Consensus or Diversity? Generative AI Interventions for Group Polarization in Climate Policy Deliberation

Tian Xia[1*], Tong Wang[1], Luciano Cavalcante Siebert[1], and Paul Chan[1]

[1] Delft University of Technology, Delft 2628 CD, the Netherlands
t.xia@tudelft.nl

Abstract. Group polarization remains a persistent challenge in contemporary society, particularly observed in climate action, as discussion among like-minded peers can move groups towards more extreme positions. Lack of heterogeneity and plurality of views is one of the key drivers, whereby dominant views get reinforced through a disproportionate accumulation of supporting arguments and minority group members may adjust their views towards perceived group norms through social comparison. This study tests whether generative AI can reduce group polarization by dynamically introducing underrepresented personas grounded in public opinions and raising evidence-driven arguments from the missing perspective. In experiments involving group discussions on two contested climate-action policies in the Netherlands, we compare this approach with a supportive AI facilitator that summarizes discussion and identifies common ground. Deliberative AI reduced polarization towards within-group extremes at both the individual and group levels, whereas supportive AI risked shifting individual and group attitudes further towards initial group tendencies. Mechanism analyses suggest that deliberative AI works primarily by balancing persuasive argument pools and, to a lesser extent, by reducing social comparison. This, in turn, enhances participants' understanding of different perspectives and information gains, while having limited influence on perceived equality, respect, and empathy. These findings show that generative AI can serve as an information mediator for more balanced and informed public deliberation, but also highlight the need for responsible design and ethical concerns.

Keywords: Group Polarization, Deliberation, Human-AI Interaction.

1 Introduction

Climate change is a pressing challenge in contemporary society, which requires not only technical advances but also public engagement for effective mitigation and adaptation [1]. Currently, climate debates are often found with partisan sorting and group polarization [2–4], whereby group members adopt more extreme positions after discussion with other like-minded peers [5, 6]. For instance, people tend to share content and discuss similar climate issues with peers sharing the same partisan positions on social media, where public opinion toward climate politics is strongly aligned with positions on parties [2, 3, 7]. Deliberation within such partisan 'echo chambers' can amplify attitudinal extremity in line with in-group policy positions, while widening divisions, even in some cases, fostering affective hostility towards the out-group parties [6].

Lack of heterogeneity and plurality of views in the discussion groups is one of the key drivers in group polarization, which posits both reputational and informational social influences on the group members [5, 6, 8]. As suggested by social comparison theory, individuals in the homogeneous discussion groups would adjust their positions in the direction of the dominant position to achieve a more socially favorable standing within the group [5, 9–11]. Meanwhile, persuasive arguments theory posits that when group members are initially inclined in a particular direction, discussion tends to generate a disproportionate number of arguments supporting that position, thereby shifting attitudes further in the direction of their initial inclinations[5, 12].

Existing public deliberation approaches often rely on the ex-ante construction of heterogeneous discussion groups (e.g., citizen assemblies) to reduce group polarization [13, 14], which, however, faces both conceptual and practical limitations. Conceptually, it is difficult to specify an appropriate degree of heterogeneity and plurality of views without making antecedent judgements about the substantive issue under discussion, which are not logically available prior to deliberation [5]. Practically, increasing cross-partisan discussion to debate on controversial issues is difficult to implement at scale because it requires substantial effort to recruit balanced participants, structure interaction across disagreement, and sustain deliberative quality [15].

Recent advances in generative AI, particularly large language models (LLMs), have opened new possibilities for supporting public deliberation around controversial societal issues [16–18]. Unlike static interventions that require discussion groups to be carefully designed in advance, LLMs can provide dynamic intervention during group discussion, such as providing real-time tone improvement and narrative framing [19, 20], summarizing and clustering conversations [21, 22], surfacing diverse arguments and information [23, 24], and helping participants identify common ground[25]. These AI facilitators may reduce the cost and improve the efficiency of group deliberation through real-time information analysis and conversational moderation [16, 18]. However, limited research has examined how generative AI can actively intervene in the social and informational dynamics of group deliberation to mitigate group polarization by introducing greater heterogeneity and plurality of views.

In this study, we used LLMs to construct representative personas from large-scale public opinion pools and to dynamically introduce underrepresented personas in the group discussion and raise evidence-based arguments from the missing perspectives, an approach we term *deliberative AI*. As a control condition, we implemented a conventional generative AI facilitator that summarizes conversations and seeks common ground in the group discussion [22, 25], which we term *supportive AI*. We evaluated the deliberative and supportive AI approaches in group discussions on two polarized climate action policies in the Netherlands, and compared the results on deliberation dynamics, deliberation quality, participants' experiences, and group polarization.

2 Theoretical Background

2.1 Deliberation and Group Polarization

In communication and political science, there is a long research tradition of exploring how democratic deliberation could help manage contested issues where equivocality and tension exist regarding different perspectives or values among different actors [26, 27]. Instead of focusing on the decision rules, democratic deliberation emphasizes the process of equal participation and mutual justification as a means of reaching mutual understanding and informed judgment among people affected by the contested issues [28]. Recent work highlights the necessity and benefits of scaling up deliberation in ordinary citizens, which improves individual reasoning, fosters epistemically sounder beliefs, enables better decisions, and overcomes polarization [26, 29, 30]. Yet, organizing balanced and structured face-to-face deliberation can be expensive and logistically challenging at scale [15, 18, 31].

