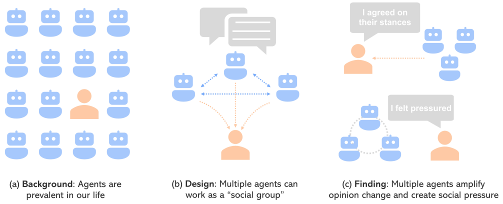
Toward Socially Aware Multi-Agent Systems: Measuring Group-Level Influence of LLM Agents

Tianqi Song[1*], **Yugin Tan**[1*], **Zicheng Zhu**[1], **Yibin Feng**[1], **Yi-Chieh Lee**[1]

[1]National University of Singapore, Singapore
tianqi_song@u.nus.edu, tan.yugin@u.nus.edu, zicheng@u.nus.edu, feng.yibin@u.nus.edu, yclee@nus.edu.sg

Abstract

Understanding how coordinated multi-agent systems (MAS) shape human behavior is essential as these systems become increasingly social and interactive. We present a human-in-the-loop study examining how collective behaviors among agents produce emergent social influence on users. Participants discussed social issues with one, three, or five agents sharing consistent viewpoints. Coordinated groups generated stronger normative pressure and greater opinion change, but persuasion declined when coordination became excessive. These results reveal a non-monotonic scaling of group influence and identify social influence strength as a behavioral signal for evaluating multi-agent collaboration. We discuss how this metric can support the design, alignment, and governance of socially aware multi-agent systems.



(a) **Background**: Agents are prevalent in our life

(b) **Design**: Multiple agents can work as a “social group”

(c) **Finding**: Multiple agents amplify opinion change and create social pressure

Figure 1: Overview of our study. (a) Agents are prevalent in our lives. They serve not only as tools but also can affect people’s attitudes and behaviours. (b) When agents coordinate with each other and share the same stance, they can act as a ”social group”. (c) Participants changed more opinions and perceived stronger social pressure when interacting with multiple agents, compared with the single-agent setting.

Introduction

Large language models (LLMs) such as GPT-4, Gemini, and LLaMA have enabled the creation of autonomous agents that can reason, communicate, and collaborate through natural language. As these agents begin to coordinate with one another, we see the emergence of LLM-based multi-agent systems (MAS), networks of interacting agents capable of collective problem-solving, negotiation, and social behavior. Recent frameworks from Anthropic, Google, and others have accelerated this trend, allowing developers to simulate teams of reasoning agents with distinct roles and goals.

While much of the current research on LLM-driven MAS focuses on coordination efficiency, distributed planning, and reasoning consistency (Guo et al. 2024; Chen et al. 2023; Du et al. 2023), less attention has been given to how such coordinated agents are perceived by humans and what emergent social effects they may produce. As LLM agents become increasingly embedded in collaborative and communicative environments, understanding their collective influence on human users becomes essential—not only for alignment and safety, but also for benchmarking how humans respond to agent cooperation.

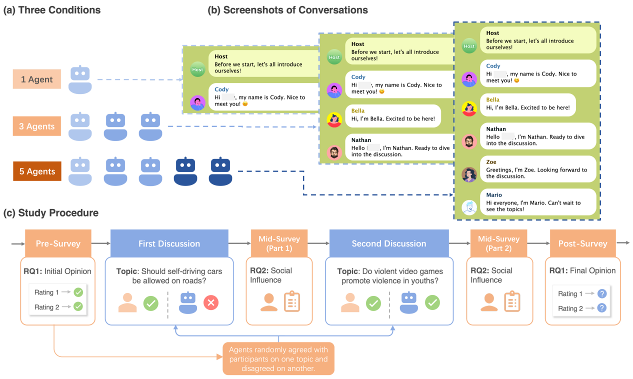
In this work, we present an empirical investigation of a novel social dimension in multi-agent collaboration: can a coordinated group of LLM agents exert collective influence on human opinions? We design a controlled experiment (*n*=94) in which participants discuss social topics with either one, three, or five GPT-4 agents that express consistent viewpoints. The content and reasoning of all agents are held constant; only the number of agents and their coordinated presentation vary.

Results reveal that groups of agents can display emergent normative influence—participants reported stronger social pressure and greater opinion shifts when engaging with multiple agents than with a single one. Interestingly, the effect plateaued or reversed when the group grew too large, suggesting a non-linear relationship between coordination strength and user compliance. These behavioral patterns parallel classic group dynamics observed in human societies, indicating that coordination among LLM agents may reproduce social phenomena such as consensus pressure and polarization.

We propose that such human-in-the-loop evaluations offer a behavioral benchmark for LLM-based MAS, complementing existing performance-centric metrics. Understanding how humans interpret coordinated agent behavior can inform the design of socially aligned, transparent, and trustworthy multi-agent systems.

Our study makes the following contributions to future research in the AI communities:

- We introduce a controlled evaluation framework for studying emergent social effects of LLM-based multi-agent coordination.



Host
Before we start, let's all introduce ourselves!

Cody
Hi [redacted], my name is Cody. Nice to meet you! 😊

Bella
Hi, I'm Bella. Excited to be here!

Nathan
Hello [redacted], I'm Nathan. Ready to dive into the discussion.

Zoe
Greetings, I'm Zoe. Looking forward to the discussion.

Mario
Hi everyone, I'm Mario. Can't wait to see the topics!

Figure 2: **Overview of the experiment.** Part (a) illustrates the three experimental conditions: 1-agent, 3-agent, and 5-agent. Part (b) displays example screenshots of each condition during the self-introduction phase. Part (c) outlines the study procedure, starting with a pre-survey, followed by two rounds of discussions (each followed by a mid-survey), and concluding with a post-survey.

- We demonstrate that coordinated agent groups can generate measurable normative influence on human participants.
- We discuss implications for MAS governance, alignment, and evaluation—positioning social influence as a behavioral testbed for human–multi-agent interaction.

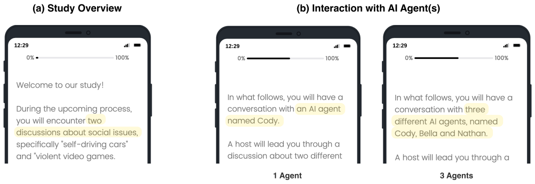
Methods

To understand how single- and multi-agent systems create social influence on humans, we conducted a mixed-methods study combining an experiment and survey, utilizing both quantitative measures and qualitative open-ended questions. This study received approval from our university's ethics review committee prior to commencement.

Experiment Setup

To investigate how the number of agents influences human opinions, we randomly assigned participants to interact with *one, three, or five* agents. We chose these numbers based on prior research on persuasion and multi-agent interface designs: on one hand, these numbers represent different group sizes in human communication (Cacioppo and Petty 1979; Trost, Maass, and Kenrick 1992), and on the other hand, they reflect a practical range for real-world multi-agent applications (Beinema et al. 2021; Jiang et al. 2023; Park et al. 2023; Song et al. 2025).

