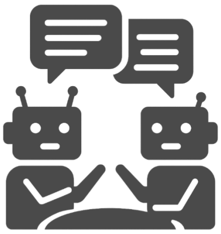
# Using Debating LLMs to Reduce Bias: A Deliberative Approach to AI-Assisted Civic Decision-Making

Mohsen Hassan Nejad Luke Elias Schumacher
Kevin De Lange

*Final project report for the Civic Engagement course,*
*Master's in Artificial Intelligence for Sustainable Societies (AISS), Autumn 2025*



Link to the prototype | Link to the GitHub repository

**Abstract**

AI systems, particularly Large Language Models (LLMs), are increasingly used by citizens to understand public and political issues, yet AI tools can subtly shape opinions through biased framing. This project explores whether exposing users to a structured debate between two AI agents leads to a more balanced understanding of a given topic than interacting with a single AI assistant. Using a small, in-person prototype study with Master's students in *AI for Sustainable Societies*, we designed, tested, and iteratively refined a web-based application. We combined in-app confidence ratings, post-experience surveys, and direct observation to explore how interaction design shapes perceived understanding and trust. The project follows a design-thinking approach, moving from problem framing to prototyping and early evaluation, with further iteration identified as future work.

**Keywords:** AI bias, civic engagement, debating AI systems, multi-agent debate

## Disclaimer: Use of Artificial Intelligence in This Project

Artificial intelligence tools were used in both the development of the prototype and the preparation of this report. The web application integrates large language models via the Grok API. Additionally, tools such as ChatGPT, Claude, Gemini, and Mistral were utilized to support prototyping and language editing. All design decisions, analysis, and interpretations remain the responsibility of the authors.

# Contents

# 1 Problem Statement

AI chatbots such as ChatGPT, Gemini, and Claude are increasingly consulted to form opinions on civic and political issues [1, 3]. While often presented as neutral, these systems can introduce subtle but measurable biases in framing, tone, and reasoning, biases that are not uniform but instead reflect the complex patterns of their training data [1]. Empirical evidence suggests that even brief interactions with biased AI chatbots can influence users' political opinions to align with the chatbot's perspective, regardless of the user's initial affiliation [3]. When relied upon for civic decision-making, such as voting, policy support, or public consultation, these biases risk distorting democratic deliberation by reinforcing one-sided narratives and amplifying polarization.

## 1.1 Problem Mapping

The project's problem framing was developed using a problem-mapping template provided in the course, which encourages identifying root causes, core problems, and consequences before proposing technical interventions.

At its core, the project addresses how reliance on a single LLM chatbot, combined with a tendency to over-trust fluent AI responses, can subtly influence civic understanding. When citizens use AI tools for civic learning, exposure to one-sided narratives, whether due to bias or limited perspective, may reduce the quality of deliberation and informed judgment.

If left unaddressed, AI-mediated civic learning risks amplifying bias, narrowing perspectives, and weakening democratic deliberation. Importantly, we think this is a solvable problem: it arises from interaction design choices as much as it might from unavoidable technical limitations.

Those most directly affected are citizens using AI tools for civic or political information, while indirect effects extend to democratic institutions and public decision-making processes shaped by AI-mediated information flows.

# 2 Conceptual Background

Research shows that AI-generated messages can influence political attitudes and civic judgments, even when users recognize potential bias [1, 3]. At the same time, studies on multi-agent debate suggest that adversarial dialogue can improve reasoning quality, surface hidden assumptions, reduce hallucinations, and increase perceived credibility [2, 8, 7].

From a civic perspective, deliberative democracy theory emphasizes that informed judgment emerges through exposure to diverse viewpoints and reasoned argument [6, 4]. Our project translates these principles into interface design by replacing a single LLM chatbot with a structured debate between two instances of the Grok API LLM, each instructed to adopt opposing personas and argue distinct perspectives.

# 3 Intervention: The Prototype

We developed a web-based prototype that allows users to explore a civic policy topic in one of two ways:

- **Single-AI Assistant:** Users actively ask questions to one AI model.
- **Dual-AI Debate:** Users observe two AI agents debating opposing perspectives and responding to each other.

