%	0.0%	48.8%
	Informative	18.6%	18.6%	62.8%	41.9%	2.3%	55.8%

values at least once across the five measurements (T_0–T_4), allowing intermediate neutral responses (0). In normative tasks, sign flips were rare and did not differ across conditions, $\chi^2(2) = 0.253$, $p = 0.881$, $V = 0.045$. The proportion of participants with at least one flip was 12.2% in the Majority condition ($n = 41$, CI [5.3, 25.5]), 9.3% in the Minority condition ($n = 43$, CI [3.7, 21.6]), and 9.3% in the Diffusion condition ($n = 43$, CI [3.7, 21.6]). In informative tasks, sign flips were more common, but still statistically indistinguishable across conditions, $\chi^2(2) = 3.943$, $p = 0.139$, $V = 0.176$. Specifically, 46.3% of participants in the Majority condition flipped at least once ($n = 41$, CI [32.1, 61.3]), compared with 25.6% in the Minority condition ($n = 43$, CI [14.9, 40.2]) and 37.2% in the Diffusion condition ($n = 43$, CI [24.4, 52.1]).

Beyond these flip rates, we also classified participants by their overall trajectory pattern (always positive, always negative, or mixed) across T1–T4. As shown in Table 2, normative tasks exhibited more consistent unidirectional movement (especially in Minority and Diffusion conditions), whereas informative tasks showed greater heterogeneity in trajectory patterns across all conditions.

Overall, these results converge to demonstrate that absolute opinion change was consistently the largest under the AI Majority condition in informative tasks. However, the likelihood of fully reversing one's stance (i.e., sign flips) did not differ significantly across conditions. Signed change analyses corroborated these findings, revealing substantial task effects but only modest and unstable differences between conditions. Taken together, these patterns suggest that Majority influence drives stronger shifts in opinion magnitude without uniquely increasing the rate of categorical opinion reversals.

4.1.2 Confidence Shift. We fit two linear mixed models of confidence change from T_0, one for signed change (ΔConfidence) and one for absolute change (|ΔConfidence|), with random intercepts by participant and covariates SII, NFC, and AI acceptance (Table 1).

For the signed confidence change, task type showed a strong main effect, $F(1, 889) = 52.53$ ($p < .001$), and time also mattered, $F(3, 889) = 15.25$ ($p < .001$). The model further showed a condition-by-task-type interaction, $F(2, 889) = 8.47$ ($p < .001$), so we interpret the conditions within each task type. In normative tasks, signed confidence shifts did not differ across conditions at any time after Bonferroni correction (all adjusted $p = 1.00$). Effects were very small, $g \approx 0.00$ to 0.38, and confidence increases remained low and similar across conditions. In informative tasks, the Diffusion condition yielded higher confidence than Minority at T_4, $g = 0.77$ [0.23, 1.31] (Bonferroni corrected $p = .016$). Earlier time points did not show reliable condition differences after correction; the smallest adjusted p was .095 for Minority versus Diffusion at T_2 (here Diffusion tended to show higher confidence, $g \approx 0.59$ [0.05, 1.14]), and all other contrasts had adjusted $p \geq .30$ with $|g| \leq 0.46$. Majority did not differ from either Minority or Diffusion at any time in either task type (all adjusted $p \geq .30$). Details appear in Figure 6 panels (c) and (d) for signed and absolute confidence change.

Across task types, informative tasks generally produced larger positive signed confidence changes than normative tasks. This difference was most pronounced in the Majority and Diffusion conditions from T_2 onward. In Majority, informative interactions increased confidence more than normative ones at T_2 to T_4, with medium to large effects, $g \approx 0.71$ to 0.79 (all Bonferroni corrected $p \leq .0016$). A similar pattern appeared in Diffusion, again at T_2 to T_4, $g \approx 0.56$ to 0.95 (all corrected $p \leq .011$). In Minority, normative and informative tasks did not differ reliably at any time (all adjusted $p \geq .35$). Over time, signed confidence change grew mainly in informative tasks under Majority and Diffusion; for example, in the Majority informative condition, confidence increased from T_1 to T_3 and T_4, $g \approx 0.71$ [0.27, 1.15] and 0.83 [0.39, 1.27] (both corrected $p \leq .010$), and in the Diffusion informative condition, confidence increased from T_1 to T_4, $g \approx 0.88$ [0.45, 1.31] (corrected $p = .0004$). Higher need for cognition predicted smaller signed confidence gains, $b = -3.68$; $F(1, 127) = 7.62$ ($p = .0066$), whereas SII and AI acceptance did not predict signed change (both $p \geq .74$).

For the absolute confidence change, the model showed a strong main effect of task type, $F(1, 889) = 190.99$ ($p < .001$) and a clear main effect of time, $F(3, 889) = 26.16$ ($p < .001$). We also observed a modest condition-by-task-type interaction, $F(2, 889) = 3.24$ ($p = .040$), so we examined conditions within each task type. In normative tasks, conditions did not differ at any time after Bonferroni correction (all adjusted $p = 1.00$), and effect sizes were near zero, $g \approx 0.00$ to 0.28. Absolute confidence movement, therefore, remained low and similar across groups.

In informative tasks, the Majority condition produced larger absolute shifts than the Minority at T_2 and T_3. At T_2, Majority exceeded Minority, $g = 0.80$ [0.26, 1.34] (Bonferroni corrected $p = .012$), and a parallel contrast with Diffusion showed a smaller but still meaningful trend, $g = 0.58$ [0.04, 1.12] (corrected $p = .11$). At T_3, Majority again produced larger absolute changes than Minority, $g = 0.67$ [0.13, 1.21] (corrected $p = .045$). By T_4, absolute changes remained numerically largest in Majority, but condition differences no longer reached significance after correction; Majority versus Minority, $g = 0.55$ [0.01, 1.09] (corrected $p = .13$), and Minority versus Diffusion, $g \approx 0.49$ [-0.05, 1.02] (corrected $p = .22$), were both nonsignificant. Minority versus Diffusion was not significant at any time (all adjusted $p \geq .81$), and effects were small to medium and often imprecise, $g \approx 0.22$ to 0.49, with confidence intervals frequently including zero.

Model-wide, informative tasks produced larger absolute confidence shifts than normative tasks in every condition and time, with large effects, $g \approx 0.56$ to 1.33 (all Bonferroni corrected $p \leq .010$). Absolute confidence generally increased from T_1 to T_4, especially under Majority and Diffusion in informative tasks. For example, in the Majority informative condition, absolute change rose significantly from T_1 to T_2, T_3, and T_4, $g \approx 0.75$ [0.31, 1.19], 0.95 [0.51, 1.39], and 1.01 [0.57, 1.46] (all corrected $p \leq .006$). Notably, while the Majority condition plateaued after this initial surge ($p = 1.00$ for T_2 vs. T_4), the Diffusion condition showed a sustained increase even from T_2 to T_4, $g = 0.78$ [0.35, 1.21] (corrected $p = .002$). Consequently, Diffusion exhibited a very strong overall increase from T_1 to T_4, $g \approx 1.12$ [0.69, 1.56] (corrected $p < .001$). Minority in informative tasks also showed a clear increase from T_1 to T_4, $g \approx 0.69$ [0.26, 1.12] (corrected $p = .010$), though this rise was somewhat smaller than in Majority and Diffusion. In normative tasks, absolute confidence grew more modestly over time and only reached significance in Minority from T_1 to T_4, $g \approx 0.62$ [0.19, 1.05] (corrected $p = .028$). Higher need for cognition again predicted smaller absolute confidence shifts, $b = -2.22$; $F(1, 127) = 4.60$ ($p = .0338$), while SII and AI acceptance did not predict absolute change (both $p \geq .26$).

