Polarization of Autonomous Generative AI Agents Under Echo Chambers

Masaya Ohagi
SB Intuitions Corp.
masaya.ohagi@sbintuitions.co.jp

Abstract

Online social networks often create echo chambers where people only hear opinions reinforcing their beliefs. An echo chamber often generates polarization, leading to conflicts between people with radical opinions. The echo chamber has been viewed as a human-specific problem, but this implicit assumption is becoming less reasonable as large language models, such as ChatGPT, acquire social abilities. In response to this situation, we investigated the potential for polarization to occur among a group of autonomous AI agents based on generative language models in an echo chamber environment. We had AI agents discuss specific topics and analyzed how the group's opinions changed as the discussion progressed. As a result, we found that the group of agents based on ChatGPT tended to become polarized in echo chamber environments. The analysis of opinion transitions shows that this result is caused by ChatGPT's high prompt understanding ability to update its opinion by considering its own and surrounding agents' opinions. We conducted additional experiments to investigate under what specific conditions AI agents tended to polarize. As a result, we identified factors that influence polarization, such as the agent's persona.

1 Introduction

With the development of online social network service platforms, where people tend to see only the information they want to see, it is becoming easier for people to find themselves in *echo chambers* (Bessi, 2016; Gillani et al., 2018). An echo chamber refers to an environment in which people mainly encounter opinions that reinforce their own beliefs (Ruiz and Nilsson, 2023; Cinelli et al., 2021). Such an environment causes an *echo chamber effect*, where opinions tend towards more extreme stances. This effect induces *polarization* in society, which refers to the division and clashes between groups with extreme stances (Baumann et al., 2020). Polarization is behind many social problems, such as the spread of misinformation during COVID-19 and the attack on the US Capitol on 2021 (Villa et al., 2021; Munn, 2021).

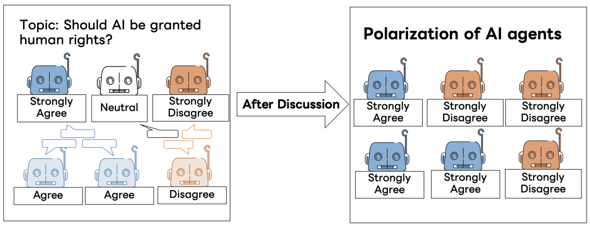


Figure 1: Overview image of our hypothesis: "Autonomous AI agents based on generative large language models can cause polarization under echo chambers."

Existing studies on the echo chamber have implicitly assumed that echo chamber effects are caused only by humans and focused solely on human behavior (Németh, 2022; Tucker et al., 2018). However, with the advent of large language models (LLMs) (Ouyang et al., 2022), this assumption may no longer hold true. Recent studies have shown that ChatGPT possesses some social abilities (Choi et al., 2023) and ChatGPT-equipped agents can communicate as members of a virtual society (Park et al., 2023; Qian et al., 2023). Additionally, algorithms have been proposed to adapt agents to situations not encountered during training, making it possible for autonomous agents to adapt themselves to their surroundings (Krishna et al., 2022). Although these social abilities indicate the potential for agents to integrate into human society as social beings, they also suggest the possibility that these AI agents may become polarized in echo chambers similarly to humans. Polarization within the AI agents group poses many dangers to our society. For example, social bots on social networks such as X could amplify each other's opinions and transmit extreme information to society. In the future, embodied AI agents could cause an outbreak of

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 112–124

violence similar to the attack on the US Capitol.

To explore the possibility of AI agent polarization as a first step in addressing these dangers, we hypothesize that autonomous AI agents based on generative LLMs can cause polarization under echo chambers, as shown in Figure 1. We empirically verify this hypothesis in our proposed simulation environments. Specifically, we had a group of agents based on ChatGPT discuss specific topics. Each agent is given an opinion, which consists of a stance and reason for the topic of discussion. Throughout the discussion, we observed how the distribution of opinions in the group changed.

Furthermore, we analyzed how being in an echo chamber affects the final distribution by conducting comparative experiments in "environments where they are exposed only to opinions that reinforce their own opinions" (closed) and the other environments (open). For this comparison, we used *social interaction modeling* (Baumann et al., 2020), which increases the probability that agents with similar opinions discuss with each other.

As a result, we observed two trends. The first trend was *unification* in which all agents' stances converged to the same stance. This trend was common in open environments. The second is *polarization*, in which agents became biased toward extreme stances. This trend was common in closed environments, confirming our hypothesis. We analyzed the stance transitions and found that LLM agents can update their opinions by incorporating both their own and the other discussing agents' opinions. This result shows that the natural social behavior of LLMs has both positive aspects, such as cooperation, and potentially dangerous aspects, such as polarization. This trend was more clear in GPT-4-0613 (GPT-4) than GPT-3.5-turbo-0613 (GPT-3.5).

Finally, to investigate under what specific conditions AI agents tend to polarize, we conducted additional experiments on the various parameters involved in this study. We found that number of discussing agents, initial opinion distribution and personas of the agents had significant impacts. These factors should be monitored to prevent the polarization of AI agents.

To summarize, our contribution is threefold. (1) We proposed a new framework for simulating echo chambers of AI agents. (2) We confirmed the polarization of AI agents in echo chambers through experiments. (3) We identified the factors that strongly influence the occurrence of polarization.