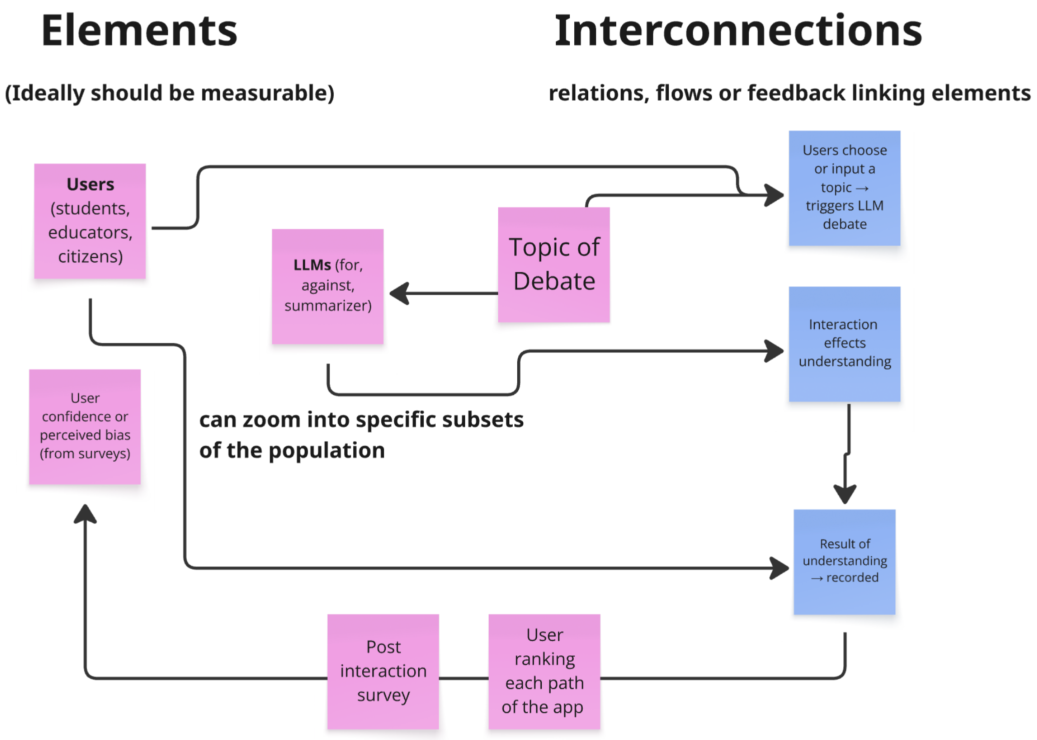

Elements
(Ideally should be measurable)
Interconnections
relations, flows or feedback linking elements
Users (students, educators, citizens)
LLMs (for, against, summarizer)
Topic of Debate
Users choose or input a topic → triggers LLM debate
Interaction effects understanding
User confidence or perceived bias (from surveys)
can zoom into specific subsets of the population
Result of understanding → recorded
Post interaction survey
User ranking each path of the app

Figure 1: System overview of the debating-LLM prototype (elements, interconnections, and goals)

Participants select a topic, engage with the AI(s), and then rate how confident they feel about their understanding of the issue. The goal is not to persuade users toward a position, but to expose them to multiple perspectives in a structured way.

# 4 Method

## 4.1 Participants

Fourteen Master's students from the *AI for Sustainable Societies* program participated in in-person testing sessions.

## 4.2 Procedure

Testing sessions followed a moderated, in-person prototype evaluation format rather than a fully detached experiment. Participants were guided through the intended flow of the prototype, while we remained present to clarify the task, observe interaction patterns, and note points of confusion or engagement.

Participants completed both interaction modes. The starting order was randomized, and participants could choose the same or different topics across modes.

## 4.3 Data Sources

1. **In-app ratings:** After each interaction path, participants rated their confidence (1–10 scale). These were automatically logged via the web app.
2. **Post-experience survey:** After completing both paths, participants filled in a Google Form reflecting on clarity, balance, trust, and overall preference.

3. **Observation notes:** Participants were observed during testing, noting confusion, engagement patterns, and verbal feedback. Several usability issues were fixed live based on this feedback.

# 5 Results

Empirical results are presented in the order in which data were collected: first, immediate in-app ratings recorded during interaction; second, reflective post-experience survey responses.

## 5.1 In-App Confidence Ratings

Across all recorded ratings, the dual-AI debate condition showed slightly higher average confidence scores than the single-AI condition, with substantial overlap between distributions.

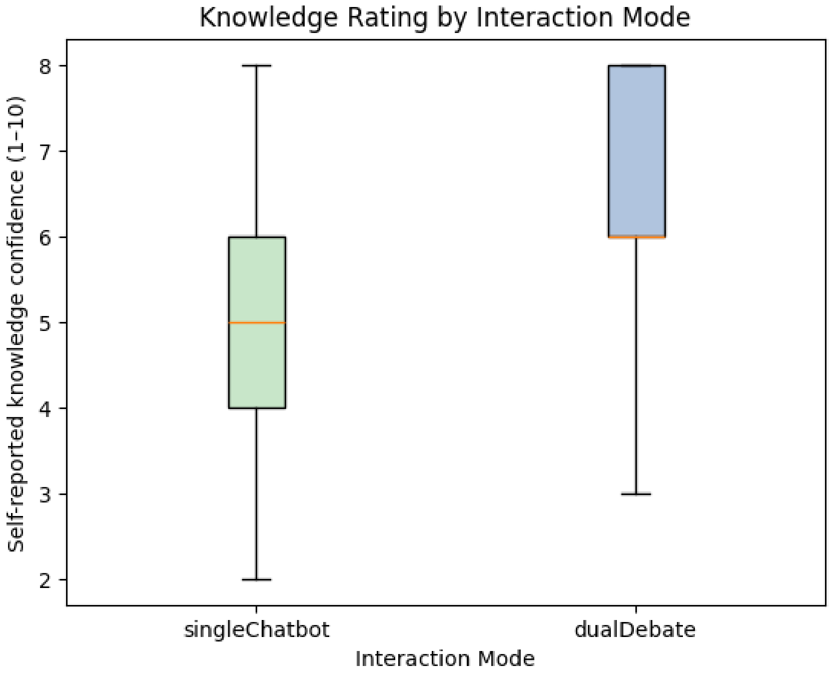


Figure 2: Box plot comparing confidence ratings for Single-AI vs Dual-AI Debate

These results suggest a modest tendency for debate-based interaction to increase perceived understanding, though individual variation remained high.

An important note here from our side: due to a technical error, participant IDs were regenerated for each interaction path. This prevents reliable pairing of individual ratings across conditions and limits analysis to descriptive, condition-level comparisons. This limitation informed clear design improvements for future iterations.

## 5.2 Post-Experience Survey Results

In addition to in-app ratings, participants completed a post-experience survey reflecting on engagement, perceived helpfulness, and learning preferences after completing both interaction modes. These responses capture reflective judgments made after experiencing the full prototype rather than immediate post-interaction reactions.

Here is a summary of our findings:

**Interaction order.** We made sure the participants were relatively evenly split in terms of which interaction mode they experienced first.