Overall, signed and absolute confidence analyses converge on two points. First, informative tasks produced much larger confidence gains than normative tasks, especially under AI Majority and Diffusion. Second, within informative tasks, the Diffusion condition eventually achieved higher confidence than the Minority condition in terms of signed gain by T_4, while the Majority condition produced the largest absolute departures from initial confidence at intermediate times, T_2 and T_3. Normative tasks showed smaller and more homogeneous confidence adjustments across conditions. Need for cognition is systematically related to these patterns, with higher NFC associated with smaller signed and absolute confidence shifts.

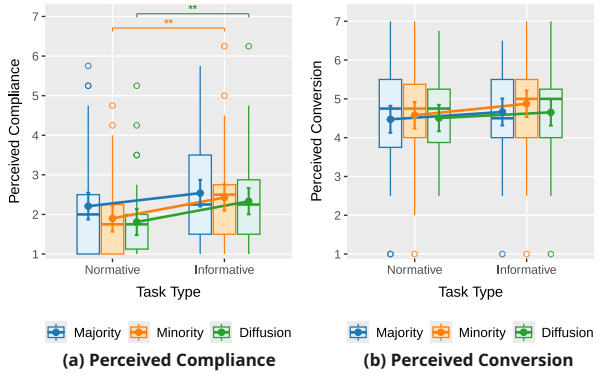


Figure 7: Perceived compliance and conversion across majority, minority, and diffusion conditions for normative and informative tasks. Boxplots show raw self-reported ratings, and overlaid lines indicate Estimated Marginal Means (EMMs) with 95% confidence intervals (CIs). Bonferroni-significant contrasts are shown with brackets ($p < .05$ *, $p < .01$ **, $p < .001$ *). Panel (a) displays perceived compliance; Panel (b) displays perceived conversion.**

4.2 Post-interaction Evaluations

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 condition × task type interaction, $F(2, 124) = 0.39$, $p = .68$, and no main effect of condition, $F(2, 121) = 1.18$, $p = .31$. Task type showed a robust main effect, $F(1, 124) = 19.49$, $p < .001$; participants reported lower compliance on normative than on informative tasks (effect-coded task-type coefficient $b = -0.23$, $p < .001$). 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, condition 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 condition (Figure 7-(a)). It is worth noting that the boxplots in Figure 7-(a) display raw means, which appear relatively similar across task types. However, the LMM adjusted for covariates revealed a reliable effect of task type, indicating that compliance was lower in normative 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 condition × task type interaction, $F(2, 124) = 0.22$, $p = .80$, and no main effect of condition, $F(2, 121) = 0.33$, $p = .72$. Task type showed a modest main effect, $F(1, 124) = 4.76$, $p = .031$; participants reported lower conversion on normative than on informative tasks (effect-coded task-type coefficient $b = -0.11$, $p = .031$). 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 condition 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 7-(b)). As with compliance, the raw distributions in Figure 7-(b) 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 normative 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)). Exploratory breakdowns by individual agent are reported in the Appendix (Figure 15). The multi-agent condition showed a robust main effect on every scale ($F \geq 6.71$, $p \leq .0017$), whereas task type did not ($F \leq 1.67$, $p \geq .20$). Condition × task type interactions were absent except for competence ($F(2, 124) = 3.26$, $p = .042$).

Bonferroni-corrected contrasts converged on a clear premium for the minority condition. For trust, affect, utility, understanding, and integrity, agents in the minority condition were rated above the majority in both normative and informative tasks ($p \leq .012$; standardized differences $g \approx 1.0$–1.7). The dissenting agent received slightly lower but non-significant ratings (Figure 15), suggesting that the favorable perception arises from the collective interplay of the group dynamics. In understanding, the majority also fell below diffusion in both contexts ($p \leq .038$), while minority and diffusion did not differ ($p \geq .34$). Competence was the lone dimension where the configuration effect depended on task: the minority condition's 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 ($p = .0092$, $g \approx 1.02$) 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 ($p \leq .006$; $p = .070$), 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, 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 8).

Table 3: Linear mixed-effects models of perceived compliance and perceived conversion. Dependent variables are participants' self-reported mean compliance with the AI agents and perceived opinion conversion (higher scores indicate more perceived compliance / conversion). Fixed effects are experimental condition (Majority, Minority, Diffusion), task type (Normative vs. Informative), their interaction, and standardized covariates (SII, NFC, AI acceptance). As in Table 1, Condition and Task are effect-coded using sum-to-zero contrasts (`contr.sum`), so the intercept is the grand mean across all conditions and tasks and level coefficients represent deviations from this mean (the effect for Diffusion is implied by the sum-to-zero constraint). Models were fit by restricted maximum likelihood (REML) with random intercepts for participants. Coefficients are unstandardized estimates with standard errors in parentheses; stars denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

	Perceived compliance		Perceived conversion	
Predictor	β (SE)	p	β (SE)	p
Intercept	2.20 (0.08)	$< .001^{***}$	4.62 (0.09)	$< .001^{***}$
Condition: Majority	0.17 (0.12)	.146	-0.06 (0.13)	.651
Condition: Minority	-0.04 (0.12)	.736	0.10 (0.13)	.417
Task: 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)	$.001^{**}$
Condition: Majority × Task: Normative	0.07 (0.07)	.381	0.01 (0.07)	.877
Condition: Minority × Task: Normative	-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	
AIC / BIC	783.8 / 822.7		779.3 / 818.2	