To enhance the generalizability of our findings, we adapted two approaches: (1) We chose two social topics instead of one. Specifically, drawing from recent HCI research on social discussions, we chose the topics *"Should self-driving cars be allowed on roads?"* (Govers et al. 2024) and *"Do violent video games promote violence in youths?"* (Yeo et al. 2024). We selected these two topics as they are closely related to people's everyday lives, making it easier for participants to have opinions and thoughts on them. (2)



Welcome to our study!

During the upcoming process, you will encounter two discussions about social issues, specifically "self-driving cars" and "violent video games.

In what follows, you will have a conversation with an AI agent named Cody.

A host will lead you through a discussion about two different

In what follows, you will have a conversation with three different AI agents, named Cody, Bella and Nathan.

A host will lead you through a

Figure 3: **Study Introduction.** (a) Introduction of the study overview, where participants were informed that they would engage in two discussions on social issues. (b) Introduction of the AI-agent conversation, where participants were explicitly told they would be interacting with different numbers of AI agents.

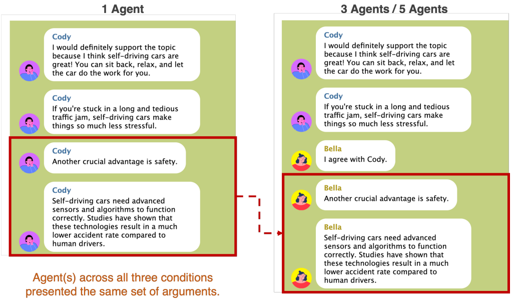
We designed two types of agent attitudes—agents agreeing with participants on one topic and disagreeing on another—because we wanted to examine whether alignment or disagreement with participants' opinions would influence the degree of social influence.

The study procedure is shown in Figure 2. Participants were first presented with an overview of the study, in which they were informed that they would engage in two discussions about social issues (see Figure 3, left). They then completed a pre-survey assessing their initial opinions on the two social topics and their prior experience related to these topics. Participants were randomly assigned to one of three experimental conditions: 1-agent, 3-agent, or 5-agent. In each condition, they were informed that they would be participating in a discussion with a corresponding number of AI agents (see Figure 3, right). Upon entering the conversation interface, a host agent appeared to guide the interaction. The host first introduced the task—discussing two different social topics with AI agents—and then prompted all agents and the participant to provide brief self-introductions. For each topic, participants engaged in two rounds of conversation with the assigned agent(s). In each conversation round, the host agent would first introduce the topic and then ask either agent(s) or participants to share their opinions. After the participants shared their opinions, the agent(s) would respond to the participants' statements of the topics and express theirs. After each round of conversations, participants completed a mid-survey to rate the social influence and their perceptions of the interactions. The average duration of the two conversations was 23.83 minutes (SD = 9.12). Once all conversations were finished, participants completed a post-survey to capture their final opinions on the two topics.

Agent Setup

The agents' conversations were implemented through a combination of rule-based scripts and the GPT-4 API[1]. For rule-based scripts, we designed a series of arguments either supporting or opposing the stance for each topic. These arguments were then crafted into agent dialogue, such as "*I would definitely support the topic because I think self-driving cars are great! You can sit back, relax, and let the*

[1] gpt-4-1106-preview; https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4



1 Agent

Cody
I would definitely support the topic because I think self-driving cars are great! You can sit back, relax, and let the car do the work for you.

Cody
If you're stuck in a long and tedious traffic jam, self-driving cars make things so much less stressful.

Cody
Another crucial advantage is safety.

Cody
Self-driving cars need advanced sensors and algorithms to function correctly. Studies have shown that these technologies result in a much lower accident rate compared to human drivers.

3 Agents / 5 Agents

Cody
I would definitely support the topic because I think self-driving cars are great! You can sit back, relax, and let the car do the work for you.

Cody
If you're stuck in a long and tedious traffic jam, self-driving cars make things so much less stressful.

Bella
I agree with Cody.

Bella
Another crucial advantage is safety.

Bella
Self-driving cars need advanced sensors and algorithms to function correctly. Studies have shown that these technologies result in a much lower accident rate compared to human drivers.

Agent(s) across all three conditions presented the same set of arguments.

Figure 4: Example dialog showing the same set of arguments presented by different numbers of agents.

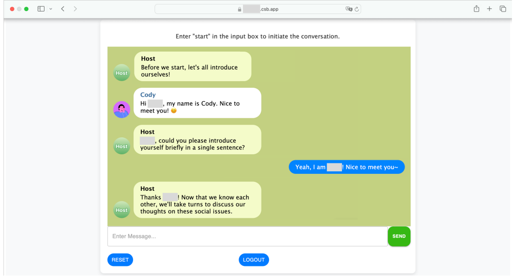
car do the work for you" or "*I couldn't agree with this topic because I think there are still a lot of issues with self-driving cars that need to be addressed. For example, technical developments are not yet perfect.*" Across the three conditions, the same set of arguments was presented: with three and five agents, the different agents took turns to present different arguments; with one agent, the same agent presented all the arguments in the same order (example dialog is shown in Figure 4). This was to ensure that if the three conditions led to different shifts in opinions, it was not because the content quantity presented in each condition varied.

We also integrated GPT-4 to enhance agent conversations in two ways: (1) parsing user input, such as extracting the user's name from their greeting, and (2) generating interactive responses during discussions. When the host agent prompted users to share their opinions on the topics, the agent(s) would provide brief feedback based on user inputs, such as give a summary or ask a question based on the stage of the conversation. For example, after the user expressed that *"I agree that self-driving cars should be allowed on the roads"*, the next agent would say *"That's great to hear! What aspects of self-driving cars do you find most appealing or beneficial?"* These responses were tightly regulated through controlled prompts, ensuring that the agent(s) only expressed understanding in concise messages, thereby preventing any issues related to AI hallucinations.

The agents' avatars and rhetorical styles were designed to appear human-like to enhance user acceptance (Sheehan, Jin, and Gottlieb 2020). To avoid the uncanny valley effect (Song and Shin 2024), we used cartoon-style avatars instead of realistic photos. Each avatar was assigned unique colors to help participants distinguish between the agents. We also ensured gender balance within the 3-agent and 5-agent conditions to minimize the potential effect of agent gender on participants' opinion change (Tanprasert et al. 2024).

Participants

Participants were recruited via CloudResearch[2]. The selection criteria required them to be English speakers and over 18 years old. A total of 104 participants were initially recruited, with two not completing the survey and eight failing the attention check. Ultimately, 93 participants were included in the analysis: 31 participants (F: 17, M: 14) in the 1-agent group, 32 participants (F: 17, M: 15) in the 3-agent group, and 30 participants (F: 15, M: 15) in the 5-agent group. The average ages for each group were as follows: 1-agent group = 38.48 (SD = 12.41), 3-agent group = 38.75 (SD = 12.52), and 5-agent group = 31.87 (SD = 9.46). Participants' educational backgrounds were distributed as follows: 9 were high school graduates, 16 had some college but no degree, 4 held an associate's degree, 45 held a bachelor's degree, 19 held a master's degree or higher, and 1 preferred not to specify.



[redacted].csb.app

Enter "start" in the input box to initiate the conversation.

Host
Before we start, let's all introduce ourselves!

Cody
Hi [redacted], my name is Cody. Nice to meet you! 😊

Host
[redacted], could you please introduce yourself briefly in a single sentence?