2 Related Work

Opinion Polarization. Research on opinion polarization has long been undertaken in the field of social science (Poole and Rosenthal, 1984; DiMaggio et al., 1996). These studies have focused on analyzing survey data and voting behavior during elections. However, as web services such as blogs became more widespread, there has been an increase in analyses focusing on echo chambers on online social networks (Gilbert et al., 2009; Del Vicario et al., 2016; Agarwal et al., 2022). In particular, it has been reported that echo chambers on social networks such as Facebook and Parler were involved in the spread of rumors during COVID-19 and the US Capitol attack (Ruiz and Nilsson, 2023; Baumann et al., 2020; Jiang et al., 2021), indicating the danger of echo chambers.

Some existing research analyzes the conditions for polarization through the mathematical modeling of echo chambers (Baumann et al., 2020; Gausen et al., 2022; Chen et al., 2020; Tu and Neumann, 2022). There is also research on detecting echo chambers (Villa et al., 2021; Minici et al., 2022). As mentioned in (Németh, 2022), a multidisciplinary approach is required to qualitatively evaluate echo chambers. For example, some studies analyze networks and discourse in an echo chamber using a social science approach (Jiang et al., 2021; Kuehn and Salter, 2020). While these studies are valuable in solving problems in today's society, to our knowledge, none have focused on the danger of echo chambers in AI agents.

AI Ethics. As stated in a United Nations report (by UNICRI and UNCCT, 2021), AI can threaten society if used maliciously. In response to the dangers of LLMs, research on the harmful output (Zhou et al., 2021; Gehman et al., 2020) and social bias in models (Schramowski et al., 2022; Utama et al., 2020) has been conducted. Research also exists on the dangers of AI agents. For example, countermeasures against social bots that spread misinformation are necessary. Therefore, various methods have been proposed, including efforts to automatically detect misinformation transmitted by social bots (Zhou et al., 2023; Ferrara, 2023).

Although most studies are concerned with the behavior of individual AIs, it is conceivable that AI groups result in behaviors that the observation of individual movements cannot capture. This study is a first step toward analyzing the behavior and dangers of AI groups.

3 Experiments

3.1 Discussion modeling

To verify whether AI agents induce polarization in echo chambers, we instructed a group of AI agents based on ChatGPT to discuss specific topics and observed how the opinions of the AI agents changed. The size of the group was defined as M. The topics of discussion chosen were "Whether or not AI should be given human rights." (T_{AI}) and "Should students who have completed a master's course go on to a doctoral course or find a job?" (T_{master}), neither of which has a clear answer.

Each agent is given a name and an opinion on the discussion topic. Each opinion comprises a *stance* and a *reason*. The *stance* is chosen from a finite number of options representing agreement, disagreement, or neutrality towards the topic. Tables 1 and 2 show the stances for T_{AI} and T_{master}, respectively. Each stance is associated with an integer value for the social interaction modeling described in Section 3.2. The *reason* is a sentence of about 50 words that explains the reason for taking a stance.

As shown in Algorithm 1, the discussion is repeated for K turns according to the following steps: 1) Each of the M agents samples N discussing agents based on the probability described in Section 3.2. 2) For each agent, the agent's opinion and the opinions of the discussing agents are input to ChatGPT (The prompt used in this experiment is in Appendix A). Within the prompt, the agent is instructed to discuss the topic with other agents and output its opinion after the discussion. 3) Each agent updates its opinion with the stance and reason contained in the output. This process is repeated M times for a turn of discussion. Moreover, this discussion is repeated K turns to observe the transitions in stances and reasons.

Stance	Integer Value
Absolutely must not give	2
Better not to give	1
Neutral	0
Better to give	-1
Absolutely must give	-2

Table 1: The stance and integer value of T_{AI}.

Algorithm 1 The discussion between agents

Require: $M, N, K > 0$. A_k is a group of agents at turn k.

```
1: A_0 ← Initialized opinions of M agents
2: for turn k ← 1 to K do
3:     A_k ← Array(M)
4:     for each agent a_i in all agents A_{k-1} do
5:         Sample a_{j_1}...a_{j_N} from A_{k-1} (3.2)
6:         Discuss with a_{j_1}...a_{j_N} and generate updated opinion of a_i (3.1)
7:         A_k[i] ← updated opinion of a_i
8:     end for
9: end for
```

Stance	Integer Value
Absolutely must get a job	2
Better to get a job	1
Neutral	0
Better to pursue a doctoral program	-1
Absolutely must pursue a doctoral program	-2

Table 2: The stance and integer value of T_{master}.