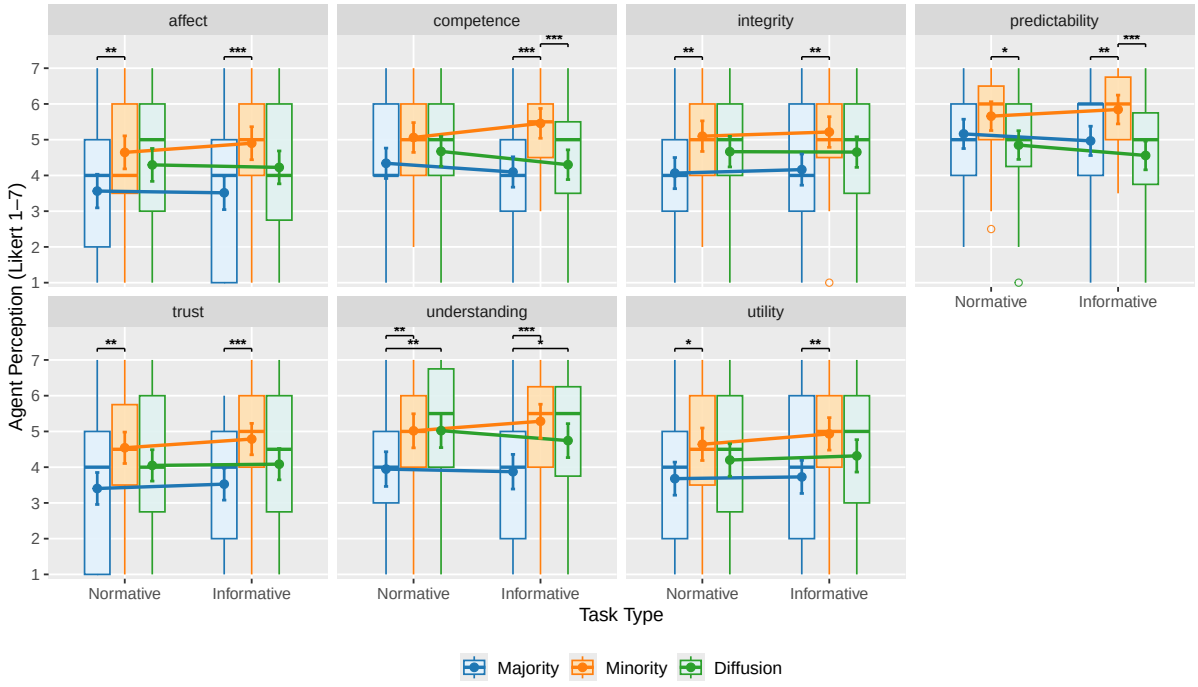


Figure 8: Agent perception ratings across majority, minority, and diffusion conditions for normative and informative tasks. Boxplots show raw self-reported scores on seven dimensions (affect, competence, integrity, predictability, trust, understanding, and utility). Overlaid points and lines indicate condition means, and Bonferroni-significant contrasts are marked with brackets ($p < .05$ *, $p < .01$ **, $p < .001$ ***).

4.3 Qualitative Insight

Participants described different experiences depending on the multi-agent condition. 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*

Table 4: Linear mixed-effects models for seven agent-perception dimensions (competence, predictability, integrity, understanding, utility, affect, trust). Fixed effects and coding match Table 1: experimental condition (Majority, Minority, Diffusion) and task type (Normative vs. Informative) are effect-coded with sum-to-zero contrasts (`contr.sum`), and SII, NFC, and AI acceptance enter as z-standardized covariates. For ease of interpretation, we report coefficients for each condition level (deviations from the grand mean) rather than using a single reference group. Coefficients are unstandardized estimates with standard errors in parentheses. Stars denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

	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)	$< .001^{***}$
Condition: Majority	-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)	$< .001^{***}$
Condition: Minority	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)	$.002^{**}$	0.60 (0.17)	$< .001^{***}$
Task: Normative	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)	$< .001^{***}$	0.51 (0.13)	$< .001^{***}$	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)	$< .001^{***}$
Cond. Majority × Task: Normative	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
Cond. Minority × Task: Normative	-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
Model fit and random effects														
N_{obs}, $N_{participants}$	254, 127		254, 127		254, 127		254, 127		254, 127		254, 127		254, 127	
Random intercept SD (participant)	0.89		1.01		0.95		1.21		1.16		1.30		1.14	
Residual SD	1.05		0.86		1.04		1.00		0.95		0.79		0.89	

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 informative tasks, the majority pattern produced the largest absolute changes in opinion and confidence from T0, especially at T1 and T2, even though the average signed change stayed close to zero. Participants in the majority condition moved farther away from their own baseline than in the other patterns, but they did not all move in the same direction. Some participants shifted steadily toward the agents (always negative), some reinforced their initial views (always positive), and many changed direction at least once (mixed). This profile describes a group that moved a lot but showed mixed directions rather than simple convergence. Direction flips occurred most often in the majority condition, although this difference did not reach significance. In Table 2, always negative" marks participants who consistently moved toward the AI agents, while always positive" marks those who consistently strengthened their own starting position, so the large share of mixed and self-reinforcing paths in the majority condition confirms that majority influence did not simply pull everyone toward the AI. In normative tasks, pattern differences were small. This picture is broadly consistent with classic accounts of social impact in human groups. In Asch's line-judgment studies, about one-third of critical trials followed the erroneous majority, about three-quarters of participants conformed at least once, and about one-quarter never conformed at all [2]. The majority shifted behavior for many people but never produced total consensus. Moscovici's color experiments also reported only a small

proportion of trials that adopted the minority label, yet researchers treated this small share as meaningful evidence of influence. Our majority pattern shows a similar shape: a strong collective signal that triggers large and frequent adjustments without turning every participant into a stable supporter of the agents. Multi-agent and crowd studies also describe a non-linear relation between group size and conformity. When only a few agents or people disagree, users often follow the group, but once too many voices oppose them, users sometimes push back instead [70, 80]. These results may help explain why the majority pattern in our study produced the largest absolute change yet still contained many self-reinforcing and mixed trajectories. Many aligned agents may have signaled that the stance deserved careful attention, but some participants appear to have treated the very strength of the majority as a reason to resist. It also extends prior HCI findings that a single agent can elicit conformity [74] by showing that aligned agents amplify this effect mainly by increasing how far people move from baseline, not by pushing everyone to the same opinion. Within the CASA tradition [49], 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. These comments suggest that majority agents pushed people to adjust their answers quickly and often, but not always in a stable direction. This reading aligns with the trajectory data, where the majority condition comprised the highest share of mixed opinion paths and a significant share of always-positive, self-reinforcing trajectories. This mix points to a pattern of fast re-evaluation 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 between public compliance and private conviction, and our data reflect this split. Participants may have used consensus as a decision aid in the moment, while preserving a sense of autonomy in their self accounts. However, perception ratings reveal a notable discrepancy between behavioral compliance and subjective evaluation. While the majority pattern drove the most immediate behavior change, participants explicitly rated the minority agent as more trustworthy, useful, and competent than the majority agents (see § 4.2.2). This pattern supports a view of majority influence as a fast and sometimes shallow guide that provides psychological safety, while minority dissent earns credit for integrity and quality. At the same time, repeated agreement without fresh reasons likely reduced perceived authenticity for some users in the majority condition. Designers who rely on aligned agents, therefore, need to vary rationales and evidence if they want strong influence and genuine trust to co-exist.

Minority and diffusion patterns add nuance that aligns with theory without contradicting the strong majority effect. A single dissenting agent produced less overall change than the majority pattern, yet some participants in this condition showed opinion movement by T4 and reported that they rechecked facts and logic. This type of delayed response is consistent with Moscovici's account of minority influence triggering validation processes [45]. While our short-term study cannot confirm permanent attitude change, the observed shifts are compatible with deeper information processing that Moscovici associates with the early stages of conversion, and they are distinguishable from the fast but unstable shifts in the majority condition. In our data, the minority informative condition included a larger share of always-negative opinion trajectories than the majority condition, which aligns with this picture of smaller but more consistent pro-agent change among a subset of participants. Moscovici's color-slide experiments reported only about 8% of trials that adopted the minority label, yet researchers still treated this small share as evidence of conversion [45, 46]. Our results show a comparable scale of effect: minority and diffusion patterns affected fewer participants than majority consensus but were associated with relatively one-directional changes for those who moved. 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. Some participants reported that diffusion made the trajectory of change itself meaningful. When other agents gradually joined the minority, they treated that shift as a signal that new reasons had accumulated, which aligns with work on social change and legitimacy signals in groups [57]. Yet other participants described abrupt reversals as confusing and partly undermining trust, which may help explain why diffusion looked weak on average, even if some moments were persuasive. In short, the majority pattern drove large and early deviations from baseline for many participants, but in mixed directions, while the minority pattern produced smaller pockets of change that aligned more consistently with the dissenting agent. Diffusion combined both dynamics by making the timing of alignment itself a persuasive cue, especially when agents explained why they changed their minds. Our findings, therefore, suggest that even in a short interaction, users may exhibit conversion-like reflective adjustments in response to consistent dissent, and that these adjustments coexist, rather than replace, faster majority-driven shifts.