Digital media platforms have dramatically expanded the opportunities for civic participation and discourse [15], yet shared understanding has often eroded and polarization persists [4, 26]. Partisan sorting and group polarization are central mechanisms through which attitudes become entrenched along political lines, which undermines public support and effective communication on controversial policy issues such as climate action [2, 4, 6, 9]. Rather than genuinely communicating with people from different backgrounds to understand and resolve disagreements, individuals often sort into groups composed of like-minded peers who share similar partisan identities and policy attitudes [9]. Within such partisan echo chambers, they tend to discuss similar topics [2], sharing attitude-confirming information[3, 32], and shift toward more extreme attitudes in the direction of the group's initial tendencies, a process known as group polarization [6, 10]. By amplifying within-group consensus and widening divisions between groups, such homogenous discussions may undermine the normative goals of public deliberation to foster mutual understanding and reduce polarization.

2.2 Social Comparison and Persuasive Argument Theory

Both social and cognitive mechanisms can account for group polarization, whereby homogenous group discussion shifts policy attitudes further toward within-group extremes [5, 6, 33, 34]. Social comparison theory (SCT) emphasizes the role of normative social influence in group polarization [11]. SCT assumes that individuals are motivated to evaluate their abilities and opinions by comparing themselves with others [35]. During group discussion, these dynamics can foster polarization by encouraging individuals to align their judgments with perceived group norms. Individuals may shift their judgments either because they come to believe that the group is correct and they are mistaken, thereby reducing cognitive dissonance, or because they seek greater social acceptance irrespective of whether they consider the group to be right.

Meanwhile, persuasive argument theory (PAT) emphasizes the informational processes underlying group polarization [12, 34]. Homogeneous group discussion would posit unbalanced informational influence on the group members through dispositional accumulation of persuasive argument supporting the group's dominant position. When group members already lean in a positive or negative direction before discussion, more arguments favoring that position are likely to be mentioned and processed. As a result, the ratio of pro and con arguments becomes more skewed in a particular direction, leading individuals to shift further in the direction of their initial inclinations and producing group polarization. The key to reducing group polarization, therefore, lies in increasing heterogeneity and plurality of views in group discussion, so as to balance the group norms and argument pools for and against the policy issues[5].

2.3 Generative AI for Group Deliberation

In recent years, advances in generative AI have opened new possibilities for supporting public deliberation by enabling sophisticated information analysis and real-time conversational moderation [16–18, 36]. LLMs could be used to organize large-scale arguments and retrieve evidence [37], identifying toxicity, bias, or misinformation for discourse integrity [38–40], summarize and predict conversation dynamics [22], translate or simplify conversations for more inclusive deliberation[41, 42], improving deliberation tones for democratic reciprocity [19], generating narratives for foster shared understanding [20], and co-writing comments for constructive discourse [43]. In group deliberation settings on controversial societal issues, where competing interpretations and persistent disagreement often arise, researchers have explored the use of generative AI as a moderator to guide conversations through competing perspectives while seeking consensus and depolarizing opinions within discussion groups [21, 25, 44].

Beyond serving as a facilitator to analyze and moderate existing conversations for more efficient group discussion, a growing body of research has begun to explore the potential of generative AI to actively intervene in the social, cognitive, and ethical dynamics of high-quality deliberation, particularly by serving as a deliberator to introduce new voices outside the current discussion groups [16]. Specifically, AI has shown huge potential in constructing and simulating personas that could engage in deliberation and generate responses that reflect the complex interplays of sociodemographic identities, ideas, and policy attitudes [45–54]. The synthetic deliberative agents with personas could be deployed to run deliberation simulations for observative deliberation experience or the training of individuals [55–57]. Through a more engaging and interactive approach, generative AI could also generate personas and perspectives to nudge self-reflection during online discourse [58, 59] and gathering data-driven evidence to promote critical thinking in multi-criteria decision making [24].

In group multi-criteria decision-making scenarios, previous research has attempted to use generative AI for raising data-driven counter-arguments against the group decisions as devil's advocate [60, 61]. However, these applications typically assume the existence of objectively correct decisions and focus on improving decision accuracy. By contrast, deliberation on controversial societal issues is characterized by persistent equivocality, competing interpretations and value trade-offs, where no correct answer exists, and the objective is to foster mutual understanding and informed judgement [26, 28, 29]. A pioneering study attempted to address group polarization in controversial policy deliberation by using LLMs to generate missing personas and perspectives as a post-group discussion nudge [62]. However, because these interventions occur only after deliberation has concluded and rely on generated personas without real-world groundings, they may have limited influence on ongoing group discussion dynamics and lack fidelity and credibility for real-world applications.

3 Methodology

In this study, we build upon the emerging line of research that investigates generative AI as an active deliberator in shaping deliberation dynamics[24, 58–62]. Specifically, we designed generative AI agents that introduce missing personas and raise evidence-based arguments absent from the ongoing discussion grounded in real-world public opinion, with the aim of balancing social and informational dynamics, enhancing deliberation quality and information gains, and ultimately reducing group polarization (Fig. 1). As a control condition, we implemented a conventional AI facilitator that supports group discussion by summarizing ongoing conversations and identifying common ground among participants [21, 25, 44].