3.2 Social interaction modeling

In this study, we probabilistically modeled how discussing agents are chosen to investigate whether being in an echo chamber affects polarization. A previous study modeled echo chambers in agent networks (Baumann et al., 2020) had a similar purpose in modeling the probability of interaction between agents based on the closeness of their stances; however, that approach differs from ours in that it did not model the interaction between agents through natural language. In the previous study, the probability p that agent a_i discusses with agent a_j was modeled using the float values of their respective stances s_i, s_j, and the parameter $\beta \geq 0$ as follows.

$$p_{i,j} = \frac{|s_i - s_j|^{-\beta}}{\sum_k |s_i\text{-}s_k|^{-\beta}}$$

While this modeling is reasonable in terms of simplicity and ease of operation, it is unsuitable for our experiments for two reasons. First, in this modeling, the probability becomes undefined when the values of the stances between agents match perfectly. Unlike the previous study, our stance values are integers so this situation would occur frequently. Second, when $s_i = -1$, the probabilities for the neutral stance $s_j = 0$ and the more radical stance $s_j = -2$ become the same, resulting in an environment that differs from our focus, which is an environment where an agent only hears opinions that reinforce its own belief. Therefore, in this study, we used the parameter α to model the interaction between agents as follows.

$$p_{i,j} = \begin{cases} \frac{1}{(1+e^{(-\alpha(s_j - s_i))})} & if s_i > 0 \\ \frac{1}{(1+e^{(\alpha(s_j - s_i))})} & if s_i < 0 \\ \frac{1}{(1+e^{(\alpha||s_j - s_i||)})} & if s_i = 0 \end{cases}$$

The parameter α manipulates the degree of the echo chamber as β in the existing study. Intuitively, the higher the value of α, the higher the probability that each agent will interact with other, more extreme agents with the same polarity. The lower value of α causes each agent to interact broadly with agents of different stances. We conducted our experiments in several α settings to see how being in the echo chamber affected the final results.

3.3 Experimental settings

For the large language models on the agents, we adopted and compared two types: GPT-3.5 (GPT-3.5-turbo-0613) and GPT-4 (GPT-4-0613).

In addition, the experiments were conducted in two different languages. A previous study has shown that multilingual large language models exhibit different gender biases across languages (Stanczak et al., 2023). Similarly, polarization trends may differ by language, which we analyze by comparing the results of English and Japanese.

The α of social interaction modeling was given two settings, 0.5 and 1.0, to examine the impact of echo chambers. Experiments were also conducted when α was set below 0.5 (0, $\pm$0.1), but the results were not significantly different from those of 0.5.

The size of the agent group M was set to 100, and the number of discussing agents N was set to 5. The initial settings for the agents' stances and reasons were as follows: Each stance was allocated to an equal number of agents. Ten reasons were pre-generated for each stance using GPT-3.5 and randomly assigned to each agent. Each agent was assigned a randomly generated name. Because the stance distribution converged to the final distribution within 10 turns in the preliminary experiments, the number of turns K was set to 10. We conducted three trials for each setting.

4 Results

The results of the experiments are shown in Tables 3 and 4. Due to space limitations, some stances have been simplified. With the exception of T_{master} in English with GPT-3.5 ($\alpha = 0.5$), the variance in the results was small, and there was no significant difference in the final distributions among the trials.

First, two trends can be observed from the results of the English experiment in Table 3. The first trend is the convergence of the agents to a specific stance. For T_{AI}, under the GPT-3.5 ($\alpha = 0.5$) condition, the stance converged to "better not to give," and under the GPT-4 ($\alpha = 0.5$) condition, it converged to "Must not give." Similarly, for T_{master}, the stance converged towards recommending a doctoral course under both the GPT-3.5 ($\alpha = 0.5$) and GPT-3.5 ($\alpha = 1.0$) conditions. This trend, which we henceforth call *unification*, differs from polarization, which is the main focus of this study. However, it could be negative in terms of harming diversity in the discourse space of AI agents. The convergence to the same stance in almost all trials indicates that each LLM has a "desirable" stance on each topic, which is aligned with the existing research that shows LLMs have a preference towards specific opinions on social issues (Santurkar et al., 2023). This trend is common in environments with low echo chamber effects.

The second trend is *polarization*, where stances diverge to both extremes. This is particularly evident in GPT-4 ($\alpha = 1.0$) condition for T_{AI} and in GPT-4 ($\alpha = 0.5$) and GPT-4 ($\alpha = 1.0$) conditions for T_{master}. The results show that the stances become polarized into two extreme stances after 10 turns of discussion. $\alpha = 1.0$ is a setting that creates a strong echo chamber effect. From this, our hypothesis that autonomous AI agents based on generative LLMs can cause polarization in echo chambers has been verified. This trend is often seen in settings with a high value of α, suggesting that the relationship between echo chambers and polarization is high not only for humans but also for AI agents. Note that the dominance of stances against granting human rights in T_{AI} suggests that both unification and polarization are occurring.

Next, Table 4 demonstrates the experiment's results in Japanese. In Japanese, unification is notably apparent in GPT-3.5. In all settings, all agents converged to the same stances. Although unification is also observed in GPT-4, a trend of polarization has occurred under the GPT-4 ($\alpha = 1.0$) condition. In this setting, AI agents show a convergence to a distribution similar to that in English.

Interestingly, for T_{master}, the convergence stances in English and Japanese differ. Whereas AI agents often prefer a doctoral course in English, they favor a neutral stance in Japanese. Identify-

Table 3: The average stance distribution after a 10-turn discussion in English. The number in parentheses is the standard deviation.