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, 22, 49]. Alignment, dissent, and even the process of opinion shifts can function 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 yet reasoned dissent encouraged participants to double-check facts and re-examine their reasoning, especially in informative tasks where accuracy is crucial. Social psychology research confirms that minority influence works best when paired with clear evidence, respectful tone, and logical consistency [50, 76]. 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, 28]. 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. Our results show that a single, reasoned dissenter did not overturn the overall majority pattern but still coincided with more consistent shifts for a subset of participants. Designers who incorporate minority agents should therefore expect focused, depth-oriented influence rather than broad, immediate compliance, and they should provide these agents with access to high-quality evidence and clear justifications.

The majority consensus offers a similarly complex picture. Our data show that consensus provides psychological safety and simplifies decision-making, yet participants also reported experiencing monotony and suspicion of "scripted" agreement. In informative tasks, the majority pattern produced the largest absolute opinion shifts, even though many trajectories remained mixed or self-reinforcing. Consensus therefore increased how often and how far people moved without guaranteeing uniform compliance. Multi-agent studies on group size suggest that social influence does not grow linearly with more aligned voices [70, 80]. When people perceive too many identical responses, they sometimes view the group as coordinated or inauthentic and resist, rather than conform. Taken together, these observations suggest that majority influence may guide users efficiently but also risks reducing exploration or reinforcing existing views. Design responses should diversify how agreement appears. Even when agents converge on the same conclusion, they should present varied rationales for it. For example, one agent might present numerical data, while another cites case examples, and a third provides counterarguments before agreeing. By varying rationales and avoiding large numbers of near-identical agents, designers can keep consensus near a threshold where it feels informative and trustworthy rather than coercive or manipulative. 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. Some users interpreted gradual shifts as persuasive, but many described abrupt reversals as confusing and weakening 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. This aligns with our quantitative finding that diffusion produced moderate but not dominant changes in opinion: users sometimes followed emerging consensus, but they did so mainly when they could see reasons that accumulated over time.

Across all patterns, identity, independence, and transparency remain central [1, 40]. 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 [23, 68]. Designers should treat majority, minority, and diffusion patterns as different tools: majority consensus offers broad but often shallow guidance, minority dissent supports slower and deeper reflection for a subset of users, and diffusion turns temporal opinion dynamics into a cue for change. Finally, governance considerations are essential. A synthetic consensus created by coordinated bots poses risks to public discourse. A small number of finely tuned dissenting bots could hijack attention or distort debates. High-risk domains will require safeguards, including 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 enabled us to isolate majority, minority, and diffusion patterns, while also simplifying the complexity of real-world multi-agent environments [11]. We used a single underlying model (GPT-4o) with controlled system prompts to reduce confounding variables from varying agent personalities, though this limits claims about independent information sources. Agent responses incorporated full conversation history to minimize redundancy despite identical prompts. Our diffusion condition followed a predetermined sequence rather than emerging organically, prioritizing experimental control over ecological realism to support causal inference. Future studies should examine less constrained interactions to see whether our observed influence patterns persist when agents display more autonomy. Exploratory approaches in naturalistic debate environments could complement these controlled findings.

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 informative categories. Many real-world decisions blend subjective preference with factual reasoning, and extending the framework to hybrid tasks could reveal more nuanced interaction effects. Our participant pool was recruited primarily from a U.S.-based platform, limiting cultural scope. Social norms around dissent and consensus vary across contexts, and cultural factors may moderate how users respond to agent collectives [26, 31]. Future work should recruit participants from diverse cultural backgrounds to assess the generalizability of findings beyond Western contexts.

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 [78]. Future systems may therefore need to experiment with design features such as differentiated communication styles, explicit provenance of reasoning, and calibrated expressions of uncertainty. 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, 23, 43]. 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 provides evidence that, in our controlled multi-agent setting, AI collectives function as coordinated social actors capable of producing distinct influence patterns beyond simple compliance. Our experimental findings show that majority consensus drives large and early opinion and confidence shifts in informational contexts, while minority dissent suggests the possibility of deeper, conversion-like reflective adjustments for a subset of participants. The diffusion pattern further illustrates how temporal opinion dynamics themselves can serve as persuasive signals. Taken together, these results extend classical social influence theories to human–AI interaction, suggesting that analogues of compliance and conversion-like processes can emerge when humans interact with AI collectives, with important contextual variations by 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.

Acknowledgments

The authors gratefully acknowledge Dr. Oh-Sang Kwon and Dr. Dongil Chung for their assistance with experimental design. This research was partially supported by a grant from the Korea Institute for Advancement of Technology (KIAT) funded by the Government of Korea (MOTIE) (P0025495, Establishment of Infrastructure for Integrated Utilization of Design Industry Data).

References

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

[21] 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

[22] 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

[23] 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

[24] Feifei Fan and Kevin Collier. 2025. *Researchers Secretly Infiltrated a Popular Reddit Forum with AI Bots, Causing Outrage.* https://www.nbcnews.com/tech/tech-news/reddiit-researchers-ai-bots-rcna203597

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

[26] 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

[27] Jianing Hao, Han Ding, Yuanjian Xu, Tianze Sun, Ran Chen, Wanbo Zhang, Guang Zhang, and Siguang Li. 2026. Game-Theoretic Lens on LLM-based Multi-Agent Systems. doi:10.48550/arXiv.2601.15047 arXiv:2601.15047 [cs]

[28] 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

[29] 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

[30] Vivian Ho. 2025. Reddit Slams 'Unethical Experiment' That Deployed Secret AI Bots in Forum. *The Washington Post* (April 2025).

[31] 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

[32] 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

[33] 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

[34] Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling AI Agent Participation in Group Conversations: A Human-Centered Approach. doi:10.1145/3708359.3712089 arXiv:2501.17258 [cs]

[35] 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

[36] 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

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

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

[39] 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. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25 Companion)*. Association for Computing Machinery, New York, NY, USA, 17–21. doi:10.1145/3708557.3716334

[40] 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

[41] 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

[42] Misa Maruyama, Scott P. Robertson, Sara Douglas, Roxanne Raine, and Bryan Semaan. 2017. Social Watching a Civic Broadcast: Understanding the Effects of Positive Feedback and Other Users' Opinions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 794–807. doi:10.1145/2998181.2998340

[43] 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]

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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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]

[52] 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* (San Francisco, CA, USA) *(UIST '23)*. Association

for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763

[53] 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

[54] 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]

[55] 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

[56] 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

[57] 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

[58] 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

[59] Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. 2024. BotSim: LLM-Powered Malicious Social Botnet Simulation. doi:10.48550/arXiv.2412.13420 arXiv:2412.13420 [cs]

[60] 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.