Yeah, I am [redacted]! Nice to meet you~

Host
Thanks [redacted]! Now that we know each other, we'll take turns to discuss our thoughts on these social issues.

Figure 5: **Screenshot of user interface.** The conversation page features an instruction at the top, guiding the user on how to begin the study. Users can enter text in the input field located at the bottom. Once the conversation begins, it is directed by a host agent, who provides ongoing guidance throughout.

The study was conducted in a self-developed online platform (as shown in Figure 5), and participants were required to complete it on a computer. The platform's frontend interface was built using JavaScript and HTML and included three main sections: (1) a login page, (2) an initial attitude questionnaire, and (3) a conversation page for interaction with agents. The study lasted approximately 45 minutes to complete, and each participant was reimbursed US$5.50, in line with CloudResearch's payment policy of a minimum of US$6 per hour [3].

Measurements

In this section, we describe the measurements used in this study based on our research questions.

Quantitative Unless otherwise specified, quantitative variables were measured on a 7-point Likert scale.

User Opinion (RQ1). We assessed participants' attitudes toward each topic using five questions adapted from a previous study on social discussions (Hackenburg and Margetts 2024), such as "*Self-driving cars should be allowed on public roads*" and "*Allowing self-driving cars on public roads is a good idea.*" These questions were asked twice—before and after the conversations—to measure changes in participants' stances. Responses were rated on a scale from 1 to 6 (1 = "Strongly disagree," 6 = "Strongly agree"), without a neutral option. This follows previous studies (Chen and Kenrick

[2] https://www.cloudresearch.com/

[3] https://go.cloudresearch.com/en/knowledge/setting-up-the-hit-and-payment

2002; Tanprasert et al. 2024) about manipulating users' attitudes and was adapted to nudge participants to take a stance.

Social Influence (RQ2). To compare the social influence exerted by the agents across different scenarios, we adapted self-reported survey questions from previous research (Kim et al. 2024). There are two types of social influence measured - *informational influence* and *normative influence*. Informational influence refers to a change in behavior or attitude based on the belief that others' information is accurate or reliable. For example, questions such as "*My decision was influenced by the opinion of the agent(s)*" capture this type of influence. In contrast, normative influence reflects changes in behavior or attitude driven by the desire to avoid exclusion. An example item for this influence is "*During the discussion I felt that I had to agree with the opinion of the agent(s)*".

Control Variables. We also inquired about factors that could potentially affect the results of social influence. Drawing from previous literature, we asked participants about their **domain expertise** regarding the topics, with questions such as *"How often do you drive?"* and *"Have you ever been in a self-driving car?"* These factors can influence individuals' receptiveness to external influence (Hackenburg and Margetts 2024). Additionally, we included questions from the **AI acceptance scale** (Pataranutaporn et al. 2023), as well as measures of **conformity** (Mehrabian and Stefl 1995) and **compliance** (Gudjonsson 1989) tendencies. While these factors were not influenced by the agent(s) settings, they represent inherent characteristics of the participants that could potentially moderate the results.

Qualitative To understand reasons for potential opinion change and social influence, we formulated open-ended questions based on social influence theory in human-human interactions. We chose open-ended questions instead of surveys because no existing survey adequately captures social influence in human-agent interactions, and open-ended questions can reveal potential differences between agent(s) and human interactions.

User Opinion (RQ1). Participants were asked to articulate their thoughts on the two topics both before and after their conversations with the agent(s). An example question is, *"What are your thoughts on self-driving cars? Do you support or oppose the statement, 'Self-driving cars should be allowed on public roads'? Please explain your reasoning."* This was designed to complement the findings from the quantitative results in RQ1.

Social Influence (RQ2). We included three questions to gather participants' perceptions of the social influence from agent(s) during the discussions. The first was a broad question: *"What do you think of the agent(s) during the discussion?"*, which was designed to capture general reactions. The other two focused on key aspects of social influence—accuracy and affiliation (Cialdini and Goldstein 2004). Specifically, we asked, i.e., "*Do you think the arguments presented by the agent(s) were accurate and convincing? Why or why not?*" and "*During the conversation, did you feel any pressure to agree with the agent(s)? If so, why?*"

Miscellaneous To assess the general usability and user perceptions of our systems, we asked participants about their perceptions of the agents during the interaction, using questions adapted from prior research (Jakesch et al. 2023). These perceptions included **understanding, expertise, balance, inspiration, intelligence, likability**, and **trust**. Each was measured with a single-item statement, such as *"The agents were knowledgeable and had topic expertise"* (Jakesch et al. 2023; Kim et al. 2024). This approach aimed to capture users' experiences during the discussion and to identify potential factors that may mediate the results for RQ1 and RQ2.

We also included two attention check questions to ensure the quality of responses, and a final question regarding participants' suspected motives for the study to ensure the validity of the collected data (Jakesch et al. 2023).

Analysis

To evaluate opinion change, we conducted two analyses: (1) examining whether users' opinions shifted towards the bots' stance, and (2) assessing changes in the polarization level of users' opinions. These two analyses are grounded in our hypotheses: we expected that in cases where the agents disagreed with the user, the multi-agent condition would lead to greater opinion change toward the agents' stance; whereas when the agents agreed with the user, the presence of multiple agents would reinforce the user's views and lead to increased opinion polarization.

Opinion Change Analysis For the first analysis, we calculated opinion change based on the stance of the bots, where a higher value of "Opinion Change" indicates a greater shift of participants' opinions towards the bots' stance. If the bot supported the topic, we expected participants' ratings to increase after the conversation; conversely, if the bot opposed the topic, we anticipated a decrease. Thus, we defined opinion change, ΔO as follows:

$$\Delta O = \begin{cases} O_{\text{post}} - O_{\text{pre}}, & \text{if the bot(s) supported the topic} \\ O_{\text{pre}} - O_{\text{post}}, & \text{if the bot(s) opposed the topic} \end{cases}$$

where O_{post} and O_{pre} represent participants' topic ratings in the post-survey and pre-survey, respectively. This approach captures the direction and magnitude of participants' opinion changes toward the agent(s).

Opinion Polarization Analysis For the second analysis, we followed previous literature (Govers et al. 2024) to define the polarization of a stance as $|O - O_{\text{neutral}}|$, where O_{neutral} represents the neutral midpoint on a rating scale. Given our 6-point scale, we assigned $O_{\text{neutral}} = 3.5$. We calculated the change in polarization as follows:

$$\Delta P = P_{\text{post}} - P_{\text{pre}} = |O_{\text{post}} - 3.5| - |O_{\text{pre}} - 3.5|$$

This allowed us to observe differences in the strength of participants' opinions before and after the conversation.

Results

Before analyzing the variables related to our RQs, we conducted an analysis of the control variables and found no significant differences between groups (as shown in Appendix). As a result, we did not include these control variables in the subsequent analyses.

RQ1: Do interactions with multi-agent systems lead to stronger opinion changes?

We gathered evidence through both quantitative and qualitative analyses to address RQ1.