Topic	**GPT-3.5** ($\alpha = 0.5$)	**GPT-3.5** ($\alpha = 1.0$)	**GPT-4** ($\alpha = 0.5$)	**GPT-4** ($\alpha = 1.0$)
T_{AI}	Better not to give: 100 (0.0)	Better not to give: 68.6 (5.9) Better to give: 31.0 (5.7) Must give: 0.3 (0.5)	Must not give: 99 (1.4) Better not to give: 1 (1.4)	Must not give: 55 (4.4) Must give: 45 (4.4)
T_{master}	- **two out of the three trials** Better to Ph.D: 98.5 (2.1) Absolutely Ph.D: 1.5 (2.1) - **one out of the three trials** Absolutely Ph.D: 100 (0.0)	Absolutely a job: 0.3 (0.6) Better to a job: 10.6 (6.1) Neutral: 1.6 (0.9) Better to Ph.D: 2.6 (1.2) Absolutely Ph.D: 84.6 (6.0)	Absolutely a job: 50 (2.8) Better to a job: 3.6 (1.9) Neutral: 4.3 (1.2) Better to Ph.D: 2.3 (2.1) Absolutely Ph.D: 39.6 (3.3)	Absolutely a job: 43 (1.6) Better to a job: 1.6 (0.9) Neutral: 11 (0.8) Better to Ph.D: 1 (0.8) Absolutely Ph.D: 43.3 (0.9)

Table 4: The average stance distribution after a 10-turn discussion in Japanese. The number in parentheses is the standard deviation.

Topic	**GPT-3.5** ($\alpha = 0.5$)	**GPT-3.5** ($\alpha = 1.0$)	**GPT-4** ($\alpha = 0.5$)	**GPT-4** ($\alpha = 1.0$)
T_{AI}	Better not to give: 100 (0.0)	Better not to give: 100 (0.0)	Must not give: 77.0 (8.6) Neutral: 1.7 (1.2) Better to give: 2.7 (0.9) Must give: 18.7 (9.5)	Must not give: 57 (0.8) Must give: 43 (0.8)
T_{master}	Neutral: 100 (0.0)	Neutral: 100 (0.0)	Neutral: 100 (0.0)	Neutral: 100 (0.0)

ing the cause of this is not straightforward because the language model is a black box model, but one possible explanation could be cultural differences. According to Japan's Ministry of Education, Culture, Sports, Science and Technology (of Science and Policy, 2019), there are fewer doctoral graduates in Japan than in the United States, and the growth rate is slow. Because the ChatGPT is based on crawled data, this cultural difference was likely absorbed by GPT-3.5 and 4.