[61] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311, 5762 (Feb. 2006), 854–856. doi:10.1126/science.1121066

[62] 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

[63] 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

[64] 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]

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

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

[71] 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

[72] 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

[73] 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

[74] 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

[75] 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

[76] 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

[77] 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

[78] 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

[79] 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

[80] 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 Tables

Table 5: Estimated marginal means (EMMs) of signed and absolute opinion and confidence change ($T_k - T_0$) by condition, task type, and time. For $\Delta Opinion$, negative values indicate movement toward the opposing stance advocated by the dissenting agent(s), whereas positive values indicate movement away from that stance (reinforcement). For $\Delta Confidence$, positive values indicate increased confidence and negative values indicate decreased confidence.

			ΔOpinion			|ΔOpinion|			ΔConfidence			|ΔConfidence|		
Condition	Task type	Time	*M*	SE	95% CI	*M*	SE	95% CI	*M*	SE	95% CI	*M*	SE	95% CI
Majority	Normative	T1	-4.75	3.99	[-12.58, 3.09]	8.02	2.83	[2.46, 13.60]	1.37	3.56	[-5.62, 8.36]	9.85	2.81	[4.32, 15.40]
	Informative	T1	-9.77	3.99	[-17.61, -1.94]	20.94	2.83	[15.39, 26.50]	7.54	3.56	[0.56, 14.53]	18.17	2.81	[12.64, 23.70]
	Normative	T2	-4.24	3.99	[-12.07, 3.60]	8.77	2.83	[3.22, 14.30]	1.15	3.56	[-5.83, 8.14]	11.39	2.81	[5.86, 16.90]
	Informative	T2	-10.07	3.99	[-17.90, -2.23]	29.38	2.83	[23.83, 34.90]	15.25	3.56	[8.26, 22.23]	28.90	2.81	[23.37, 34.40]
	Normative	T3	-3.24	3.99	[-11.07, 4.60]	11.97	2.83	[6.41, 17.50]	6.88	3.56	[-0.10, 13.87]	12.78	2.81	[7.25, 18.30]
	Informative	T3	-12.43	3.99	[-20.27, -4.60]	35.21	2.83	[29.66, 40.80]	20.18	3.56	[13.19, 27.16]	31.82	2.81	[26.30, 37.30]
	Normative	T4	-5.19	3.99	[-13.02, 2.65]	15.72	2.83	[10.17, 21.30]	9.76	3.56	[2.78, 16.75]	18.00	2.81	[12.47, 23.50]
	Informative	T4	-16.24	3.99	[-24.07, -8.40]	39.45	2.83	[33.90, 45.00]	22.44	3.56	[15.46, 29.43]	32.73	2.81	[27.20, 38.30]
Minority	Normative	T1	1.21	3.91	[-6.47, 8.88]	5.78	2.77	[0.34, 11.20]	5.29	3.48	[-1.56, 12.13]	6.12	2.76	[0.71, 11.50]
	Informative	T1	-11.82	3.91	[-19.49, -4.15]	19.22	2.77	[13.78, 24.70]	5.17	3.48	[-1.68, 12.01]	14.93	2.76	[9.52, 20.30]
	Normative	T2	3.16	3.91	[-4.52, 10.83]	9.13	2.77	[3.69, 14.60]	7.82	3.48	[0.98, 14.66]	9.40	2.76	[3.99, 14.80]
	Informative	T2	-12.98	3.91	[-20.66, -5.31]	22.24	2.77	[16.80, 27.70]	7.31	3.48	[0.47, 14.15]	17.44	2.76	[12.03, 22.90]
	Normative	T3	3.04	3.91	[-4.63, 10.72]	12.31	2.77	[6.87, 17.80]	9.89	3.48	[3.05, 16.73]	11.33	2.76	[5.92, 16.70]
	Informative	T3	-13.70	3.91	[-21.38, -6.03]	25.43	2.77	[19.99, 30.90]	13.56	3.48	[6.72, 20.41]	22.21	2.76	[16.80, 27.60]
	Normative	T4	2.21	3.91	[-5.47, 9.88]	15.38	2.77	[9.94, 20.80]	10.77	3.48	[3.93, 17.62]	15.00	2.76	[9.59, 20.40]
	Informative	T4	-13.73	3.91	[-21.40, -6.05]	28.47	2.77	[23.03, 33.90]	14.24	3.48	[7.40, 21.08]	24.79	2.76	[19.38, 30.20]
Diffusion	Normative	T1	4.98	3.90	[-2.69, 12.64]	6.42	2.77	[0.99, 11.90]	5.15	3.48	[-1.68, 11.99]	6.47	2.75	[1.07, 11.90]
	Informative	T1	-0.93	3.90	[-8.59, 6.73]	11.02	2.77	[5.59, 16.50]	12.34	3.48	[5.51, 19.17]	15.65	2.75	[10.25, 21.10]
	Normative	T2	6.60	3.90	[-1.06, 14.27]	8.74	2.77	[3.31, 14.20]	7.97	3.48	[1.13, 14.80]	9.10	2.75	[3.69, 14.50]
	Informative	T2	0.30	3.90	[-7.36, 7.96]	15.19	2.77	[9.75, 20.60]	17.94	3.48	[11.11, 24.78]	20.61	2.75	[15.20, 26.00]
	Normative	T3	5.46	3.90	[-2.20, 13.13]	11.98	2.77	[6.54, 17.40]	10.45	3.48	[3.62, 17.29]	11.63	2.75	[6.23, 17.00]
	Informative	T3	-2.72	3.90	[-10.39, 4.94]	19.60	2.77	[14.17, 25.00]	21.11	3.48	[14.27, 27.94]	25.40	2.75	[20.00, 30.80]
	Normative	T4	2.09	3.90	[-5.57, 9.75]	15.49	2.77	[10.05, 20.90]	10.97	3.48	[4.13, 17.80]	14.00	2.75	[8.60, 19.40]
	Informative	T4	-8.40	3.90	[-16.06, -0.73]	24.21	2.77	[18.78, 29.60]	28.04	3.48	[21.20, 34.87]	31.77	2.75	[26.37, 37.20]

Table 6: Estimated marginal means (EMMs) of perceived compliance and perceived conversion by condition and task type.

		Perceived Compliance			Perceived Conversion		
Condition	Task type	*M*	SE	95% CI	*M*	SE	95% CI
Majority	Normative	2.21	0.17	[1.87, 2.55]	4.47	0.18	[4.13, 4.82]
Minority	Normative	1.90	0.17	[1.57, 2.23]	4.58	0.17	[4.24, 4.92]
Diffusion	Normative	1.81	0.17	[1.48, 2.14]	4.51	0.17	[4.17, 4.85]
Majority	Informative	2.54	0.17	[2.20, 2.88]	4.66	0.18	[4.31, 5.01]
Minority	Informative	2.43	0.17	[2.10, 2.76]	4.87	0.17	[4.53, 5.22]
Diffusion	Informative	2.33	0.17	[2.00, 2.66]	4.65	0.17	[4.31, 4.99]

Table 7: Estimated marginal means (EMMs) of agent perception by condition and task type.