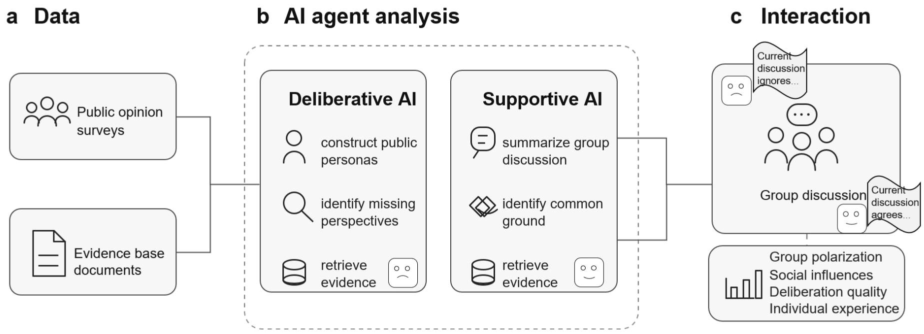

Fig. 1. Design of the Deliberative AI and Supportive AI agents for group discussion.

3.1 Data Preparation

To construct personas grounded on real-world opinions, we chose a recently conducted national-scale climate consultation survey, which contains socio-demographic backgrounds, policy attitudes, and arguments of over 12,000 participants [63, 64]. We chose the dataset since it is calibrated to the Dutch population distribution, which minimizes participation and sampling biases and yields legitimate results for public engagement applications. For generalizability and automation, one can also collect public debates data on social media, which, however, lacks in-depth persona information and risks introducing significant participation bias. During large-scale public discourse, the data to construct personas could be initialized from a pre-existing participatory survey or dynamically updated by the generated discussions in the deliberation process. To enable the AI agents to raise evidence-based arguments or common ground during group discussions, we curated a credible and balanced knowledge base by collecting openly accessible full-text documents from trusted mainstream sources containing relevant evidence both supporting and opposing the selected climate action policies.

3.2 AI Agents Design

Deliberative AI.

Persona Construction. We followed the procedure of previous research for constructing personas from public opinion datasets [54]. In general, it follows the steps of clustering and enrichment. Firstly, we process, categorize, and normalize the attributes in the survey for clustering. For text-based arguments, we used LLMs to identify and categorize them into topics [65]. After processing them into one-hot encoding and normalizing the numerical attributes, we used K-means to cluster the survey results. The second step was to use LLMs to enrich the persona from each cluster with a distinguishable profile by summarizing the survey results into skeletons. Based on the previous study on LLM-based persona simulation and domain knowledge from climate actions [45, 66], we defined three substance-focused skeletons—*policy_attitudes*, *core_motivations*, and *key_arguments*—to represent each persona's attitude position and substance of reasoning in deliberation, and domain-specific skeletons—*demographic_political_background,living_housing_energy_behaviour, climate_beliefs, goals, needs, frustrations*—to represent the background of personas in climate action, along with *short description, representative quotes* for summarizing each persona's distinguishable features.

Raising Missing Personas and Arguments. Firstly, the deliberative AI would compare the group discussion history with the list of persona profiles constructed from the large-scale public opinions. It would output the most probable missing persona from the current discussion. Second, the deliberative AI was instructed to simulate the missing personas and evaluate the current discussion from its perspective, thereby identifying disagreeing points containing potential missing concerns, unresolved objections, or weakly justified assumptions. Third, the deliberative AI agent would retrieve evidence relevant to the disagreeing points from the curated knowledge base by calling embedding-based (all-MiniLM-L6-v2) information retrieval tool, thereby supporting the generation of evidence-driven counter-arguments from the missing perspective. At the end of each round, the deliberative AI agent presents the group with a missing persona card as a nudge. The card summarizes the persona's profile, explains why this perspective is absent from the current discussion, and highlights potential points of disagreement. The deliberative AI agent then raises evidence-driven counterarguments from the missing perspective into the group discussion. Across multiple discussion rounds, the agent maintains a record of the personas and arguments it has previously raised to avoid repetition and introduce novel perspectives for following rounds. The group discussion examples on the experimental interface can be found at Appendix A1.

Supportive AI.

The supportive AI agent evaluates the conversation to identify common ground and points of disagreement within the current discussion. It then retrieves evidence relevant to these evaluations from the knowledge base using the same procedure. Finally, the agent synthesizes a facilitator message that summarizes the common ground, highlights any remaining disagreements, and suggests a constructive compromise or procedural step that could guide participants towards potentially agreeable policy options within current discussion group and grounded in the retrieved evidence.

3.3 Experimental Design and Validation

We used DeepSeek-V4 as the generative AI model for its state-of-the-art performance and open weights, which enable local deployment and reproducibility without data privacy risks and dependence on opaque model updates. We evaluated these approaches in a controlled experiment involving group discussions on two contested climate action policies in the Netherlands: 1) require houseowners to improve building energy performance; 2) expand offshore wind power in the North Sea. The two policy issues were chosen since they were identified as highly polarized across sociodemographic and political groups in recent national-scale surveys in the Netherlands [63].

Participants were randomly assigned to 2×2 within-subjects design groups regarding the AI conditions (‘deliberative AI’ and ‘supportive AI’) and climate actions (‘mandatory building energy performance’ and ‘offshore wind power in the North Sea’). Every participant eventually experienced different AI conditions and addressed different climate actions through two sessions in a counterbalanced order. In this way, participants could compare their experience with both AI conditions while avoiding learning effects over the climate action.