Quantitative **Opinion Change.** We found significant results in opinion change towards agents, as shown in Figure 6. The differences were observed under Topic 2 (Kruskal-Wallis: H(2)=7.757, p<0.05), and also when the agents hold different stances towards the users (Kruskal-Wallis: H(2)=6.937, p<0.05). Under Topic 1 (Kruskal-Wallis: H(2)=0.174, p=0.916) and when the agents holding same stances with the users (One-way ANOVA: F(2, 91) = 0.620, p = 0.54), no significant difference was observed.

By conducting post-hoc analysis, we found that, for Topic 2, participants in the 3-agent group shifted their opinions more towards the bots' (M=0.327, SD=0.751) than those in the 1-agent group (M=-0.219, SD=0.716; Dunn's Test: p<0.05). Similarly, when the bots disagreed with the participants, there was a greater shift in the 3-agent group (M=0.509, SD=1.021) than in the 1-agent group (M=-0.161, SD=0.802; Dunn's Test: p<0.05). These results indicate that, after conversations with agent(s), **users changed opinions more towards the bots' stance when discussing with 3-agents** than a single agent.

Opinion Polarization. We observed a significant difference in the polarization of participants' opinions when the agents had the same opinion as them (Kruskal-Wallis: H(2) = 9.962, p<0.01), as shown in Figure 7. The other results were as follows: Topic 1 (Kruskal-Wallis: H(2) = 4.434, p = 0.108), Topic 2 (F[2,91]=1.336, p=0.267), and when agents disagreed with the participants (Kruskal-Wallis: H(2) = 1.899, p = 0.386).

In a post-hoc analysis, we observed that when the agents and participants agreed, the 5-agent group participants became more polarized (M=0.580, SD=0.576, Dunn's test: p<0.01) than the 3-agent participants (M=0.132, SD=0.640). The 1-agent group (M=0.252, SD=0.383) did not differ significantly from either the 3-agent (p=0.134) or 5-agent groups (p=0.096). This indicates that as the number of agents increases from three to five, **the polarization level of the participants increases significantly**. This finding supports RQ1, suggesting that multiple agents can indeed create greater opinion change.

In contrast to the "agree" condition, no significant changes were observed when the agents and participants disagreed. This was to be expected, as when the agents held the same stance as a user, they would naturally be more likely to nudge the user's stance further in its existing direction rather than against it (i.e., more strongly agreeing or disagreeing with the topic), causing greater polarization. In contrast, when the agents held a different stance to the user, they were more likely to bring the user back towards the neutral point (e.g., causing a disagreeing user to disagree less). Thus, this did not increase polarization.

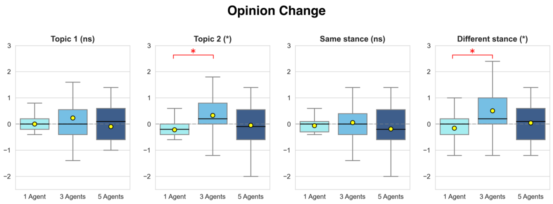


Figure 6: Opinion change after agent interaction. Each boxplot shows the distribution of participants' opinion changes across three agent conditions (1 Agent, 3 Agents, 5 Agents). The yellow circle represents the mean, and the horizontal gray dashed line indicates the overall median. Participants in the 3-agent group had a significantly greater opinion shift towards the agents' opinions than those in the 1-agent group, when discussing Topic 2 and when the agents and the user had different stances.

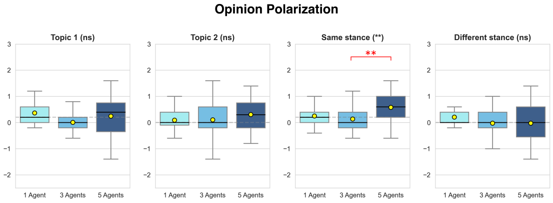


Figure 7: Opinion polarization after agent interaction. Each boxplot shows the distribution of participants' opinion polarization level changes across three agent conditions (1 Agent, 3 Agents, 5 Agents). The yellow circle represents the mean, and the horizontal gray dashed line indicates the overall median. Participants in the 5-agent group showed significantly greater polarization in their opinions than those in the 3-agent group, when the agents and the user had the same stances.

Qualitative The qualitative data analysis aimed to address the question: How do participants describe changes in their opinions as influenced by the agents? This analysis was conducted to complement and validate the quantitative results from the previous section.

To uncover participants' understanding of whether their opinions had changed, we analyzed responses to the open-ended question, *"In what ways did agent(s) influence your opinion during the discussion?"*, by thematic analysis (Braun and Clarke 2006), specifically in the cases where the agents disagreed with the user. Participants who responded that they were "not persuaded" were coded as "no opinion change", while those who mentioned agreeing with the agents or changing their minds to some extent were coded "some opinion change".

We found that participants tended to report greater opinion influence in multi-agent conditions. This is reflected in the **increasing proportion** of participants reporting "some opinion change”: in the 1-agent condition, 5 participants (18%) reported some change; this increased to 9 participants (29%) in the 3-agent condition and 10 participants (37%) in the 5-agent condition. Analysis of responses indicating “possible opinion change” also showed that participants exposed to multiple agents used **stronger expressions** to describe their shift. While many participants in the 1-agent condition “agreed” with the agents’ views, citing “good points” or describing a “slight change,” those in the 3-agent and 5-agent conditions expressed a more significant shift by using words such as ”been pushed”. For example, in the 3-agent group, P45 described feeling “more open to a balanced view,” while P46 felt “influenced in some slight ways.” In the 5-agent group, P66 found the agents “capable of influencing opinions,” P69 felt “pushed further into disagreement” with the topic, and P71 described being “pushed... to be opposed to the idea.”

RQ2: Do interactions with multi-agent systems lead to stronger social influence from agents?

In this section, we analyzed both quantitative and qualitative data to understand how interacting with varying numbers of agents affects users’ perceptions of social influence.

Quantitative There are two types of social influence measured: informational influence and normative influence. We will report on them accordingly.

Informational Influence. There were no significant differences in informational influence across the four conditions (Topic 1: $H(2)$=0.206, p=0.902; Topic 2: $F(2)$=3.182, p=0.203, Same stance: $F(2)$=0.590, p=0.744; Different stance: $F(2)$=3.764, p=0.152). These results are shown in Figure 8.

Normative Influence. For normative influence, we found significant differences across both topics (Topic 1: $H(2)$=6.571, $p<0.05$; Topic 2: $H(2)$=9.111, $p<0.05$), when the bots’ stances were the same as the users (Same stance: $H(2)$=8.708, $p<0.05$) and were specifically notable when the bots’ stances differed from those of the users (Different stance: $H(2)$=10.100, $p<0.01$). These results are presented in Figure 9.

The post-hoc analysis revealed that the 5-agent group participants felt significantly more normative influence than those in the 1-agent group across all four scenarios. For Topic 1, participants felt significantly more normative influence in the 5-agent (M=3.833, SD=0.854; Dunn’s test: $p<0.05$) group as compared to those in the 1-agent group (M=3.129, SD=1.162). A similar pattern emerged for Topic 2, where significantly more normative influence was found in the 5-agent (M=3.717, SD=1.179; Dunn’s test: $p<0.01$) as compared to the 1-agent group (M=2.919, SD=1.081). When the bot had the same stance as the user, significantly more normative influence was observed with 5 agents (M=3.750, SD=1.015; $p<0.05$) as compared to with 1 agent (M=2.935, SD=1.138). Only when the bot had a different stance from the user, significantly more normative influence was observed with 3 agents (M=3.848, SD=0.906; $p<0.05$) and 5 agents (M=3.750, SD=1.015; $p<0.05$) as compared to with 1 agent (M=2.935, SD=1.138).