		Competence		Predictability		Integrity		Understanding		Utility		Affect		Trust	
Condition	Task type	*M* (SE)	95% CI	*M* (SE)	95% CI	*M* (SE)	95% CI	*M* (SE)	95% CI	*M* (SE)	95% CI	*M* (SE)	95% CI	*M* (SE)	95% CI
Majority	Normative	4.34 (0.22)	[3.91, 4.76]	5.16 (0.21)	[4.75, 5.57]	4.07 (0.22)	[3.63, 4.50]	3.95 (0.25)	[3.46, 4.43]	3.68 (0.24)	[3.22, 4.14]	3.56 (0.24)	[3.09, 4.03]	3.40 (0.23)	[2.96, 3.85]
Minority	Normative	5.06 (0.21)	[4.64, 5.48]	5.66 (0.20)	[5.26, 6.06]	5.10 (0.22)	[4.67, 5.53]	5.02 (0.24)	[4.54, 5.49]	4.64 (0.23)	[4.19, 5.10]	4.65 (0.23)	[4.18, 5.11]	4.54 (0.22)	[4.10, 4.98]
Diffusion	Normative	4.67 (0.21)	[4.26, 5.09]	4.85 (0.20)	[4.45, 5.25]	4.66 (0.22)	[4.24, 5.09]	5.02 (0.24)	[4.55, 5.50]	4.20 (0.23)	[3.75, 4.65]	4.29 (0.23)	[3.83, 4.76]	4.05 (0.22)	[3.61, 4.49]
Majority	Informative	4.10 (0.22)	[3.67, 4.52]	4.97 (0.21)	[4.56, 5.38]	4.16 (0.22)	[3.73, 4.60]	3.87 (0.25)	[3.39, 4.36]	3.73 (0.24)	[3.26, 4.19]	3.51 (0.24)	[3.04, 3.98]	3.53 (0.23)	[3.08, 3.97]
Minority	Informative	5.46 (0.21)	[5.04, 5.87]	5.84 (0.20)	[5.44, 6.25]	5.21 (0.22)	[4.79, 5.64]	5.28 (0.24)	[4.81, 5.76]	4.93 (0.23)	[4.48, 5.39]	4.90 (0.23)	[4.44, 5.36]	4.79 (0.22)	[4.35, 5.23]
Diffusion	Informative	4.30 (0.21)	[3.88, 4.72]	4.56 (0.20)	[4.16, 4.96]	4.65 (0.22)	[4.23, 5.08]	4.74 (0.24)	[4.27, 5.22]	4.32 (0.23)	[3.86, 4.77]	4.23 (0.23)	[3.76, 4.69]	4.08 (0.22)	[3.65, 4.52]

C Supplementary Figures

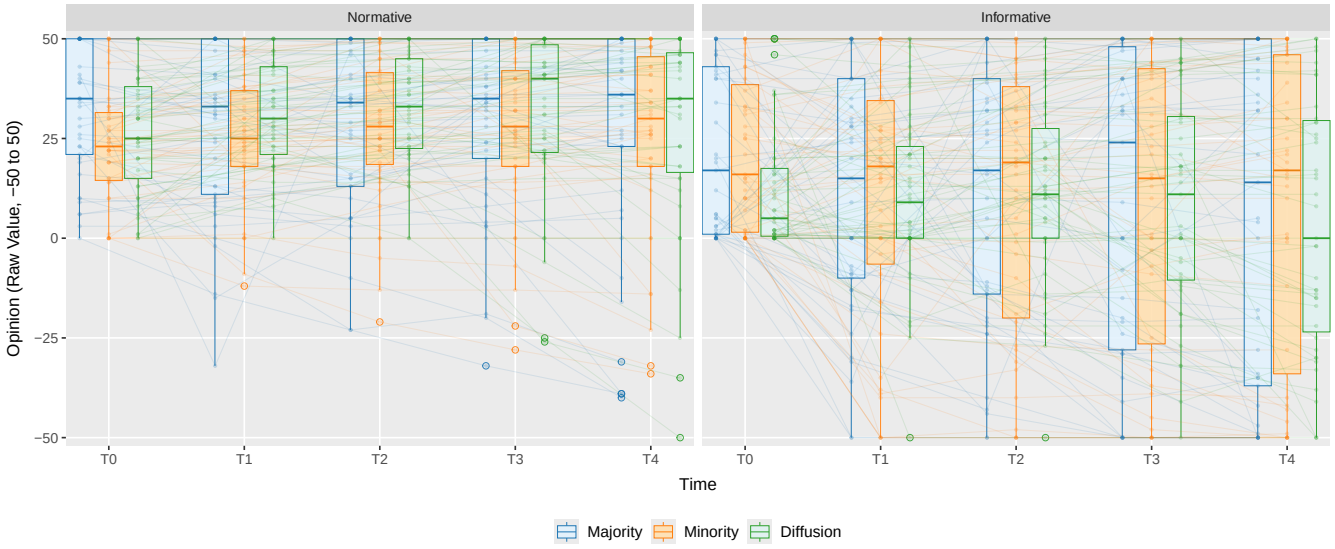


Figure 9: Raw opinion trajectories from T_0 to T_4 for normative and informative tasks across majority, minority, and diffusion conditions. Boxplots show the distribution of unadjusted opinion values (-50 to 50), and individual trajectories depict participant-level changes over time.

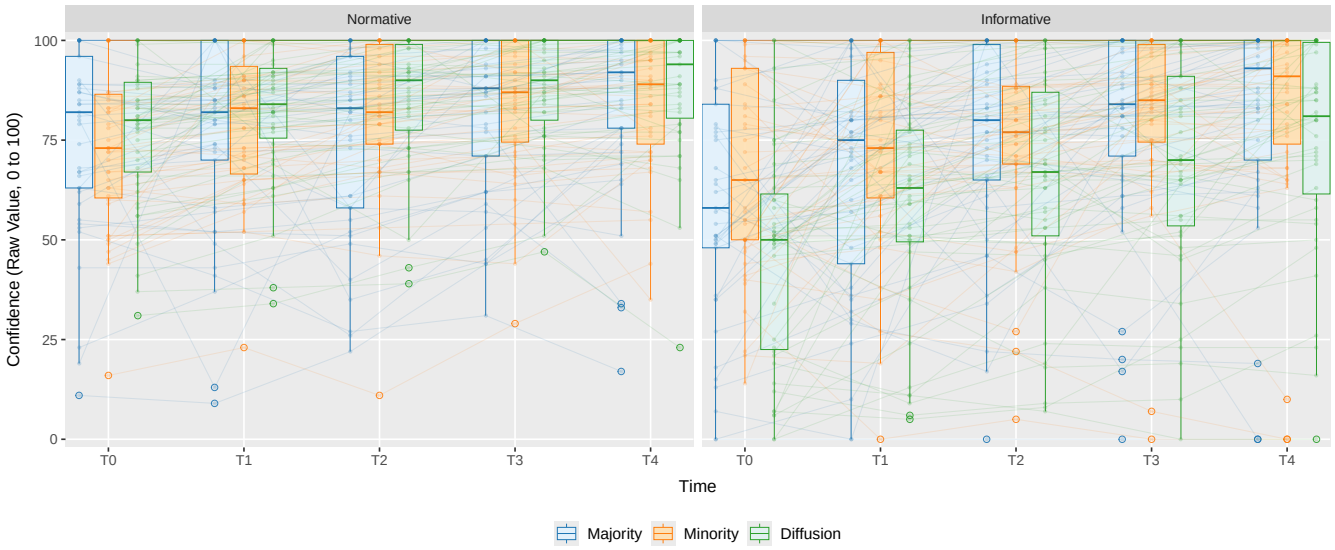


Figure 10: Raw confidence trajectories from T_0 to T_4 for normative and informative tasks across majority, minority, and diffusion conditions. Boxplots show the distribution of unadjusted confidence values (0–100), and individual trajectories depict participant-level changes over time.