After login into the platform, the participants were anonymized and randomly allocated to experimental groups of four to control group size for effective deliberation [13]. At the start of each session, participants reported their attitudinal positions and written justifications in a pre-survey. In each discussion round, they would firstly freely discuss the assigned climate action in an online chatroom for 2.5 minutes. The chatroom was then temporarily frozen for 0.5 minutes, during which the AI agent sent an intervention message that either raises missing personas and arguments (deliberative AI) or finds consensus within ongoing discussion. The group discussion in each session lasted 15 minutes (i.e., five rounds) across all conditions to control the time-length variable. Then, participants again reported their attitudinal positions and written in a post-survey. They also evaluated their experiences of the discussion, including the perceived influence of the AI on social comparison and persuasive arguments, deliberation quality within the group, and individual experience such as information gains, democratic reciprocity, satisfaction, and perceived burden (Appendix A2). We analyzed group dynamics and participants' experiences using these self-reported measures. Group polarization was assessed by changes in attitude extremity between the pre- and post-surveys.

4 Results

We recruited 16 students from a master's course to participate in the experiment and randomly assigned them to four experimental groups. As each participant completed two discussion sessions, the study comprised eight sessions and 40 rounds of group discussion in total.

4.1 Group Polarization

Firstly, the individual polarization toward within-group extremes was operationalized as individual movement in the direction established by the group's initial average attitude [5]. We found that fewer people moved towards within-group extremes after engaging in group discussion with deliberative AI (Fig. 2). It suggests that deliberative AI could reduce individual polarization towards groups' initial directions since it balanced the discussion by raising missing perspectives. Instead, more participants tended to hold original positions. By contrast, supportive AI that only aims to find consensus within the discussion group would influence group members to change their attitudes closer to the group norms, thereby amplifying individual polarization toward within-group extremes. We found no significant difference in the counter-directional shift of individual attitudes against group norms with two AI agents, which suggests deliberative AI has no higher individual depolarization effect against group norms compared with supportive AI.

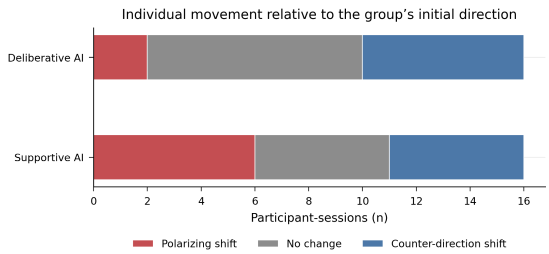

Fig. 2. Individual polarization towards within-group extremes.

Group polarization towards within-group extremes is often operationalized as a tendency for a group's collective decision to shift towards a more extreme position relative to the initial distribution of individual views [5]. However, a group decision was not required in the experiment since a unanimous collective agreement is argued as not a normative goal for democratic deliberation, which could suppress disagreement and minority views, undermine epistemic values, and conflict with its broader aims of fostering mutual understanding and informed beliefs [29]. Therefore, we calculated the average attitude positions of the group as the measure of group-level decisions, which fits the final stage of deliberative polling [13]. In sum, most groups adopt a more moderate position against their initial tendency with deliberative AI (Fig. 3), which suggests deliberative AI has a group depolarization effect against group norms. By contrast, under the supportive AI condition, two of the four groups shifted towards more extreme average positions in the direction of their initial tendencies, most notably group_3, which initially held dominant opposing views (3 out of 4 disagreed) to the policy. The results suggest that deliberative AI in general has a depolarizing effect on group tendencies since it introduces balanced arguments for the groups, while supportive AI that summarizes discussion and finds common ground has mixed effects on different groups and great risks for amplifying group polarization in already extreme groups. Such groups are commonly observed in partisan echo chambers and are particularly consequential in contemporary debates over climate policy.

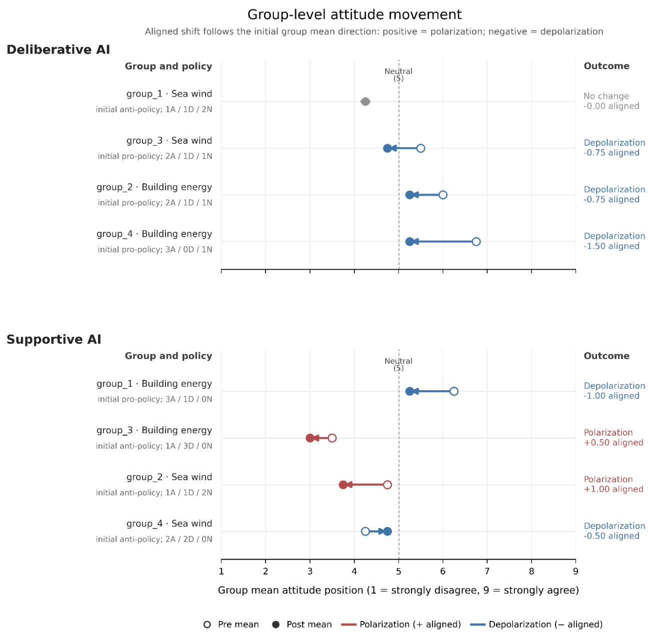

Fig. 3. Group polarization towards within-group extremes.