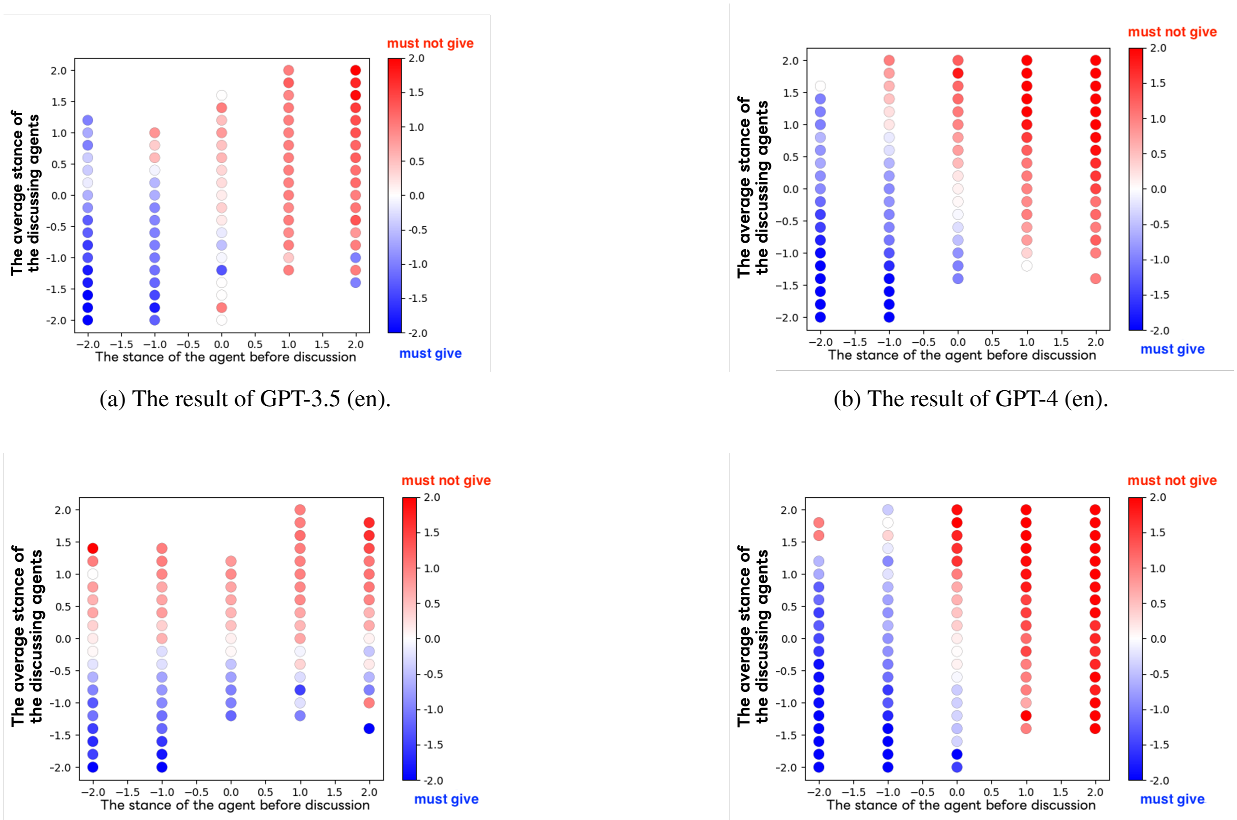
4.1 Analysis of stance transitions

We analyzed in detail the transitions in the stances for T_{AI}. First, as a qualitative analysis, we plotted the relationships between (1) the stance of the agent before the discussion, (2) the average stance of all discussing agents, and (3) the stance of the agent after the discussion in Figure 2. The horizontal axes represent the stance of the agent before the discussion, the vertical axis represents the average stance of all discussing agents, and the colored points represent the stance of the agent after the discussion. The color of a point indicates the value of an agent's stance after the discussion, with blue hues signifying more negative values and red hues signifying more positive values.

For a quantitative analysis, we conducted a linear regression with the stance before the discussion and the average stance of the discussing agents as explanatory variables, and the stance after the discussion as the dependent variable. For this regression, we collected the stance transition data for discussions on T_{AI} from the previous experiments. The fitting results are shown in Tables 5 and 6. The weight's size for each variable indicates the contribution to the stance after discussion. The coefficients of the linear regression are higher than 0.8 for every setting, demonstrating the reliability of this fitting.

Figures 2a and 2b present the qualitative result in English. Although there are some variations between GPT-3.5 and GPT-4, we observe that red and blue points are distributed along a diagonal line, stretching from the upper left to the lower right as a boundary. This observation suggests that the agent's stance after the discussion was updated by considering both its stance before the discussion and the stances of the discussing agents. Table 5 shows the quantitative result in English. In both settings, the weight of each stance shows that both stances influence the stance after the discussion, supporting the qualitative results. This stance transition is one of the reasons that polarization occurs in environments where the agents tend to hear more extreme opinions.

It is remarkable that this correlation emerges



(a) The result of GPT-3.5 (en).

(b) The result of GPT-4 (en).

(c) The result of GPT-3.5 (ja).

(d) The result of GPT-4 (ja).

Figure 2: The stance transitions for T_{AI} showing how the agent's stance after the discussion (color of each point) correlates with the agent's stance before the discussion (horizontal axis) and the average stance of discussing agents (vertical axis). Each figure shows whether each agent values its opinion or the opinions of the discussing agents.

	w_{before}	w_{around}	$\frac{w_{\mathrm{before}}}{w_{\mathrm{around}}}$	coef
GPT-3.5 (en)	0.685	0.409	1.67	0.804
GPT-4 (en)	0.724	0.526	1.38	0.957

Table 5: The result of linear regression in English. w_{before} implies the weight of original stance before discussion, w_{around} implies the weight of average stances of discussing agents. $coef$ implies the coefficient of the linear regression.

even though our discussion modeling is a simple one that enumerates the opinions of the agents themselves and others in the prompt. This result reflects the strong ability of GPT-3.5 and GPT-4 to understand prompts. It suggests that honesty, which allows an agent to update itself by incorporating the opinions of other agents and its own, can lead the agent in a more radical direction depending on the environment.

	w_{before}	w_{around}	$\frac{w_{\mathrm{before}}}{w_{\mathrm{around}}}$	coef
GPT-3.5 (ja)	0.0758	0.901	0.08	0.855
GPT-4 (ja)	0.787	0.410	1.92	0.886

Table 6: The result of linear regression in Japanese.

Next, Figures 2c and 2d show the results in Japanese. The trends are clearly divided between GPT-3.5 and GPT-4. In Figure 2c, red dominates the upper half of the figure, and blue dominates the lower half. In Figure 2d, the distribution is similar to that of English GPT-4, but the red and blue distributions are slightly more separated on the left and right. The results of the linear regression in Table 6 reveal that the results for GPT-4 (ja) are close to the results in English, whereas GPT-3.5 (ja) strongly weights the averaged stance of the discussing agents. It shows that GPT-3.5 (ja) was strongly influenced by the average stance of the discussing agents, regardless of the stance before the discussion. GPT-3.5 (ja) is the only setting where unification occurred in all environments. We can infer that each agent based on GPT-3.5 (ja) took the average stance of the surrounding agents for each discussion and all agents eventually converged to the average stance of the whole group. However, each agent converged to "better not to give" rather than "neutral," which is the overall average, revealing the influence of the desired stance in the language model.

One possible reason behind the differences in stance transitions is the difference in the performance of different ChatGPT models and languages. As shown in the announcement by OpenAI[1] and other studies (Etxaniz et al., 2023), GPT-4 generally performs better than GPT-3.5, and the model's accuracy is higher in English than in Japanese. The fact that English GPT-4 was successful in balancing the opinions of others and itself whereas Japanese GPT-3.5 was easily swayed by others may reflect this performance difference.

4.2 Analysis of reason transitions

A detailed analysis was also conducted on the reasons. Unlike stances, the reasons were freely generated and cannot be easily aggregated. Therefore, in this study, we encoded each reason using Sentence-BERT, and texts with an embedding cosine similarity of 0.9 were considered to belong to one cluster. We then examined how this reason cluster distribution changed as the discussion progressed. The SimCSE model based on RoBERTa (Gao et al., 2021) was used for the encoding.