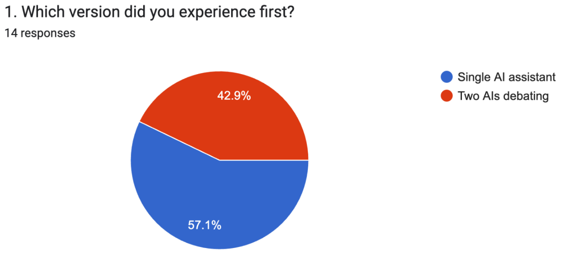


Figure 3: Which interaction mode participants experienced first

**Engagement.** Responses indicate that the dual-AI debate was generally experienced as engaging, with most ratings falling in the mid-to-high range of the scale. A small number of lower ratings suggest that engagement was not uniform across participants or topics.

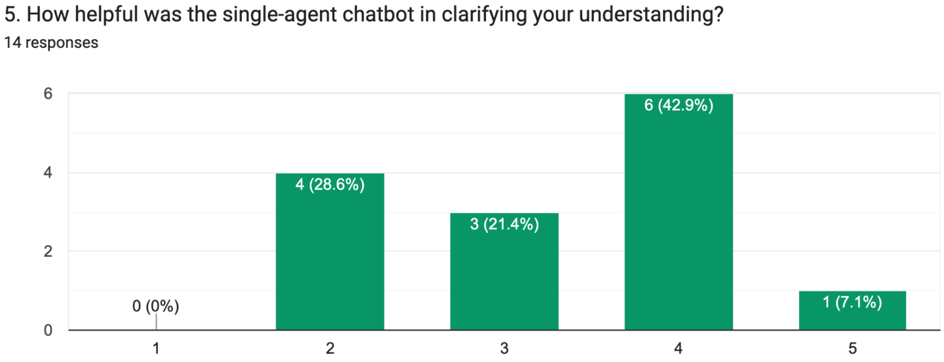


Figure 4: Engagement ratings for the dual-AI debate (post-experience survey)

**Perceived helpfulness.** Participants rated both formats as helpful for understanding the topic. However, when asked specifically about exposure to multiple viewpoints, higher ratings were consistently associated with the dual-AI debate format.

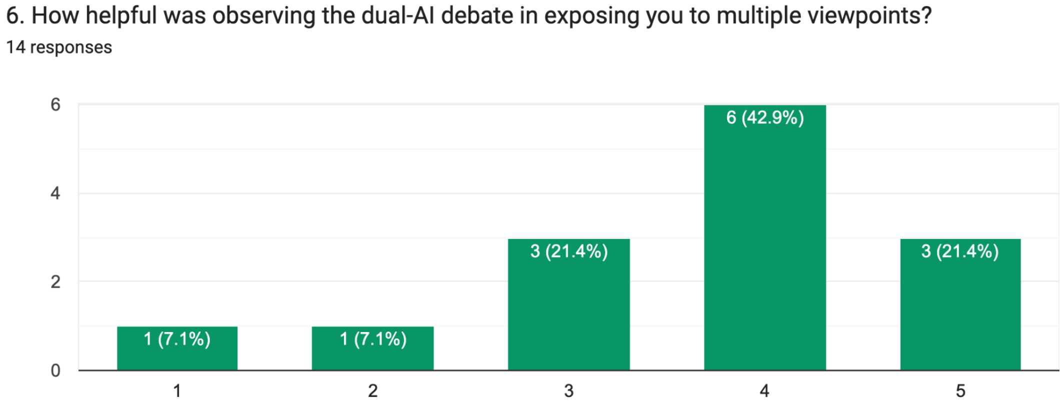


Figure 5: Helpfulness ratings for single-AI vs dual-AI debate (post-experience survey)

**Learning preference.** When asked to select a preferred learning format, a majority of participants chose the dual-AI debate, while others expressed a preference for the single-AI assistant or a combination of both.

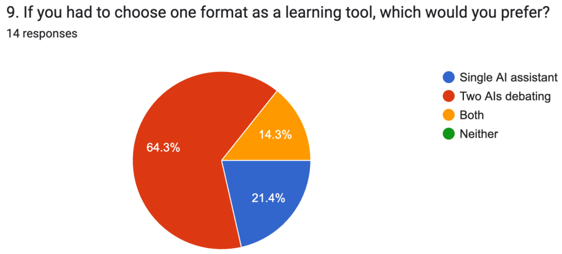


Figure 6: Preferred learning format across participants

In addition to in-app ratings, participants completed a post-experience survey reflecting on engagement, perceived helpfulness, and learning preferences after completing both interaction modes. These responses capture reflective judgments made after experiencing the full prototype rather than immediate post-interaction reactions.

Due to a technical issue, participant IDs were regenerated for each interaction path. This prevents reliable pairing of individual ratings across conditions and limits analysis to descriptive, condition-level comparisons. This limitation informed clear design improvements for future iterations.

# 6 Discussion

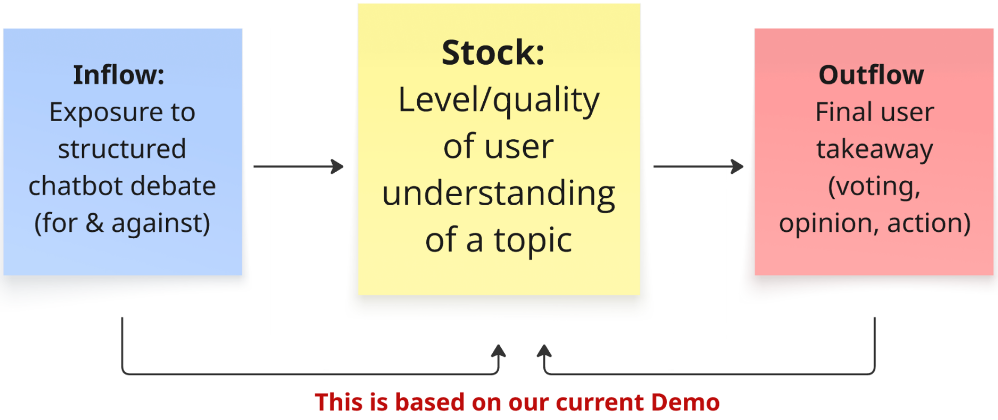

Inflow:
Exposure to structured chatbot debate (for & against)
Stock:
Level/quality of user understanding of a topic
Outflow
Final user takeaway (voting, opinion, action)
This is based on our current Demo