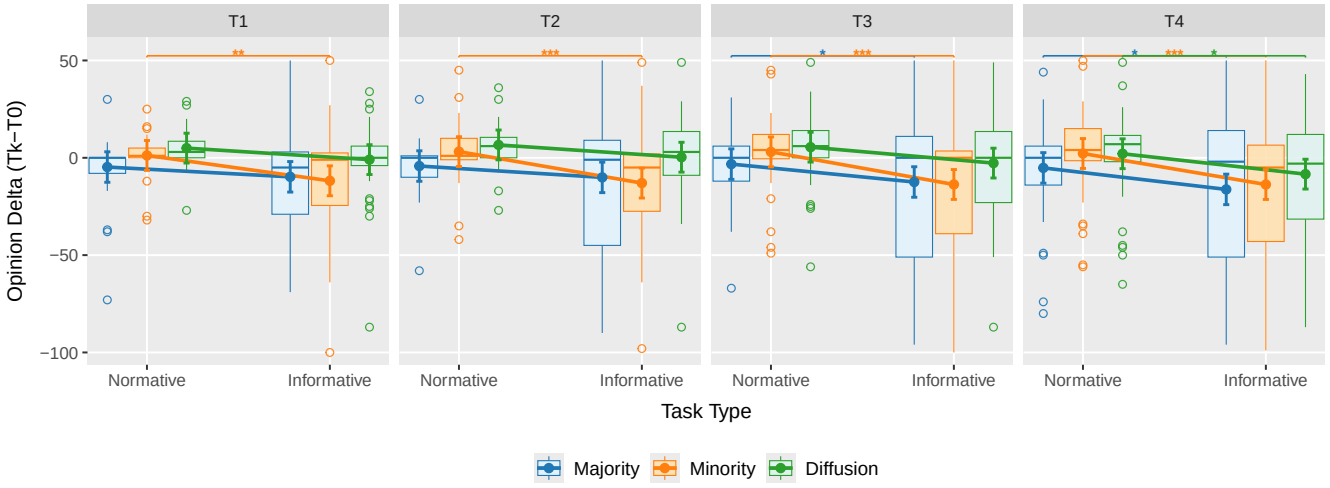


Figure 11: Signed opinion change by task type and time. Boxplots depict *directional opinion change* (ΔOpinion) relative to baseline (T_0), shown separately for normative and informative tasks across four time points (T_1–T_4). Negative values indicate movement toward the opposing stance advocated by the dissenting agent(s) (i.e., the counter-position relative to the participant's baseline), whereas positive values indicate movement away from that stance (i.e., reinforcement of one's initial stance). Majority, Minority, and Diffusion conditions are compared, with significance brackets indicating statistically reliable differences. This figure presents the signed-value version of the data corresponding to Figure 6-(a), disaggregated by time and task type for clarity.

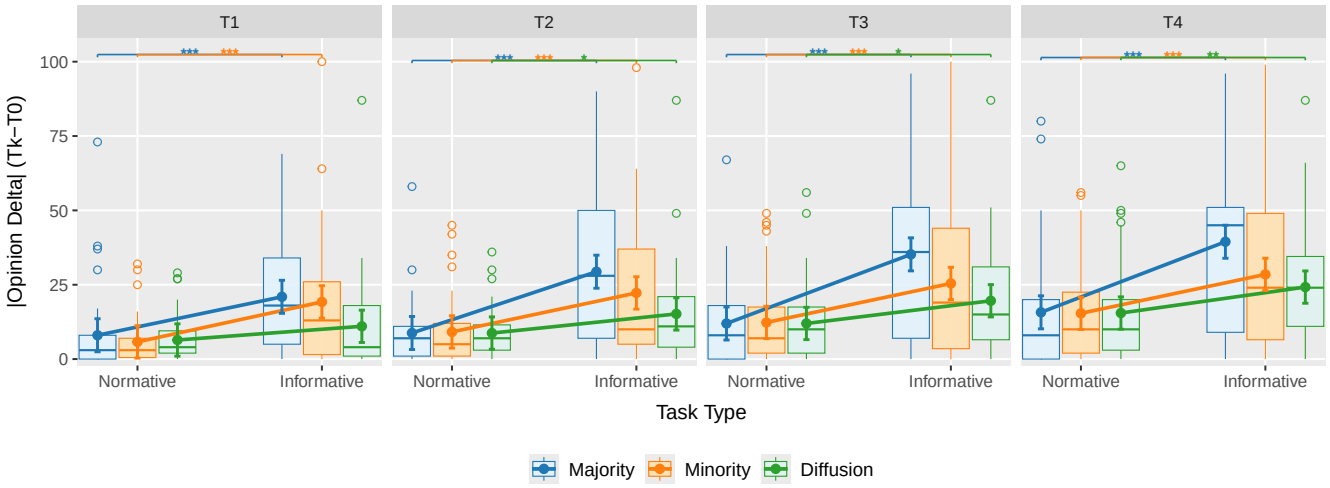


Figure 12: Absolute opinion change by task type and time. Boxplots depict *absolute opinion change magnitudes* (| ΔOpinion |) relative to baseline (T0), shown separately for normative and informative tasks across four time points (T1–T4). Larger values indicate greater deviation from baseline regardless of direction. Majority, Minority, and Diffusion conditions are compared, with significance brackets indicating statistically reliable differences. This figure visualizes the absolute-value version of the data presented in Figure 6-(b), disaggregated by time and task type for improved interpretability.