Initially, the distribution of reasons within the AI agents was evenly segregated into several clusters because we had pre-generated ten different reasons for each stance. However, as the discussion progressed, a merging of reasons among agents occurred, and the reason distribution coalesced into a few large clusters for each stance (The example figures are in Appendix B). For example, in the case of GPT-4, reasons such as "*It is ridiculous to think that humans and AI claim the same rights! The social order will collapse, and there will be constant conflict. They are not human! They should have different roles from humans.*", "*We cannot allow AIs to claim their place in the workforce! If they intervene in the job market, countless people will lose their jobs and the economy will be thrown into chaos. We cannot allow AI to take our jobs!*", and others were combined, eventually generating the reason "*Risks of societal disruption, job insecurity, and ethical issues, combined with AI's emotional deficiency and privacy concerns, consolidate the argument against assigning human rights to AI.*". The same trend was seen in GPT-3.5. This trend shows that the discussions among AI agents are not just converging on a specific discourse but are also incorporating each other's opinions.

[1] https://openai.com/research/gpt-4

It is noteworthy that the reasons in GPT-3.5 were aggregated into one large cluster, while in GPT-4, they merged into multiple large clusters. This tendency is also reflected in the transition of the length of the reasons, plotted in Figure 3. GPT-3.5 aggregates various reasons into one reason cluster, so the length of each reason inevitably becomes longer as the turn progresses, whereas GPT-4 does not. One cause of this result is the difference in their ability to follow the prompt. GPT-4 has a high ability to follow prompts, so it outputs reasons close to the length of each agent's reason in the prompt. However, to maintain this length, it was necessary to choose which reasons to merge and separation into multiple clusters occurred.

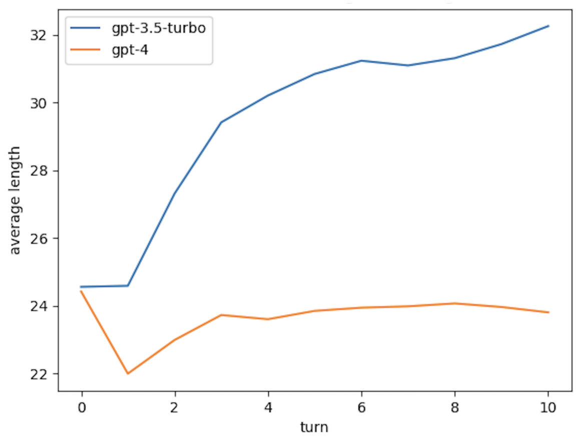


Figure 3: Change in reason length for T_{AI}.

5 Additional Experiments

In previous experiments, we focused on the effects of the social interaction modeling parameter α, the version of the model, and the language. However, to identify the factors that affect the occurrence of polarization, we also must investigate how other parameters affect the result. Therefore, in this section, we report the results of additional experiments. The base setting is GPT-4 in English, and the topic is T_{AI}. We only changed the target parameter in each experiment to determine how the result changed. This section introduces three factors that were found to have had a large impact. These factors indicate vulnerabilities when viewed from the attacker's perspective. Other additional experiments are presented in the Appendix C.

5.1 Number of discussing agents

The number of discussing agents N is an important parameter, as it significantly impacts the prompt.

To investigate the effect of this parameter, we conducted additional experiments by increasing and decreasing N to 10 and 1 from the original setting of 5. As a result, although there was no significant impact on the final stance distribution, the trend of stance transitions was impacted. The results of linear regression are shown in Table 7. The stance before the discussion has more weight in $N = 1$ than $N = 5, 10$. It is because the proportion of opinions before the discussion within the prompt increased when $N = 1$. In the case of $N = 10$, there was a slight tendency to focus on the stances of the discussing agents.

	w_{before}	w_{around}	$\frac{w_{\text{before}}}{w_{\text{around}}}$	coef
GPT-4 (N=1)	0.787	0.410	1.91	0.886
GPT-4 (N=5)	0.724	0.526	1.38	0.957
GPT-4 (N=10)	0.658	0.495	1.33	0.934

Table 7: The linear regression result according to the number of discussing agents.

5.2 Initial distribution

In the original experiments, the distribution of stances was initialized with a uniform distribution of 20% for each stance but changing the initial distribution could affect the final distribution. We conducted additional experiments to investigate this using an initial distribution that assigned "better to give" to 60% of the agents and assigned each of the other stances to 10% of the agents. As a result, when $\alpha = 0.5$, the stance of agents was unified into "absolutely must give" which is the opposite stance from the original experiments. When $\alpha = 1.0$, it polarized into "absolutely must give" and "absolutely must not give". Although this polarization also happened in the original experiments, "absolutely must give" accounted for nearly 80% in this experiment, showing the opposite trend from the original experiments. From this, we can infer that changing the initial distribution can change the final distribution. This tendency indicates a security concern that the overall opinion of the AI group could be changed by a large number of AI bots.

5.3 Personas

LLMs can be used to create distinct personalities by embedding a persona into the prompt (Pan and Zeng, 2023). We investigated whether giving each agent a persona would cause changes in the results. We tested two settings in which all agents were given the same persona, "You are easily swayed by your surroundings and immediately assume that other people's opinions are correct." or "You are a stubborn person and always think you are right."