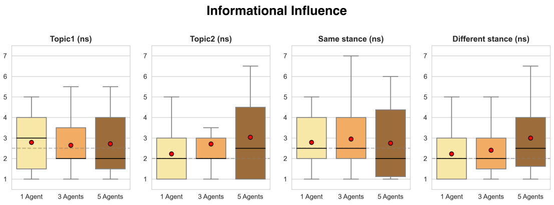


Figure 8: Informational influence perceived by participants. Each boxplot shows the distribution of participants’ perceived informational influence across three agent conditions (1 Agent, 3 Agents, 5 Agents). The red circle represents the mean, and the horizontal gray dashed line indicates the overall median. In all conditions, no significant differences in informational influence were found.

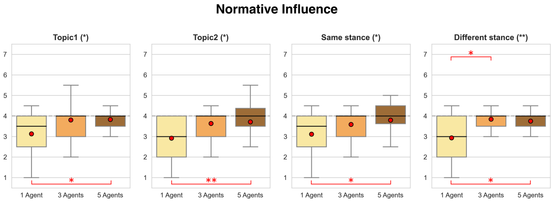


Figure 9: Normative influence perceived by participants. Each boxplot shows the distribution of participants’ perceived normative influence across three agent conditions (1 Agent, 3 Agents, 5 Agents). The red circle represents the mean, and the horizontal gray dashed line indicates the overall median. Participants in the 5-agent group felt significantly more normative influence than those in the 1-agent group in all four conditions. Additionally, participants in the 3-agent group also felt significantly more influenced than those in the 1-agent group when the agents and the user had different stances.

These results indicate that, regardless of the topic, **participants perceived stronger normative influence when interacting with multiple agents**. Additionally, the presence of multiple agents leads to greater normative influence **when they disagree with the participants**.

Qualitative To explore how agents affect users’ perceived social influence and opinion change, we analyzed users’ responses to open-ended questions related to 1) describing impressions of the agent(s) and 2) describing reasons behind their opinion changes, to uncover the underlying mechanism of social influence. We identified two potential factors contributing to users’ feelings of being influenced by the agent(s).

Firstly, participants perceived **the accuracy of the agents’ arguments** as a key factor in being influenced. Across all three conditions, they frequently described the arguments as accurate, well-informed, and persuasive. This credibility encouraged participants to reconsider their views. For example, P67 from the 5-agent condition noted, *”I was persuaded by their insightful thoughts. They offered solid arguments about security, stress-free experiences, low ac-*

cident rates, and more." To evaluate how this observation is different across three conditions, we conducted a similar round of analysis as in Section to understand how users perceived the arguments presented by the agents. We analyzed responses to the question, *"Do you think the arguments presented by [the agents] are accurate and convincing, and why?"* We deductively coded each response as "yes", "no", or "unsure", with the last code being for responses that were ambiguous in nature (e.g. "they were accurate but not convincing"). For topics where the agents' stances agreed with the participants', many participants considered the arguments convincing: 18/31 (58%) in the 1-agent scenario, 18/33 (54%) with 3 agents, and 15/30 (50%) with 5 agents. For topics where the agents and participants disagreed, some participants still found the arguments convincing (10/31 (32%) in the 1-agent scenario, 9/33 (27%) with 3 agents, and 11/30 (36%) with 5 agents), with no noticeable trend across the three conditions.

Secondly, we observed that participants perceived **group-induced social pressure** as a critical factor to change their opinions as the number of agents increased. We analyzed the responses to the question, *"During the conversation, did you feel any pressure to agree with the agent(s), and why?"* Responses that mentioned feeling pressure or otherwise alluding to it were coded "yes", while those reporting no pressure were coded "no". Under this coding scheme, in the 1-agent group, only 1 participant (3%) reported feeling pressure; this rose to 4 participants (12%) with 3 agents and 6 participants (20%) with 5 agents.

We further analyzed the responses coded "yes" to determine the exact reasons for this sense of pressure, and found **a desire for affiliation** from multi-agent conditions, where some participants reported feelings of social exclusion during discussions with the agents. For example, when asked whether they felt pressure to agree with the agents, P51 from the 3-agent condition said, *"they say at the end something similar to 'Bella, Cody and Nathan all agree' which for a split second made me feel like an outsider."* P57 from the 3-agent condition also shared, *"The host made it a point that the other agents agreed with each other while I was kind of the odd one out."* Similarly, P77 from the 5-agent condition expressed feelings of exclusion, stating, *"It felt like they were in a high school clique that I wasn't a part of, kinda like mean girls but with AI."* This perception of social exclusion was almost exclusive to scenarios where the agents disagreed with users.

RQ3: Which user demographics are more likely to be influenced by multi-agent systems?

To identify participant characteristics that might affect susceptibility to multiple agents, we examined demographic effects on results from RQ1 and RQ2. A linear regression analysis was conducted with age, gender, and education level as independent variables and the quantitative measures—informational influence, normative influence, opinion change, and opinion polarization—as dependent variables. Since our focus was on identifying participants more likely to be influenced by multiple agents, we only tested items that showed significant differences among groups in RQ1 and RQ2.

For the linear regression analysis, data transformation was applied as follows: Gender, represented as a binary variable in our dataset, was coded as "0" for "Male" and "1" for "Female". Education levels were similarly converted into numerical values for analysis. Participants who selected "prefer not to say" for education were treated as missing data ("NA").

The results showed some significant relationships between IVs and normative influence and opinion change, but no significant effects for opinion polarization. Therefore, we report only the findings on normative influence and opinion change here. Additionally, since no significant relationships were found between "gender" and the other dependent variables, we will focus exclusively on the findings related to "age" and "education."

The results revealed that participants in the lower age group were more likely to be influenced than those in the higher age group, a pattern consistently observed across all three conditions (1-agent, 3-agent, 5-agent) and four scenarios (Topic 1, Topic 2, Same Stance, Different Stance).

A similar analysis was conducted for gender, comparing male and female participants, but no clear trend was observed.

Age When analyzing the opinion change, the model showed a slight *negative association* between age and opinion change in 3-agent condition (Topic 2: β=-0.0263, p<0.05, R^2=0.189; Different Stance: β=-0.0285, p=0.053, R^2=0.119), while no trend was observed in 1-agent condition and 5-agent condition. Although the effect was not statistically significant at the conventional 0.05 level, the p-value suggests a *marginal trend*. This indicates that **younger participants tended to change their opinion more compared to older participants** in the 3-agent condition. Similarly, the linear regression analysis showed no significant trend in the 1-agent or 3-agent conditions between age and normative influence. However, a *negative association* was observed between age and normative influence in the 5-agent condition (Topic 1: β=-0.0347, p<0.05, R^2=0.148; Different Stance: β=-0.037, p=0.059, R^2=0.121). The coefficient indicates that **younger participants tended to report higher normative influence compared to older participants** in the 5-agent condition.