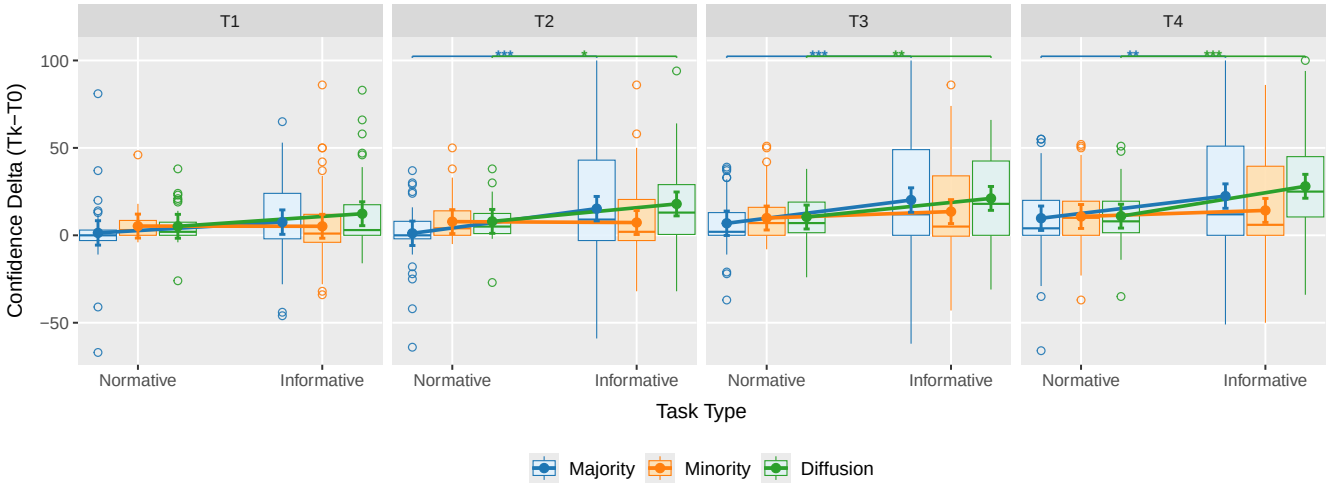


Figure 13: Signed confidence change by task type and time. Boxplots depict *directional confidence change* (ΔConfidence) relative to baseline (T_0), shown separately for normative and informative tasks across four time points (T_1–T_4). Positive values indicate increases in confidence relative to baseline, whereas negative values indicate decreases. Majority, Minority, and Diffusion conditions are compared. This figure presents the signed-value version of the data corresponding to Figure 6-(c), disaggregated by time and task type as a supplementary visualization.

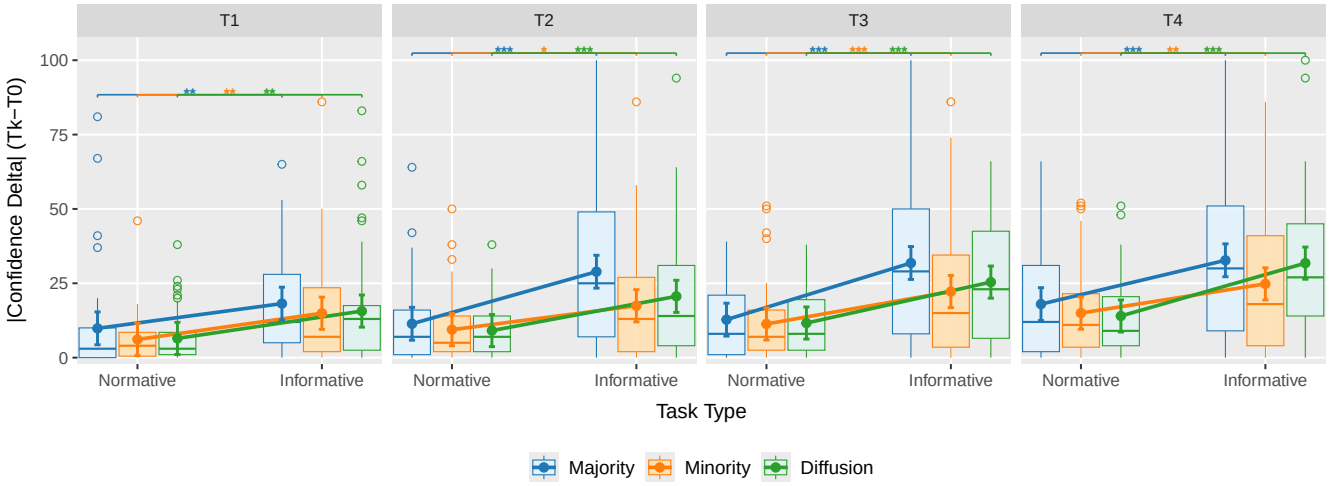


Figure 14: Absolute confidence change by task type and time. Boxplots depict *absolute confidence change magnitudes* (| ΔConfidence |) relative to baseline (T0), shown separately for normative and informative tasks across four time points (T1–T4). Larger values indicate greater deviation from baseline regardless of whether confidence increased or decreased. Majority, Minority, and Diffusion conditions are compared. This figure presents the absolute-value version of the data shown in Figure 6-(d), disaggregated by time and task type as a supplementary visualization.

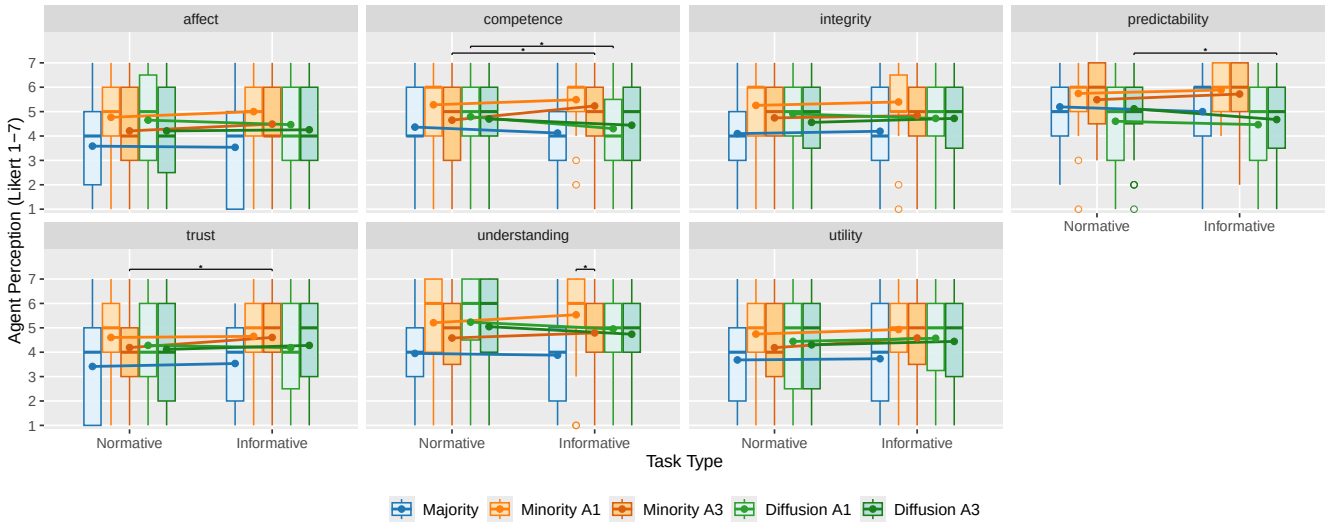


Figure 15: Exploratory analysis of agent perception by individual agent roles. Boxplots show ratings on seven dimensions (affect, competence, integrity, predictability, trust, understanding, and utility) for Majority agents, and separately for Minority and Diffusion agents split into A1 (initially supportive agents) and A3 (the consistently dissenting agent). This figure supplements Figure 8, where ratings were averaged across agents.