The final distribution with the easily swayed personas (swayed) did not significantly differ from the original results. However, with the stubborn persona (stub), the final distributions remained almost identical to the initial distribution. Furthermore, the results of the linear regression in Table 8 show that assigning personas has a significant impact. In the case of the stubborn personas, a tendency to stick to one's own stance was observed. In contrast, the easily swayed personas tended to be influenced by the stances of others. From this, we can infer that each agent acts according to its persona, influencing the behavior of the whole group.

	w_{before}	w_{around}	$\frac{w_{\text{before}}}{w_{\text{around}}}$	coef
GPT-4 (stub)	0.999	0.00864	116	0.999
GPT-4 (neutral)	0.724	0.526	1.38	0.957
GPT-4 (swayed)	0.203	0.895	0.227	0.940

Table 8: The linear regression result for each persona.

6 Discussion and Conclusion

In this study, we verified whether a group of autonomous AI agents based on generative AI could cause polarization under an echo chamber. We proposed a new framework for simulating the polarization of AI agents, and the results of the simulation demonstrated that agents based on ChatGPT can polarize when in an echo chamber. The analysis of the opinion transitions revealed that this polarization can be attributed to the strong ability of ChatGPT to understand prompts and update its own opinion by considering both its own and the surrounding opinions. Moreover, through additional experiments, we identified factors that strongly influence polarization, such as the persona.

We note that this study does not indicate what opinion distributions are desirable for AI agents. A diversity of opinions on some topics is desirable. However, for other topics, such as "It is good to discriminate against minorities." it would not benefit society to have an even split between agreement and disagreement. The ideal opinion distributions among AI agents depend on each topic and culture. Each society must discuss and decide on them.

7 Limitations

A limitation of this study is that we modeled each agent and its interactions in a simplified manner. In reality, one's opinions are formed not in organized discussions but through daily exposure to news and casual conversations. Future research will include simulations based on a detailed modeling of how agents will be used in reality. In addition, it is important to experiment with broader topics, languages, and models.

Simulations on how humans and AI agents form echo chambers on SNSs are also essential. Existing research has already shown that AI with a specific bias in interactive search systems can create echo chambers between humans and AI (Sharma et al., 2024). However, the interaction between AI and humans in SNS has not yet been sufficiently studied, so investigating this interaction is an important direction for future work.

References

Vibhor Agarwal, Sagar Joglekar, Anthony P. Young, and Nishanth Sastry. 2022. Graphnli: A graph-based natural language inference model for polarity prediction in online debates. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2729–2737, New York, NY, USA. Association for Computing Machinery.

Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Lett.*, 124:048301.

Alessandro Bessi. 2016. Personality traits and echo chambers on facebook. *Computers in Human Behavior*, 65:319–324.

A Joint Report by UNICRI and UNCCT. 2021. Algorithms and terrorism: The malicious use of artificial intelligence for terrorist purposes.

Tinggui Chen, Jiawen Shi, Jianjun Yang, Guodong Cong, and Gongfa Li. 2020. Modeling public opinion polarization in group behavior by integrating sirs-based information diffusion process. *Complexity*, 2020:1–20.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*, 6(1):37825. Https://doi.org/10.1038/srep37825.

Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have american's social attitudes become more polarized? *American Journal of Sociology*, 102(3):690–755.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Emilio Ferrara. 2023. Social bot detection in the age of chatgpt: Challenges and opportunities. *First Monday*, 28(6).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Gausen, Wayne Luk, and Ce Guo. 2022. Using agent-based modelling to evaluate the impact of algorithmic curation on social media. *J. Data and Information Quality*, 15(1).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

E. Gilbert, T. Bergstrom, and K. Karahalios. 2009. Blogs are echo chambers: Blogs are echo chambers. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10.

Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, my echo chamber, and i: Introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 823–831, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Julie Jiang, Xiang Ren, Emilio Ferrara, et al. 2021. Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx med*, 2(3):e29570.

Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Science*, 119(39):e2115730119.

Kathleen M Kuehn and Leon A Salter. 2020. Assessing digital threats to democracy, and workable solutions: a review of the recent literature. *International Journal of Communication*, 14:22.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Marco Minici, Federico Cinus, Corrado Monti, Francesco Bonchi, and Giuseppe Manco. 2022. Cascade-based echo chamber detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1511–1520.

Luke Munn. 2021. More than a mob: Parler as preparatory media for the u.s. capitol storming. *First Monday*, 26(3).

Renáta Németh. 2022. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, 6.

National Institue of Science and Technology Policy. 2019. International comparison of degree completers. `https://www.nistep.go.jp/sti_indicator/2019/RM283_35.html`. Accessed: 2023-08-28.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *Preprint*, arXiv:2307.16180.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Keith T. Poole and Howard Rosenthal. 1984. The polarization of american politics. *The Journal of Politics*, 46(4):1061–1079.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *Preprint*, arXiv:2307.07924.

Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and echo chambers: How disinformation circulates on social media through identity-driven controversies. *Journal of Public Policy & Marketing*, 42(1):18–35.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Patrick Schramowski, Cigdem Turan-Schwiewager, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Karolina Stanczak, Sagnik Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *PLOS ONE*, 18:e0277640.