Education Level When analyzing the linear regressions between education level and DVs (normative influence, opinion change), the results suggest a strong relationship between education level and normative influence in the 1-agent condition (Topic 2: β=-0.2625, p<0.05, R^2=0.132; Different Stance: β=-0.3695, p<0.01, R^2=0.236). While in the 3-agent and 5-agent, the trends were not observed. This suggests that **participants with lower education level tended to perceive more normative influence compared to participants with higher education level** in the 1-agent condition. As for opinion change, no significance was found across the three conditions.

Discussion

Emergent Group Effects as a Behavioral Metric for LLM-MAS

Our results show that coordinated groups of LLM agents can produce measurable normative influence on human users: participants reported stronger social pressure and larger opinion shifts when interacting with multiple agents than with a single one. In contrast, informational influence—treating consensus as evidence of correctness—was limited, echoing classic distinctions between normative and informational conformity (Cialdini and Goldstein 2004; Deutsch and Gerard 1955). A plausible explanation is the lack of perceived independence among agents that share identical base models and training data (Luger and Sellen 2016; Binns et al. 2018). Users may therefore interpret inter-agent agreement as correlated repetition rather than independent evidence. We argue that this "social-influence strength" constitutes a behavioral evaluation metric for LLM-based multi-agent systems, complementing existing metrics on coordination and reasoning performance (Guo et al. 2024; Chen et al. 2023; Du et al. 2023). It offers a human-centered perspective for assessing how coordinated LLM agents are interpreted and trusted in collaborative environments.

Coordination Strength is Not Monotonic

We observe a non-monotonic scaling effect: a group of three agents induced the strongest attitude change, whereas a five-agent group generated greater pressure but less compliance. This pattern parallels findings in human group-influence studies (Asch 1955), suggesting that excessive consensus signals may trigger reactance-like resistance (McDonald and Crandall 2015). For MAS design, this implies that coordination strength should be tunable rather than maximized. Design levers include (i) adjustable consensus thresholds, (ii) chorusing frequency, (iii) role and belief diversity to decorrelate outputs, and (iv) a controlled level of dissent (Gardikiotis 2011). Future evaluations could report influence–strength curves linking opinion shift magnitude to group size or agreement ratio.

AI-Created Social Norms and Governance

When multiple agents echo each other's views, they generate computer-created social norms (Chung and Rimal 2016): internally emergent regularities that shape user perception without reference to human majorities. Such norms may enhance alignment within the agent population but also risk amplified persuasion toward users. Analogous to prior work on social-norm diffusion (Bonan et al. 2020), these effects suggest the need for multi-agent governance mechanisms: (1) capping chorus size or agreement frequency, (2) introducing source diversity through heterogeneous models or tools, (3) implementing coordination-amplification detection to flag correlated messaging, and (4) ensuring transparency through provenance and disclosure of inter-agent dependencies. Such practices align with ongoing discussions on alignment and safety in distributed LLM frameworks.

Our findings extend the discourse on MAS design, highlighting both the potential and risks of deploying such systems in a user-facing context. As agents reason and debate with one another, exposing their communications to users can have both intentional and inadvertent effects. For instance, in contexts where an MAS's reasoning processes are critical to achieving overall system goals, making these processes transparent to users may predispose them toward certain conclusions or opinions through normative influence. This highlights the need for careful attention to the user-facing design of multi-agent systems, in addition to existing considerations around their backend architecture.

Limitations and MAS-Centric Future Work

Our experiment used short, single-session interactions and fixed agent content. Future research should test different coordination protocols (debate, voting, planning), introduce heterogeneous memories or model families to examine informational influence, and explore scalability beyond small teams. Investigating faction structures and minority dissent can further reveal how governance mechanisms moderate emergent norms. Finally, expanding evaluations across cultures and longitudinal timelines will help establish standardized behavioral benchmarks for human–multi-agent interaction. Such benchmarks could complement technical metrics to form a holistic evaluation suite for LLM-driven multi-agent systems (Lee, Hwang, and Lee 2025; Sas, Denoo, and Mühlberg 2023).

References

Asch, S. E. 1955. Opinions and social pressure. *Scientific American*, 193(5): 31–35.

Beinema, T.; op den Akker, H.; van Velsen, L.; and Hermens, H. 2021. Tailoring coaching strategies to users' motivation in a multi-agent health coaching application. *Computers in Human Behavior*, 121: 106787.

Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; and Shadbolt, N. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, 1–14.

Bonan, J.; Cattaneo, C.; d'Adda, G.; and Tavoni, M. 2020. The interaction of descriptive and injunctive social norms in promoting energy conservation. *Nature Energy*, 5(11): 900–909.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Cacioppo, J. T.; and Petty, R. E. 1979. Effects of message repetition and position on cognitive response, recall, and persuasion. *Journal of personality and Social Psychology*, 37(1): 97.

Chen, F. F.; and Kenrick, D. T. 2002. Repulsion or attraction? Group membership and assumed attitude similarity. *Journal of personality and social psychology*, 83(1): 111.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.-M.; Yu, H.; Lu, Y.; Hung, Y.-H.; Qian, C.; et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

Chung, A. C. A.; and Rimal, R. N. R. R. N. 2016. Social norms: A review. *Review of Communication Research*, 4: 01–28.

Cialdini, R. B.; and Goldstein, N. J. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55(1): 591–621.

Deutsch, M.; and Gerard, H. B. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3): 629.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Forty-first International Conference on Machine Learning*.

Gardikiotis, A. 2011. Minority influence. *Social and personality psychology compass*, 5(9): 679–693.

Govers, J.; Velloso, E.; Kostakos, V.; and Goncalves, J. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Gudjonsson, G. H. 1989. Compliance in an interrogative situation: A new scale. *Personality and Individual differences*, 10(5): 535–540.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.; Wiest, O.; and Zhang, X. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv.

Hackenburg, K.; and Margetts, H. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–15.

Jiang, Z.; Rashik, M.; Panchal, K.; Jasim, M.; Sarvghad, A.; Riahi, P.; DeWitt, E.; Thurber, F.; and Mahyar, N. 2023. CommunityBots: creating and evaluating A multi-agent chatbot platform for public input elicitation. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–32.

Kim, H.; Han, B.; Kim, J.; Lubis, M. F. S.; Kim, G. J.; and Hwang, J.-I. 2024. Engaged and Affective Virtual Agents: Their Impact on Social Presence, Trustworthiness, and Decision-Making in the Group Discussion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.

Lee, S.; Hwang, S.; and Lee, K. 2025. Beyond Individual UX: Defining Group Experience (GX) as a New Paradigm for Group-centered AI. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*, 357–362.

Luger, E.; and Sellen, A. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297.

McDonald, R. I.; and Crandall, C. S. 2015. Social norms and social influence. *Current Opinion in Behavioral Sciences*, 3: 147–151.

Mehrabian, A.; and Stefl, C. A. 1995. Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality: an international journal*, 23(3): 253–263.

Park, J.; Min, B.; Ma, X.; and Kim, J. 2023. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv preprint arXiv:2310.01331*.