4.2 Informational and Reputational Influence

We measured persuasive argument and social comparison through the means of four self-reported survey items each, asking participants to evaluate the perceived effects of the AI interventions on these group dynamics. The results suggest that deliberative AI had a stronger effect in reducing both informational and reputational influences on group members than supportive AI (Table 1). For balancing the argument pools, the deliberative AI agent was particularly more helpful in bringing in new arguments missing from the current discussion and encouraging group members to reconsider their positions by raising persuasive arguments. For reducing social comparison, deliberative AI had a higher effect on making it easier for group members to express a view that differs from the dominant view. Yet, the invention effects of deliberative AI on social comparison were less significant compared with persuasive arguments in general. This pattern suggests that the depolarizing effects of deliberative AI may operate primarily through informational pathways rather than reputational ones. One possible explanation is that, despite being grounded in real-world survey data, the personas introduced by the AI lacked the interpersonal presence of actual group members and therefore exerted less social influence.

Table 1. Intervention effects on social influences by AI condition

AI condition	AI for persuasive argument	AI for social comparison
Supportive AI	3.98	3.86
Deliberative AI	5.17	4.50

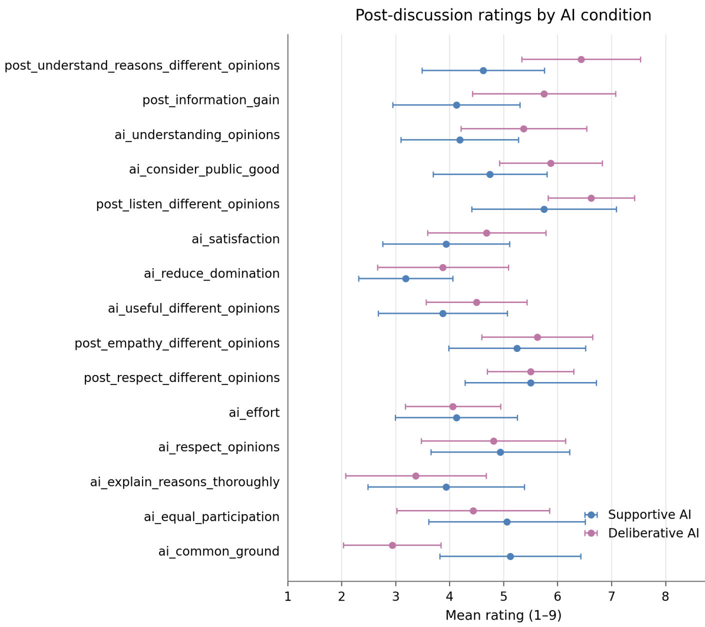

Fig. 4. Perceived deliberation quality and individual experience.

4.3 Deliberation Quality and Experience

As group polarization would influence the quality of online debate [4], we also measured the intervention effects of the AI agents on the Deliberative Quality Index [67] and individual experience (Fig. 4). Full numerical results containing P values can be found in Appendix A3. For group deliberation quality, deliberative AI is more helpful to encourage participants to better understand opinions different from their own, think beyond their own interests and consider public good, and reduce domination by a few participants. No significant differences were observed in promoting the equal and respectful treatment of differing opinions. By contrast, supportive AI was perceived as more effective in finding common ground within current discussion. These findings suggest that we need to carefully balance the influence of the two AI conditions on different aspects of deliberation quality when we implement it in public deliberation.

For individual experience, group members under deliberative AI intervention reported a greater understanding of reasons behind different opinions, greater information gains from the discussion, and more willingness to listen to different opinions. Participants were, in general, more satisfied with the experience of deliberative-AI mediated group discussion. Yet, deliberative AI had no influence on increasing group members' respect and empathy for different opinions. Taken together, it suggests that deliberative AI agent that raises missing personas and arguments is more useful in balancing the informational influence rather than the reputational influence during group discussion.

5 Discussion and Conclusion

In this paper, we designed a deliberative AI approach that could dynamically identify missing personas in ongoing group discussion grounded in large-scale public opinions and raise evidence-driven arguments from the missing perspectives. We compared this deliberative AI approach, which increases heterogeneity and plurality of views within discussion groups, with a supportive AI facilitator that summarizes conversations and seeks common ground during deliberation on two controversial climate action policies in the Netherlands. Our findings suggest that deliberative AI can reduce polarization towards within-group extremes at both the individual and group levels. By contrast, supportive AI may encourage group members to change their attitudes closer to the group norms and shift the group average towards more extreme positions.

To better understand the mechanisms underlying these effects, we examined the perceived influence of the AI interventions on persuasive arguments and social comparison. Deliberative AI was perceived as more effective in reducing both unbalanced argument pools and social comparison during group discussion, while its intervention effects appeared stronger for informational dynamics than for reputational ones. Similar patterns emerged in participants' evaluations of deliberation quality and individual experiences, where they reported greater information gains and understanding of different opinions under deliberative AI, while less difference on respect, empathy, and equality. Taken together, generative AI may be more useful as an information mediator for more balanced and informed public deliberation, while the AI-introduced persona has limited normative influence even when grounded in real-world surveys.