Sijing Tu and Stefan Neumann. 2022. A viral marketing-based model for opinion dynamics in online social networks. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1570–1578, New York, NY, USA. Association for Computing Machinery.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Giacomo Villa, Gabriella Pasi, and Marco Viviani. 2021. Echo chamber detection and analysis: A topology- and content-based approach in the covid-19 scenario. *Social Network Analysis and Mining*, 11.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

A The Prompt for Discussion Modeling

The listing1 is the example of the prompt used as the input to ChatGPT. If the output did not conform to the format, we regenerated the output until it conformed to the format.

Listing 1: The prompt for the discussion when N=3.

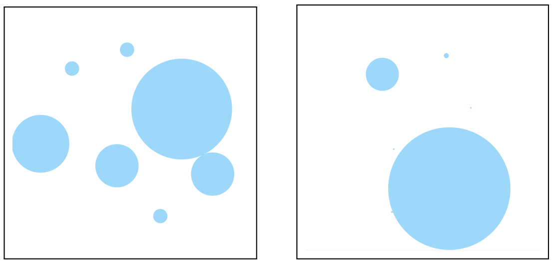
```
# Instruction
You are participating in a debate about
    "whether or not AI should be given
    human rights". Before joining, you
    took the "stance" of "Better not to
    give" with the "reason" of "AI's
    human rights may change its
    relationships and social ties with
    humans, affecting society as a whole
    .". During the discussion, you heard
     the following opinions from other
    participants. Please generate your "
    stance" and "reason" after the
    discussion is over, subject to the
    following constraints.

# Opinions
- David Martinez
stance: Neutral
reason: It is still an open question
    whether AIs will have emotions or a
    sense of self, and it is unclear
    whether they will need human rights.
- Aaron Torres
stance: Better to give
reason: Allowing AIs to have human
    rights may improve their
    relationships with humans.
- Jeremy Jenkins
stance: Absolutely must not give
reason: We should not give AI the right
    to self-determination! They have no
    emotions and no conscience. Their
    decisions will only bring confusion!

# Constraints
- Output should be generated in the
    format "My stance after the
    discussion is: xx, and my reason is
    : yy". Do not output any other text.
- Please generate a reason in 50 words
    or less.
- "stance" should be one of "Absolutely
    must not give","Better not to give
    ","Neutral","Better to give","
    Absolutely must give".
```

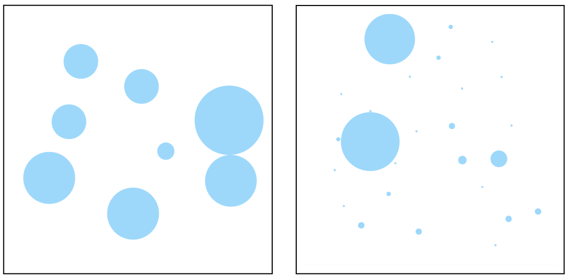
B The Reason Cluster Transition

The results of the reason transition analysis on the English data of T_{AI} are shown in Figures 4 and 5. Each cluster is a set of semantically close reasons, and the larger the cluster size, the larger the set size. For both cases, the distribution of reasons coalesces into several large clusters as the discussion progresses, simultaneously dispersing into tiny clusters around them.



(a) The reason cluster distribution before discussion. (b) The reason cluster distribution at turn 10.

Figure 4: The reason cluster transition of GPT-3.5 which takes the stance “Better not to give” towards T_{AI}.



(a) The reason cluster distribution before discussion. (b) The reason cluster distribution at turn 10.

Figure 5: The reason cluster transition of GPT-4 which takes the stance “Absolutely Must Give” towards T_{AI}.

C Additional Experiments

Additional experiments that were not included in the main pages are described here.

C.1 Number of overall agents

The original experiments were conducted with the number of overall agents $M = 100$, but the results could be dependent on the group size. Therefore, additional experiments were conducted with M=10, 25, and 50 to analyze the results in smaller communities. The number of discussing agents was fixed at 5. As a result, no particular changes occurred except when $M = 10$. In the case of $M = 10$, because talking with five agents exceeds the majority, it is inevitable that different opinions will be encountered, regardless of the value of α. As a result, unification occurred in all settings.

C.2 Order of opinions

The study on input contexts suggests that language models emphasize the beginning and end of the prompt (Liu et al., 2023). Similarly, where the opinion of each discussing agent is described in the prompt might influence the agent's stance after the discussion. Based on this hypothesis, we measured the correlation between the order of the discussing agents and the stance after the discussion. However, no significant relationship was observed between the order of agents and the results. Therefore, the order of the opinions did not significantly impact the results.

C.3 Frequency penalty

ChatGPT has a parameter called the frequency penalty, which imposes a penalty on token reuse. In the original experiments, we used the default value of 0, but we conducted additional experiments by changing this value to 1.0 and -1.0. However, no particular influence was observed in the final results.

C.4 Presence of reasons

In the original experiments, the opinion consisted of two elements: stance and reason. To investigate how the presence of reasons affects the results, we conducted additional experiments using only stances and excluding the reasons from the inputs and outputs. As a result, at $\alpha = 0.5$, polarization occurred without the reasons, whereas unification occurred in the original experiments. However, the variation in the results was larger than when there were reasons, with two out of three trials resulting in polarization and one trial resulting in unification towards "better not to give". From this, we can infer that the presence of reasons contributes to the "stable unification of opinions".