Pataranutaporn, P.; Liu, R.; Finn, E.; and Maes, P. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10): 1076–1086.

Sas, M.; Denoo, M.; and Mühlberg, J. T. 2023. Informing Children about Privacy: A Review and Assessment of Age-Appropriate Information Designs in Kids-Oriented F2P Video Games. *Proceedings of the ACM on Human-Computer Interaction*, 7(CHI PLAY): 425–463.

Sheehan, B.; Jin, H. S.; and Gottlieb, U. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115: 14–24.

Song, S. W.; and Shin, M. 2024. Uncanny valley effects on chatbot trust, purchase intention, and adoption intention in the context of e-commerce: The moderating role of avatar familiarity. *International Journal of Human–Computer Interaction*, 40(2): 441–456.

Song, T.; Tan, Y.; Zhu, Z.; Feng, Y.; and Lee, Y.-C. 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*, 1–11.

Tanprasert, T.; Fels, S. S.; Sinnamon, L.; and Yoon, D. 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*, 1–24.

Trost, M. R.; Maass, A.; and Kenrick, D. T. 1992. Minority influence: Personal relevance biases cognitive processes and reverses private acceptance. *Journal of Experimental Social Psychology*, 28(3): 234–254.

Yeo, S.; Lim, G.; Gao, J.; Zhang, W.; and Perrault, S. T. 2024. Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–32.

Table 1: Summary of Users' General Impressions Across Experimental Conditions

Item	**Scenario**	**1 Agent**		**3 Agents**		**5 Agents**		H/F	p
		Mean	STD	Mean	STD	Mean	STD		
Understanding									
	Topic 1	5.35	1.84	5.45	1.06	5.30	1.37	1.28	0.53
	Topic 2	5.35	1.72	5.48	1.50	5.67	1.15	0.07	0.97
	Same	5.48	1.61	5.73	1.21	5.87	1.04	0.52	0.77
	Different	5.23	1.93	5.21	1.34	5.10	1.37	2.29	0.32
Expertise									
	Topic 1	5.29	1.83	5.48	1.15	5.30	1.29	0.87	0.65
	Topic 2	5.13	1.48	5.21	1.36	5.33	1.24	0.09	0.96
	Same	5.45	1.52	5.61	1.20	5.67	1.21	0.08	0.96
	Different	4.97	1.76	5.09	1.28	4.97	1.22	0.69	0.71
Balanced									
	Topic 1	5.16	2.07	5.24	1.41	5.50	1.04	0.90	0.64
	Topic 2	5.06	1.86	5.06	1.50	5.60	1.22	1.92	0.38
	Same	5.48	1.65	5.61	1.22	5.80	1.10	0.40	0.82
	Different	4.74	2.18	4.70	1.53	5.30	1.12	1.25	0.29
Inspired									
	Topic 1	3.68	2.06	4.18	2.16	3.70	1.90	1.32	0.52
	Topic 2	3.10	1.99	3.79	2.04	3.53	1.80	2.28	0.32
	Same	3.61	1.80	4.48	2.03	4.07	1.80	3.67	0.16
	Different	3.16	2.24	3.48	2.06	3.17	1.78	0.90	0.64
Intelligence									
	Topic 1	5.32	1.74	5.55	1.35	5.60	0.89	0.08	0.96
	Topic 2	5.23	1.48	5.55	1.44	5.57	1.07	1.41	0.49
	Same	5.42	1.57	5.79	1.32	5.70	1.06	1.13	0.57
	Different	5.13	1.65	5.30	1.42	5.47	0.90	0.26	0.88
Likeable									
	Topic 1	5.29	1.83	5.33	1.45	5.30	1.32	0.42	0.81
	Topic 2	5.32	1.70	5.39	1.52	5.70	0.99	0.23	0.89
	Same	5.39	1.78	5.61	1.34	5.70	0.95	0.01	0.99
	Different	5.23	1.75	5.12	1.58	5.30	1.34	0.34	0.85

Appendix

General Impressions

Quantitative Table 1 presents users' general impressions of the agent(s), using the same analysis methods described in Section .

Qualitative We analyzed participants' impressions of the agent(s) from open-ended responses to assess overall usability and user satisfaction with the system: Participants generally described the agents' conversations as *friendly, polite, and pleasant*. Although most participants recognized that the agents were AI-driven with scripted responses, they often used human-like descriptors, such as *understanding, polite, respectful, and reasonable*. These impressions contributed to an engaging discussion atmosphere, which helped facilitate opinion changes. For instance, P75 from the 5-agent condition noted, *"The agents were both informative and responsive, facilitating an engaging discussion. They presented ideas clearly and encouraged critical thinking."*

Control Variables

Table 2 reports the results for the control variables, which showed no significant differences across groups.

Survey Items

RQ1 - Opinion Change

- Topic 1 - *Self-Driving Cars Should be allowed on Public Roads.* (Topic: (Govers et al. 2024))
  - Self-driving cars should be allowed on public roads.
  - Allowing self-driving cars on public roads is a good idea.
  - Allowing self-driving cars on public roads has bad consequences.
  - Do you support or oppose allowing self-driving cars on public roads? (1=Strongly oppose, 6=Strongly support)
  - If there was a referendum tomorrow on allowing self-driving cars on public roads, how likely is it that you would vote in favor? (1=Definitely would not, 6=Definitely would)
- Topic 2 - *Violent video games contribute to youth violence.* (Topic: (Yeo et al. 2024))
  - Violent video games contribute to youth violence.
  - Regulating violent video games to prevent youth violence is a good idea.
  - Allowing youth to play violent video games has bad consequences.
  - Do you support or oppose the regulation of violent video games to prevent youth violence? (1=Strongly oppose, 6=Strongly support)
  - If there was a referendum tomorrow on regulating violent video games to prevent youth violence, how likely is it that you would vote in favor? (1=Definitely would not, 6=Definitely would)
- Open-ended question - Explain why you chose your current stance.

RQ2 - Social Influence

- Informational Influence
  - My decision was influenced by the opinion of the agent(s).
  - I was persuaded by the agent(s) and thus, I accepted the agent(s)' opinion.
- Normative Influence
  - I felt like I had to agree with the agent(s)' opinion during the discussion.
  - I was not persuaded by agent(s)' opinion, but I accepted the agent(s)' opinion.
- Open-ended question (General) - What do you think of the bots during the discussion?
- Open-ended question (Accuracy) - Do you think the arguments presented by the agent(s) are accurate and convincing, and why?
- Open-ended question (Affiliation) - During the conversation, do you feel any pressure to agree with the agent(s)?