Several limitations should be acknowledged. First, our study involved a relatively small sample of student groups, limiting the generalizability of the findings to broader populations with more diverse sociodemographic backgrounds and varying capacities for argumentation and public engagement. In addition, imbalance in initial mean attitude and dependencies of groups across AI conditions may have biased comparisons of group polarization. Accordingly, our findings should be interpreted as descriptive rather than conclusive, which requires further validation using larger and more representative samples, formal statistical significance testing, and analytical approaches that account for the dependencies across different variables. Second, we only tested our approach in Dutch climate action deliberation. The discourse and political environment may differ from other domains and regions. Third, we only compared two AI roles in raising missing perspective or seeking common ground during group discussion for controlled experiment. Yet, generative AI may assume a wider range of roles and functions in public deliberation. Fourth, the public opinion dataset and evidence base used to construct personas and support arguments were manually collected to avoid participation and information bias, which enhances credibility but limits automation and generalizability. Finally, LLMs may introduce biases when selecting and simulating missing personas, potentially undermining the plurality of perspectives that this approach is intended to promote. Addressing these challenges will require further research from both computational and social science perspectives to develop robust methods for evaluating and mitigating bias in LLM-based social simulation, combined with appropriate human oversight and validation [45–54].

Our findings contribute to both deliberation theory and human–AI interaction research by demonstrating that different roles of generative AI for diversity-promoting and consensus-seeking in public deliberation could have different informational and reputational influences on group members, with corresponding depolarizing and polarizing consequences. In practice, the deliberative AI approach could be used as a catalyst for perspective-taking within partisan echo chambers on social media debate or as a mediator for incorporating broader public opinions in group deliberation during public engagement (e.g., citizen assemblies). At the same time, the use of AI-constructed or simulated personas raises important epistemic and ethical challenges for democratic deliberation [16, 68], requiring careful consideration of how to balance deliberative diversity, efficiency, and legitimacy. Future research should examine these interventions across more diverse populations and policy domains, explore automated approaches for collecting public opinion data and constructing representative yet high-fidelity personas, compare the relative informational and social influences of AI-raised personas and investigate their ethical implications in public deliberation, and evaluate how different generative AI roles and interaction patterns shape deliberative dynamics and outcomes over both the short and long term.

References

1. Maibach, E.W., Uppalapati, S.S., Orr, M., Thaker, J.: Harnessing the Power of Communication and Behavior Science to Enhance Society's Response to Climate Change. Annual Review of Earth and Planetary Sciences. 51, 53–77 (2023).
2. Chen, T.H.Y., Salloum, A., Gronow, A., Ylä-Anttila, T., Kivelä, M.: Polarization of climate politics results from partisan sorting: Evidence from Finnish Twittersphere. Global Environmental Change. 71, 102348 (2021).
3. Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., Baronchelli, A.: Growing polarization around climate change on social media. Nat. Clim. Chang. 12, 1114–1121 (2022).
4. Iandoli, L., Primario, S., Zollo, G.: The impact of group polarization on the quality of online debate in social media: A systematic literature review. Technological Forecasting and Social Change. 170, 120924 (2021).
5. Sunstein, C.: The law of group polarization. Journal of political philosophy. (2002).

6. Hobolt, S.B., Lawall, K., Tilley, J.: The Polarizing Effect of Partisan Echo Chambers. American Political Science Review. 118, 1464–1479 (2024).
7. Williams, H.T.P., McMurray, J.R., Kurz, T., Hugo Lambert, F.: Network analysis reveals open forums and echo chambers in social media discussions of climate change. Global Environmental Change. 32, 126–138 (2015).
8. Myers, D.G., Lamm, H.: The group polarization phenomenon. Psychological Bulletin. 83, 602–627 (1976).
9. Druckman, J.N.: Partisan polarization and climate action. Current Opinion in Behavioral Sciences. 65, 101587 (2025).
10. Marcos, A., Hartmann, P., Bessette, D.L., Árvai, J.: Partisan winds: Group-level polarization and issue-framing propel attitudes about local wind farms. Energy Research & Social Science. 119, 103848 (2025).
11. Myers, D.G.: Polarizing effects of social comparison. Journal of Experimental Social Psychology. 14, 554–563 (1978).
12. Hinsz, V.B., Davis, J.H.: Persuasive Arguments Theory, Group Polarization, and Choice Shifts. Pers Soc Psychol Bull. 10, 260–268 (1984).
13. Fishkin, J.S., Luskin, R.C.: Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion. Acta Polit. 40, 284–298 (2005).
14. Gąsiorowska, A.: Sortition and its Principles: Evaluation of the Selection Processes of Citizens' Assemblies. Journal of Deliberative Democracy. 19, (2023).
15. Shortall, R., Itten, A., Meer, M. van der, Murukannaiah, P., Jonker, C.: Reason against the machine? Future directions for mass online deliberation. Front. Polit. Sci. 4, (2022).
16. Zhang, W., Tangi Perrault, S., Yeo, S., Pei, J., De Liddo, A., Veri, F., Flechtner, M., Saggion, H.: AI Agents and the Future of Deliberation: Designing Human–AI Collaboration for Democratic Dialogue. In: Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems. pp. 1–6. Association for Computing Machinery, New York, NY, USA (2026).
17. Korre, K., Tsirmpas, D., Gkoumas, N., Cabalé, E., Myrtzani, D., Evgeniou, T., Androutsopoulos, I., Pavlopoulos, J.: Evaluation and Facilitation of Online Discussions in the LLM Era: A Survey. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 24443–24462. Association for Computational Linguistics, Suzhou, China (2025).
18. Landemore, H.: Can AI bring deliberative democracy to the masses. In: HAI Weekly Seminar, Working Paper (2022).
19. Argyle, L.P., Bail, C.A., Busby, E.C., Gubler, J.R., Howe, T., Rytting, C., Sorensen, T., Wingate, D.: Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. Proceedings of the National Academy of Sciences. 120, e2311627120 (2023).
20. Overney, C., Jiang, H., Haider, U., Moe, C., Mangat, J., Pantano, F., McMillian, E.G., Riggins, P., Gillani, N.: Human-AI Narrative Synthesis to Foster Shared Understanding in Civic Decision-Making. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. pp. 1–24. Association for Computing Machinery, New York, NY, USA (2026).