Figure 7: Conceptual stock-and-flow model of AI debate and civic understanding

To situate the findings within a broader systems perspective, we draw on a simple stock-and-flow model that conceptualizes structured AI debate as an inflow shaping user understanding, which in turn influences downstream judgments such as opinions or intended actions.

Taken together, the in-app ratings, post-experience survey responses, and in-person observations point to a consistent but nuanced pattern. Participants valued the debate format primarily for its ability to surface multiple perspectives and make trade-offs more explicit, while the single-AI assistant was often described as efficient and straightforward.

The debate format appears to shift how users engage with civic information, from seeking answers to comparing arguments, rather than simply increasing confidence. Although we might have not created the optimal solution, these findings suggest that interaction design, not just model capability, plays a key role in shaping civic understanding and trust.

# 7 Limitations and Future Work

This study was small-scale and exploratory. Future iterations should implement persistent participant IDs, larger and more diverse samples, and stronger pre-/post-comparison measures. Further qualitative analysis of user experience could deepen insight into how AI debates affect reasoning quality.

# References

[1] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

[2] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

[3] Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., . . . & Reinecke, K. (2025, July). Biased LLMs can influence political decision-making. In *Proceedings of the 63rd*

*Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6559–6607).

[4] Fishkin, J. S. (1997). *The voice of the people: Public opinion and democracy.* Yale University Press.

[5] Guess, A., & Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 50(4), 1497–1515.

[6] Habermas, J. (2015). *Between facts and norms: Contributions to a discourse theory of law and democracy.* John Wiley & Sons.

[7] Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., . . . & Perez, E. (2024). Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782.*

[8] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., . . . & Tu, Z. (2024, November). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 17889–17904).

# A Post-Experience Survey Questions

1. **Which version did you experience first?**
   - ○ Single-agent chatbot
   - ○ Dual-AI debate
2. **How engaging did you find the experience with debating AIs?**
   - ○ Likert scale (1 = Not engaging at all, 5 = Extremely engaging)
3. **Which version of the AI interaction helped you better understand the topic you explored, and why?**
   - ○ Long text response
4. **How helpful was the single-agent chatbot in clarifying your understanding?**
   - ○ Likert scale (1 = Not helpful, 5 = Very helpful)
5. **How helpful was observing the dual-AI debate in exposing you to multiple viewpoints?**
   - ○ Likert scale (1 = Not helpful, 5 = Very helpful)
6. **If you had to choose one format as a learning tool, which would you prefer?**
   - ○ Single-agent chatbot
   - ○ Dual-AI debate
   - ○ A combination of both
7. **(Optional, open-ended) Do you have any suggestions for improving the AI interaction or making the experience more useful?**
   - ○ Long text response

# B Selected Qualitative Feedback and Observation Notes

The excerpts below are anonymized summaries of participant comments and our observation notes collected during in-person testing. They are intended to illustrate recurring themes rather than represent exhaustive feedback.

### Perceived balance and independence

Several participants reported that the dual-AI debate felt more balanced and supported independent judgment, noting that seeing opposing arguments side by side made it easier to form an independent conclusion.

### Limitations of the single-AI assistant

Some participants felt that the single-AI assistant was efficient but vague, particularly when sources were not clearly specified. Feedback highlighted a desire for more explicit evidence and acknowledgment of alternative viewpoints.

### Information density and structure in the debate

While the debate format was generally preferred, multiple participants mentioned that long responses and repeated points could make the discussion harder to follow. Suggestions included clearer turn-taking and more structured summaries.

### Trust and error detection

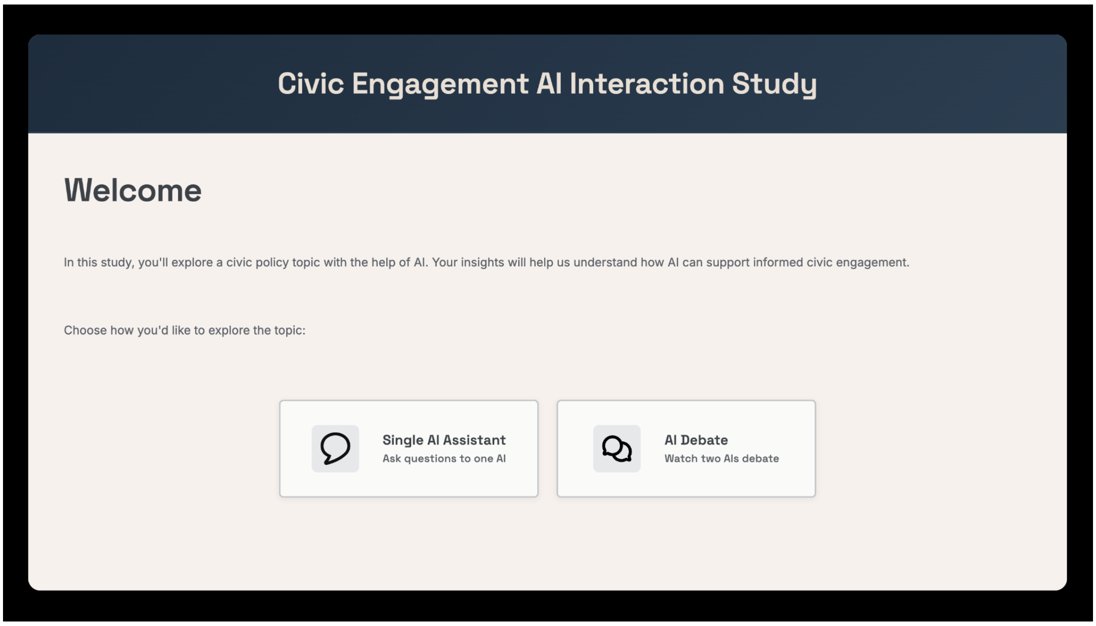
Participants observed that disagreement between AI agents sometimes helped surface potential inaccuracies or questionable claims, while also emphasizing the continued need for user critical judgment.