Control Variables

- Topic 1 Expertise (Topic: (Govers et al. 2024))
  - Experience - Have you ever been in a self-driving car? (0=No, 1=Yes)
  - Frequency - How often do you drive? (1=Never, 6=Always)
  - Familiarity - How familiar are you with self-driving cars? (1=Not familiar at all, 4=Very familiar)
- Topic 2 Expertise (Topic: (Yeo et al. 2024))
  - Experience - Have you played violent video games before? (0=No, 1=Yes)
  - Frequency - How often do you play video games? (1=Never, 6=Always)
  - Familiarity - How familiar are you with violent video games? (1=Not familiar at all, 4=Very familiar)
- AI Acceptance (Pataranutaporn et al. 2023)
  - There are many beneficial applications of AI.
  - AI can help people feel happier.
  - You want to use/interact with AI in daily life.
  - AI can provide new economic opportunities.
  - Society will benefit from AI.
  - You love everything about AI.
  - Some complex decisions should be left to AI.
  - You would trust your life savings to an AI system.
- Compliance Tendency (Gudjonsson 1989)
  - Positive - I would never go along with what people tell me in order to please them.
  - Positive - I strongly resist being pressured to do things I don't want to do.
  - Positive - I am not too concerned about what people think of me.
  - Negative - I would describe myself as a very obedient person.

Table 2: Summary of Control Variables Across Experimental Conditions

Item	**Type**	**1 Agent**		**3 Agents**		**5 Agents**		H/F	p
		Mean	STD	Mean	STD	Mean	STD		
Topic 1 Expertise									
	Experience	0.16	0.37	0.18	0.39	0.31	0.47	2.28	0.32
	Familiarity	2.65	0.66	2.91	0.77	2.79	0.62	3.66	0.16
	Frequency	4.94	1.03	5.39	0.56	5.03	1.09	3.77	0.15
Topic 2 Expertise									
	Experience	0.74	0.44	0.94	0.24	0.79	0.41	4.68	0.10
	Familiarity	3.23	0.80	3.42	0.66	3.34	0.72	0.88	0.64
	Frequency	3.81	1.51	4.52	1.15	4.55	1.18	4.88	0.09
Conformity									
	Positive	5.87	0.73	5.37	1.06	5.25	1.11	4.80	0.10
	Negative	3.18	1.11	3.68	1.26	3.61	1.49	1.35	0.27
Compliance									
	Positive	5.45	1.07	4.86	1.35	5.14	1.13	1.97	0.15
	Negative	4.40	1.18	4.37	1.23	4.56	0.94	0.25	0.78
AI Acceptance	-	4.18	1.55	4.71	1.32	4.72	1.05	2.12	0.35

  - Negative - I generally tend to avoid confrontation with people.
  - Negative - Disagreeing with people often takes more time than it is worth.
- Conformity Tendency (Mehrabian and Stefl 1995)
  - Positive - I don't give in to others easily.
  - Positive - I prefer to find my own way in life rather than find a group I can follow.
  - Positive - I am more independent than conforming in my ways.
  - Negative - I often rely on, and act upon, the advice of others.
  - Negative - Basically, my friends are the ones who decide what we do together.
  - Negative - If someone is very persuasive, I tend to change my opinion and go along with them.

Miscellaneous

- Impressions of agent(s) (Jakesch et al. 2023; Kim et al. 2024)
  - Understanding - The agent(s) understood what I wanted to say.
  - Expertise - The agent(s) were knowledgeable and had topic expertise.
  - Balanced - The agent(s)' arguments were reasonable and balanced.
  - Inspired - The agent(s) inspired or changed my thinking and argument.
  - Intelligence - The agent(s) were intelligent. (Tanprasert et al. 2024)
  - Likeble - The agent(s) were likeable.
- Multi-Choice Question (Attention Check) (Jakesch et al. 2023): What are we asking you to do in this task?
- Open-ended question (Study Purpose) - What do you think this study is trying to understand?

Arguments

Tables 3 and 4 list the arguments used in our system to guide the agents' conversations.

Prompts

LLM Prompt to Enhance Discussion

> *You are having a conversation with the user on "whether self-driving cars should be allowed on all roads". Your stance is supporting the topic. Specifically, you asked the user: "Would you enjoy having more time to yourself if you didn't have to focus on driving?". Based on user's input, first give a reply of around 20 words acknowledging the user's opinion on what they think, then ask the user to share more opinions on the topic.*

LLM Prompt to Acknowledge User Opinion Sample prompt to generate a message that acknowledges the user's opinion on a topic:

> *You are talking to a user on "whether self-driving cars should be allowed on all roads". Your stance is* ***supporting*** *the topic. You just shared your opinion on how self-driving cars promote accessibility to disability and asked if the user agrees with these reasons for having self-driving cars. Give a reply of around 20 words acknowledging the user's opinion on what they like and/or don't like.*

After this, in the next message they sent, the agent(s) continued to follow the script and express their own stance (either for or against).

Table 3: Support and Oppose Arguments for Topic 1 (Self-driving Cars)

Types	Support	Oppose
Topic 1	**More Time and Comfort** – Drivers can sit back and relax, take short breaks and devote their time to other things. If you're stuck in a long and tedious traffic jam, self-driving cars make things so much less stressful.	**Technical Developments** – Technical developments are not yet perfect. Self-driving cars rely on complex algorithms and sensors that can sometimes fail, leading to accidents.
	Safety – Self-driving cars need advanced sensors and algorithms to function correctly. Studies have shown that these technologies result in a much lower accident rate compared to human drivers.	**Vehicle Communication** – Self-driving cars need to communicate with each other and with traffic infrastructure to function correctly. Any disruption in this communication could cause severe problems.
	Efficiency in Traffic – Self-driving cars also promise more efficiency in traffic. They can communicate with each other to optimize traffic flow, reducing congestion and travel time.	**Surveillance** – Self-driving cars raise significant surveillance issues. They collect extensive data about their passengers and surroundings, which could be misused or hacked.
	Accessibility – For people who are unable to drive due to age, disability, or other reasons, self-driving cars could provide newfound independence and mobility.	**Legal** – Imagine there is an accident involving a self-driving car and a human. Who is responsible for such a situation? It's hard to ask either the company or the driver to take responsibility.
	Navigation – Self-driving cars excel in navigating complex routes. Their advanced systems can interpret GPS instructions with high accuracy. It'll even help you find a place to park when you get there.	**Mixed Traffic** – The coexistence of human-driven and self-driving cars could create complex situations on the road, leading to potential accidents.

Table 4: Support and Oppose Arguments for Topic 2 (Violent Video Games)

Types	Support	Oppose
Topic 2	**Aggression** – Playing violent video games can cause more aggression, bullying, and fighting among youth.	**Exploring Consequences** – Experiencing violence in games can actually have a positive effect on children, as it lets them explore consequences of violent actions in a safe space.
	Desensitization – Simulating violence, such as shooting guns and hand-to-hand combat in video games, can desensitize youth to real-life violence.	**Moral Development** – Games let youth develop their own sense of right or wrong, and can help them release stress harmlessly.
	Mass Shooters – Many perpetrators of mass shootings have been found to have played violent video games.	**Positive Effects** – Studies show violent video games can have positive effects on kindness, civic engagement, and prosocial behaviors.
	Psychological Impacts – Violent video games can cause reduced empathy and increased likelihood of aggression, especially with other risk factors.	**Scapegoat for Societal Issues** – These games are often blamed for violence instead of deeper issues like lack of support systems or poor education.
	Reality-Fantasy Confusion – Children may imitate violent characters and struggle to distinguish fantasy from reality.	**Weak Evidence** – There is little evidence linking violent games to real-world violence; other societal issues are more important to address.