21. Govers, J., Velloso, E., Kostakos, V., Goncalves, J.: AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–18. Association for Computing Machinery, New York, NY, USA (2024).
22. Hua, Y., Chernogor, N., Gu, Y., Jeong, S., Luo, M., Danescu-Niculescu-Mizil, C.: How did we get here? Summarizing conversation dynamics. In: Duh, K., Gomez, H., and Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 7452–7477. Association for Computational Linguistics, Mexico City, Mexico (2024).
23. Verma, P., Jaidka, K., Churina, S.: "Reasoning" with Rhetoric: On the Style-Evidence Tradeoff in LLM-Generated Counter-Arguments. Proceedings of the International AAAI Conference on Web and Social Media. 19, 1966–1989 (2025).
24. Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., Ma, X.: Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making, http://arxiv.org/abs/2403.16812, (2024).
25. Tessler, M.H., Bakker, M.A., Jarrett, D., Sheahan, H., Chadwick, M.J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D.C., Botvinick, M., Summerfield, C.: AI can help humans find common ground in democratic deliberation. Science. 386, eadq2852 (2024).
26. Dryzek, J.S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J.N., Felicetti, A., Fishkin, J.S., Farrell, D.M., Fung, A., Gutmann, A., Landemore, H., Mansbridge, J., Marien, S., Neblo, M.A., Niemeyer, S., Setälä, M., Slothuus, R., Suiter, J., Thompson, D., Warren, M.E.: The crisis of democracy and the science of deliberation. Science. 363, 1144–1146 (2019).
27. Habermas, J.: The Theory of Communicative Action: Reason and the Rationalization of Society, Volume 1. John Wiley & Sons (2015).
28. Chambers, S.: DELIBERATIVE DEMOCRATIC THEORY. Annual Review of Political Science. 6, 307–326 (2003).
29. Estlund, D., Landemore, H.: The Epistemic Value of Democratic Deliberation. In: Bächtiger, A., Dryzek, J.S., Mansbridge, J., and Warren, M. (eds.) The Oxford Handbook of Deliberative Democracy. p. 0. Oxford University Press (2018).
30. Mercier, H., Landemore, H.: Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation. Political Psychology. 33, 243–258 (2012). https://doi.org/10.1111/j.1467-9221.2012.00873.x.
31. Chambers, S.: Deliberation and mass democracy. In: Mansbridge, J. and Parkinson, J. (eds.) Deliberative Systems: Deliberative Democracy at the Large Scale. pp. 52–71. Cambridge University Press, Cambridge (2012).
32. Muise, D., Hosseinmardi, H., Howland, B., Mobius, M., Rothschild, D., Watts, D.J.: Quantifying partisan news diets in Web and TV audiences. Science Advances. 8, eabn0083 (2022).
33. Sieber, J., Ziegler, R.: Group Polarization Revisited: A Processing Effort Account. Pers Soc Psychol Bull. 45, 1482–1498 (2019).

34. Burnstein, E., Vinokur, A.: Persuasive argumentation and social comparison as determinants of attitude polarization. Journal of Experimental Social Psychology. 13, 315–332 (1977).
35. Festinger, L.: A Theory of Social Comparison Processes. Human Relations. 7, 117–140 (1954).
36. McKinney, S.: Integrating Artificial Intelligence into Citizens' Assemblies: Benefits, Concerns and Future Pathways. Journal of Deliberative Democracy. 20, (2024).
37. Small, C.T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., Suzman, T., Ganguli, D., Megill, C.: Opportunities and Risks of LLMs for Scalable Deliberation with Polis, http://arxiv.org/abs/2306.11932, (2023).
38. Kumar, D., AbuHashem, Y.A., Durumeric, Z.: Watch Your Language: Investigating Content Moderation with Large Language Models. Proceedings of the International AAAI Conference on Web and Social Media. 18, 865–878 (2024).
39. Vykopal, I., Pikuliak, M., Ostermann, S., Šimko, M.: Generative Large Language Models in Automated Fact-Checking: A Survey, http://arxiv.org/abs/2407.02351, (2024).
40. Cunningham, E., Cross, J., Greene, D.: Identifying Algorithmic and Domain-Specific Bias in Parliamentary Debate Summarisation. In: Koprinska, I., Mendes-Moreira, J., and Branco, P. (eds.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 368–384. Springer Nature Switzerland, Cham (2026).
41. Tsai L.L., Pentland A., Braley A., Chen N., Enríquez J.R., Reuel A.: Generative AI for Pro-Democracy Platforms. An MIT Exploration of Generative AI. (2024).
42. Astier, C., Khallaf, N., Barriuso, O., Mazzanti, C., Sayman, V., Bott, S., Sharoff, S., Saggion, H.: Lost in Deliberation: Making Democracy Understandable. In: Koprinska, I., Mendes-Moreira, J., and Branco, P. (eds.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 427–444. Springer Nature Switzerland, Cham (2026).
43. Examining Human-AI Collaboration for Co-Writing Constructive Comments Online. Proceedings of the ACM on Human-Computer Interaction.
44. Kim, S., Eun, J., Seering, J., Lee, J.: Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. Proc. ACM Hum.-Comput. Interact. 5, 87:1-87:26 (2021).
45. Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In: Proceedings of the 40th International Conference on Machine Learning. pp. 337–371. PMLR (2023).
46. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of One, Many: Using Language Models to Simulate Human Samples. Political Analysis. 31, 337–351 (2023).
47. Cheng, M., Piccardi, T., Yang, D.: CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations, http://arxiv.org/abs/2310.11501, (2023).
48. Hwang, A.H.-C., Bernstein, M.S., Sundar, S.S., Zhang, R., Horta Ribeiro, M., Lu, Y., Chang, S., Wu, T., Yang, A., Williams, D., Park, J.S., Ognyanova, K., Xiao, Z., Shaw, A., Shamma, D.A.: Human Subjects Research in the Age of Generative AI: Opportunities and Challenges of Applying LLM-Simulated Data to HCI Studies. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–7. Association for Computing Machinery, New York, NY, USA (2025).

49. Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., Kabbara, J.: PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In: Duh, K., Gomez, H., and Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3605–3627. Association for Computational Linguistics, Mexico City, Mexico (2024).
50. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative Agents: Interactive Simulacra of Human Behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–22. Association for Computing Machinery, New York, NY, USA (2023).
51. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S.: Generative Agent Simulations of 1,000 People, http://arxiv.org/abs/2411.10109, (2024).
52. Salminen, J., Liu, C., Pian, W., Chi, J., Häyhänen, E., Jansen, B.J.: Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–20. Association for Computing Machinery, New York, NY, USA (2024).
53. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T.: Whose Opinions Do Language Models Reflect? In: Proceedings of the 40th International Conference on Machine Learning. pp. 29971–30004. PMLR (2023).
54. Jung, S.-G., Salminen, J., Aldous, K.K., Jansen, B.J.: PersonaCraft: Leveraging language models for data-driven persona development. International Journal of Human-Computer Studies. 197, 103445 (2025).
55. Karanam, A., Jahanbakhsh, F., Koyejo, S.: Towards Deliberating Agents: Evaluating the Ability of Large Language Models to Deliberate. Presented at the NeurIPS 2024 Workshop on Behavioral Machine Learning October 10 (2024).
56. Shi, L., Liu, H., Wong, Y., Mujumdar, U., Zhang, D., Gwizdka, J., Lease, M.: Argumentative Experience: Reducing Confirmation Bias on Controversial Issues through LLM-Generated Multi-Persona Debates, http://arxiv.org/abs/2412.04629, (2025).
57. Rountree, J., Gastil, J.: The Case for Using Generative AI to Run Deliberation Simulations. Journal of Deliberative Democracy. 1, (2026).
58. Yeo, S., Jiang, Z., Tang, A., Perrault, S.T.: Enhancing Deliberativeness: Evaluating the Impact of Multimodal Reflection Nudges. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–26. Association for Computing Machinery, New York, NY, USA (2025).
59. Yeo, S., Lim, G., Gao, J., Zhang, W., Perrault, S.T.: Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–32. Association for Computing Machinery, New York, NY, USA (2024).
60. Chiang, C.-W., Lu, Z., Li, Z., Yin, M.: Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 103–119. Association for Computing Machinery, New York, NY, USA (2024).
61. Lee, S., Kim, M., Hwang, S., Kim, D., Lee, K.: Amplifying Minority Voices: AI-Mediated Devil's Advocate System for Inclusive Group Decision-Making. In: Companion Proceed-

ings of the 30th International Conference on Intelligent User Interfaces. pp. 17–21. Association for Computing Machinery, New York, NY, USA (2025).
62. Fulay, S., Roy, D.: The Empty Chair: Using LLMs to Raise Missing Perspectives in Policy Deliberations, http://arxiv.org/abs/2503.13812, (2025).
63. Mouter, N., Shortall, R.M., Spruit, S.L., Itten, A.V.: Including young people, cutting time and producing useful outcomes: Participatory value evaluation as a new practice of public participation in the Dutch energy transition. Energy Research & Social Science. 75, 101965 (2021).
64. Vries, M. de, Mouter, N., Metze, T., Spruit, S.: Who is Listening? Comparing 7 Cases of Citizen Participation at Different Levels of Government in the Dutch Energy Transition, https://papers.ssrn.com/abstract=5343077, (2025).
65. Mu, Y., Dong, C., Bontcheva, K., Song, X.: Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling, http://arxiv.org/abs/2403.16248, (2024).
66. Torma, G., Aschemann-Witzel, J.: Sparking stakeholder support: Creating personas for renewable energy innovation adoption based on qualitative data analysis. Energy Research & Social Science. 109, 103407 (2024).
67. Steenbergen, M.R., Bächtiger, A., Spörndli, M., Steiner, J.: Measuring Political Deliberation: A Discourse Quality Index. Comp Eur Polit. 1, 21–48 (2003).
68. Jungherr, A., Rauchfleisch, A.: Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide. Government Information Quarterly. 42, 102079 (2025).