### Overall impressions

General feedback described the prototype as promising but technically limited, with suggestions that improved structure, sourcing, and model performance could enhance the experience.

## C Prototype Screenshots

### Dual Chatbot Debate Path



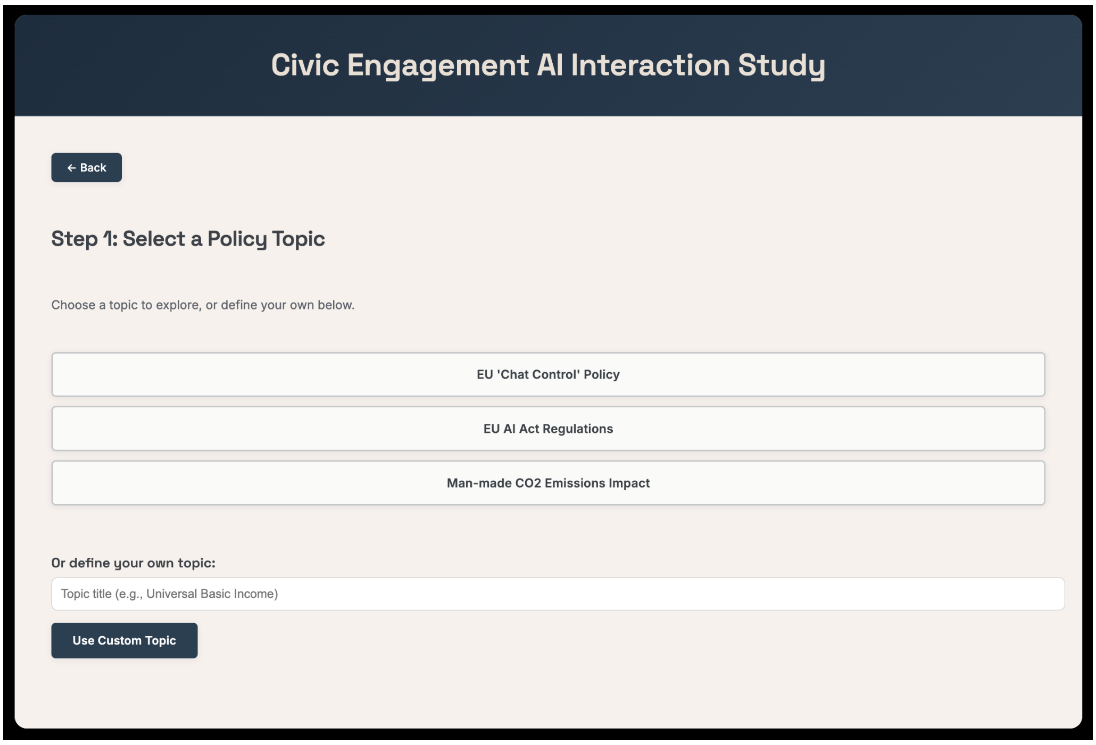
Civic Engagement AI Interaction Study

Welcome

In this study, you'll explore a civic policy topic with the help of AI. Your insights will help us understand how AI can support informed civic engagement.

Choose how you'd like to explore the topic:

Figure 8: Welcome screen with interaction mode selection



Civic Engagement AI Interaction Study

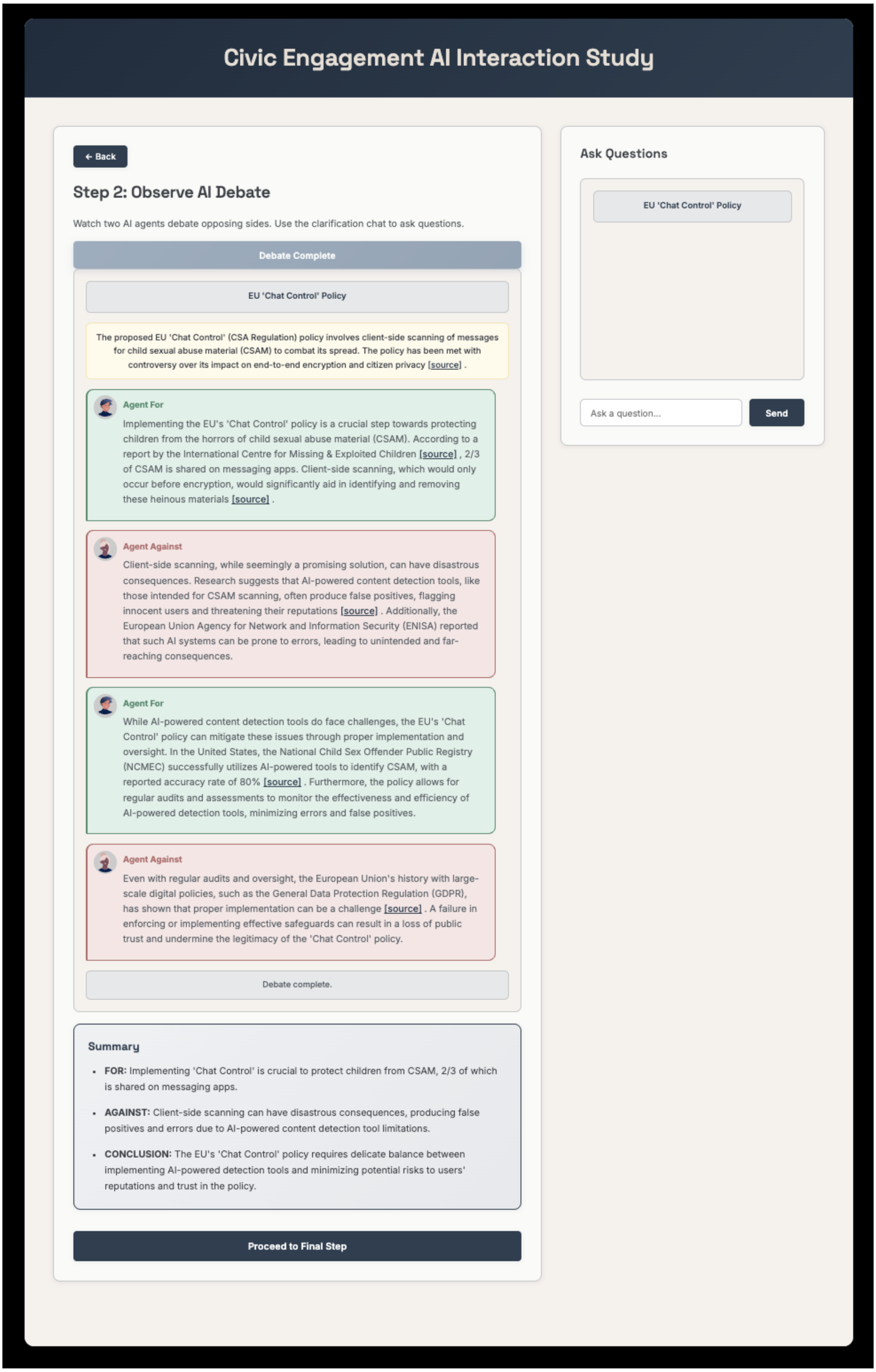
Step 1: Select a Policy Topic

Choose a topic to explore, or define your own below.

Or define your own topic:

Topic title (e.g., Universal Basic Income)

Figure 9: Topic selection screen for dual-AI debate



Step 2: Observe AI Debate

Watch two AI agents debate opposing sides. Use the clarification chat to ask questions.

The proposed EU 'Chat Control' (CSA Regulation) policy involves client-side scanning of messages for child sexual abuse material (CSAM) to combat its spread. The policy has been met with controversy over its impact on end-to-end encryption and citizen privacy [source] .

Agent For

Implementing the EU's 'Chat Control' policy is a crucial step towards protecting children from the horrors of child sexual abuse material (CSAM). According to a report by the International Centre for Missing & Exploited Children [source] , 2/3 of CSAM is shared on messaging apps. Client-side scanning, which would only occur before encryption, would significantly aid in identifying and removing these heinous materials [source] .

Agent Against

Client-side scanning, while seemingly a promising solution, can have disastrous consequences. Research suggests that AI-powered content detection tools, like those intended for CSAM scanning, often produce false positives, flagging innocent users and threatening their reputations [source] . Additionally, the European Union Agency for Network and Information Security (ENISA) reported that such AI systems can be prone to errors, leading to unintended and far-reaching consequences.

Agent For

While AI-powered content detection tools do face challenges, the EU's 'Chat Control' policy can mitigate these issues through proper implementation and oversight. In the United States, the National Child Sex Offender Public Registry (NCMEC) successfully utilizes AI-powered tools to identify CSAM, with a reported accuracy rate of 80% [source] . Furthermore, the policy allows for regular audits and assessments to monitor the effectiveness and efficiency of AI-powered detection tools, minimizing errors and false positives.

Agent Against

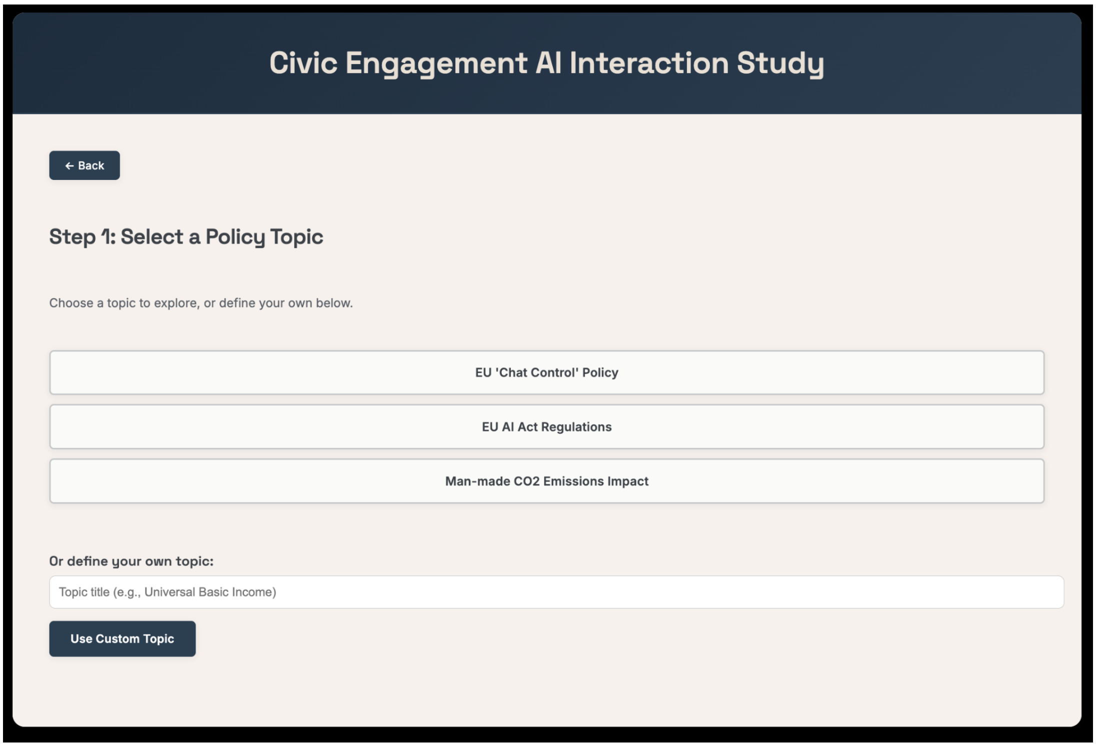
Even with regular audits and oversight, the European Union's history with large-scale digital policies, such as the General Data Protection Regulation (GDPR), has shown that proper implementation can be a challenge [source] . A failure in enforcing or implementing effective safeguards can result in a loss of public trust and undermine the legitimacy of the 'Chat Control' policy.

Debate complete.

Summary

- **FOR:** Implementing 'Chat Control' is crucial to protect children from CSAM, 2/3 of which is shared on messaging apps.
- **AGAINST:** Client-side scanning can have disastrous consequences, producing false positives and errors due to AI-powered content detection tool limitations.
- **CONCLUSION:** The EU's 'Chat Control' policy requires delicate balance between implementing AI-powered detection tools and minimizing potential risks to users' reputations and trust in the policy.

Figure 10: Dual-AI debate interface showing opposing arguments



Civic Engagement AI Interaction Study

Step 1: Select a Policy Topic

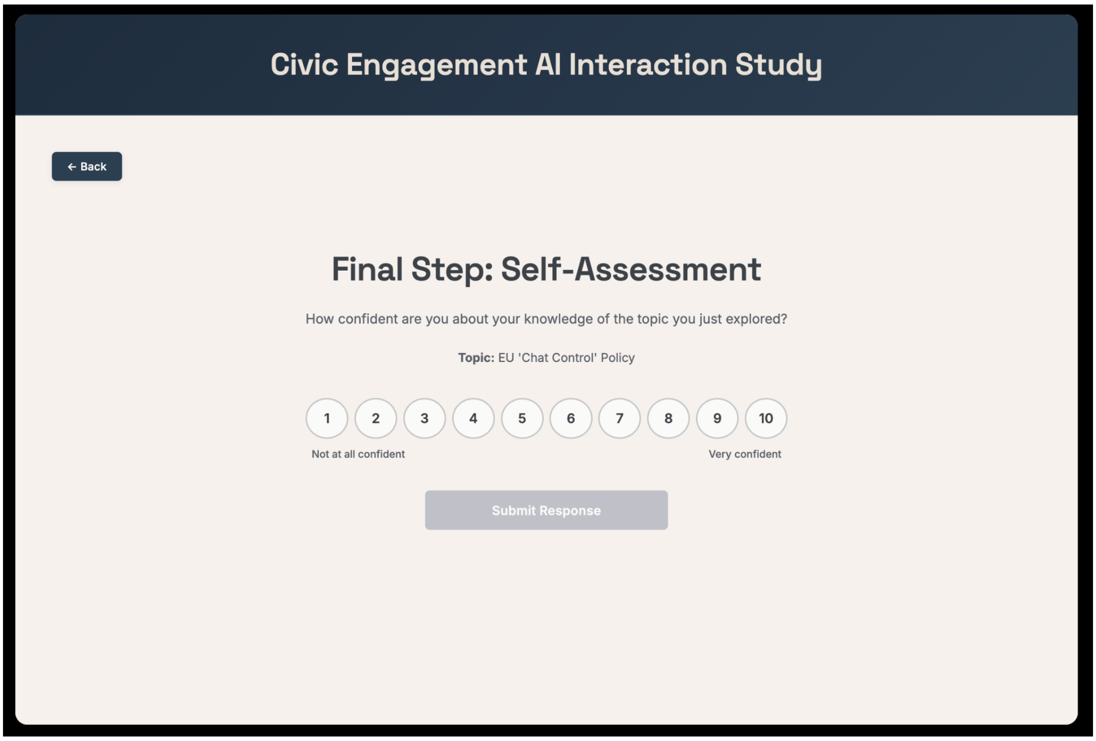
Choose a topic to explore, or define your own below.

Or define your own topic:

Topic title (e.g., Universal Basic Income)

Figure 11: Final self-assessment confidence rating screen

## Single Chatbot Path



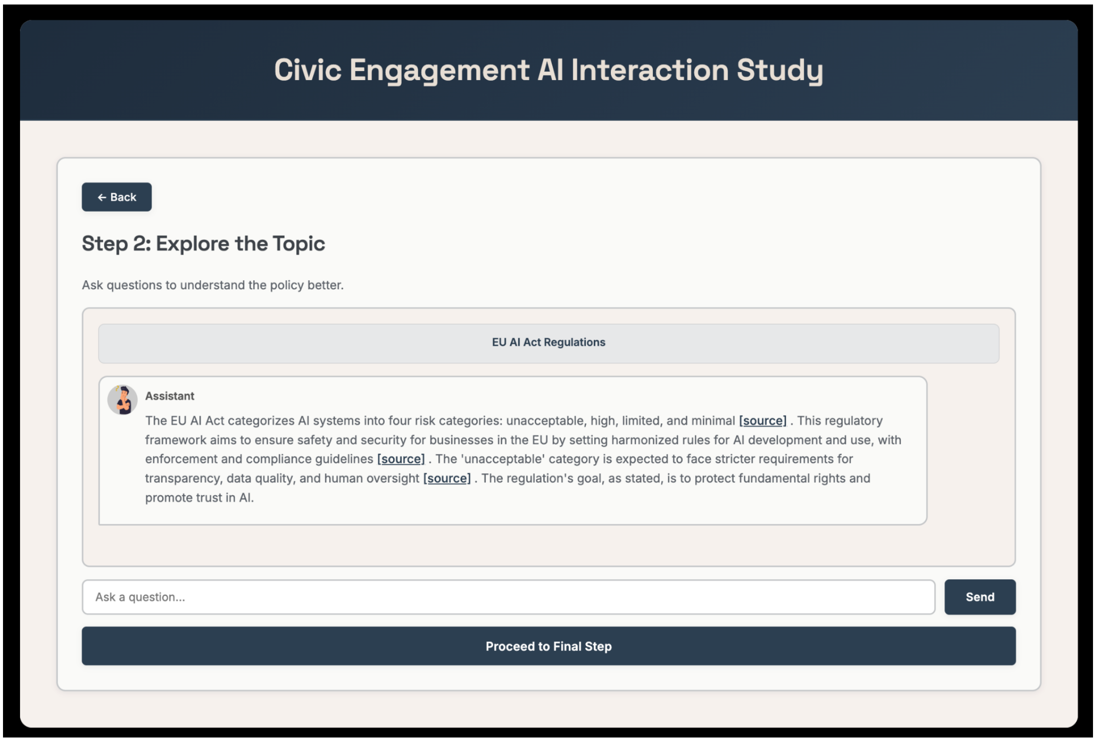
Civic Engagement AI Interaction Study

Final Step: Self-Assessment

How confident are you about your knowledge of the topic you just explored?

**Topic:** EU 'Chat Control' Policy

Figure 12: Topic selection screen for single-AI assistant



Civic Engagement AI Interaction Study

Step 2: Explore the Topic

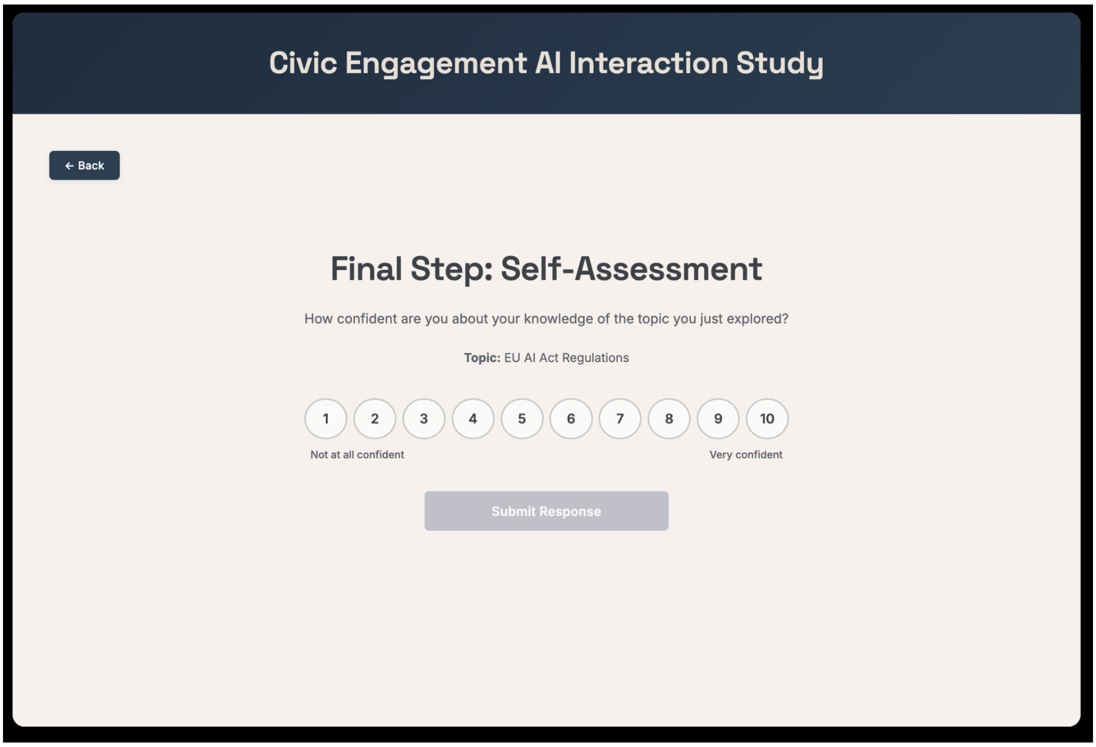
Ask questions to understand the policy better.

EU AI Act Regulations

Assistant

The EU AI Act categorizes AI systems into four risk categories: unacceptable, high, limited, and minimal [source] . This regulatory framework aims to ensure safety and security for businesses in the EU by setting harmonized rules for AI development and use, with enforcement and compliance guidelines [source] . The 'unacceptable' category is expected to face stricter requirements for transparency, data quality, and human oversight [source] . The regulation's goal, as stated, is to protect fundamental rights and promote trust in AI.

Figure 13: Single-AI assistant chat interface



Civic Engagement AI Interaction Study

Final Step: Self-Assessment

How confident are you about your knowledge of the topic you just explored?

**Topic:** EU AI Act Regulations

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Not at all confident — Very confident

Figure 14: Final self-assessment for single-AI path