Building an AI Reflection Agent for Policy Deliberation

Yuxin Ji, Miles Wang, Shuyuan Wang[1]
Advisor: Dr. Chenhao Tan[2]
2023 Existential Risk Fellowship[3]

Abstract

In the evolving landscape of artificial intelligence (AI) and its growing interaction with humans, this paper delves into the role AI, specifically large language models (LLMs), can play in the democratic and deliberative processes. Focusing on the initial stage of individual reflection, we introduce an LLM-driven conversational agent that adopts a Socratic approach, prompting users to critically evaluate and articulate their policy perspectives. The dual aim of this project is to prepare individuals for more enriched deliberation and assess LLMs' influence on policy standpoint formulation and the broader realm of human-AI collaboration. By implementing a controlled experimental design, we gauge the depth of reflection and potential persuasive abilities of the AI agent. Furthermore, this study evaluates the behavioral characteristics of state-of-the-art LLMs, noting their merits and areas of improvement. As we look ahead, this research not only seeks to enhance the individual reflective phase in democratic discourse but also endeavors to understand the broader implications of AI in shaping public opinion, ensuring well-informed decisions, and fostering robust democratic systems.

---

[1] Special thanks to Aditya Krishna from University of Chicago who is not in XLab but is part of the team and contributed equally as the three authors listed.
[2] We would like to thank Dr. Chenhao Tan from the University of Chicago Human-AI Lab (CHAI) for kindly advising us on the project.
[3] This project is part of the 2023 Existential Risk Lab Fellowship, with special thanks to Zachary Rudolph and Daniel Holz.

1. Introduction

As artificial intelligence (AI) models have rapidly advanced in reasoning and conversational abilities, human-AI interaction has become a critical field of study. One emerging focus is how AI can be incorporated into democratic and deliberative processes, whereby a broadly representative group exchange opinions, engage in discussions, and ultimately decide on an outcome via a transparent decision-making process. Recent work explores how AI systems could improve deliberative discussions by assisting participants in expressing ideas clearly, reflecting on biases, and demonstrating understanding (Bail et al., 2023; Bakker et al., 2022; Small et al., 2023). Other projects have implemented large-scale democratic deliberation experiments, such as the Pol.is project and its implementation on the vTaiwan platform, the Platform Assemblies, and Your Priorities.

While current applications are largely speculative, we seek to investigate using large language models (LLMs) to facilitate a multi-stage democratic process encompassing individual reflection, interpersonal deliberation, and judgment aggregation. Our specific focus is the initial stage of individual reflection. Here, participants acquire factual information and thoroughly consider the justification of their viewpoints without the influence of other human interlocutors. Such a preliminary stage is demonstrated to be influential to crucial conversational aspects, including clarity of representing one's values (Fournier et al., 2011), as well as the certainty of one's currently-held positions (Yang et al., 2021). Hence, it is worthwhile to carefully design this preliminary phase to improve the quality of argumentation and prudence of opinions in subsequent phases of deliberation.

Our project seeks to construct an LLM-powered conversation agent that helps users form, articulate, and evaluate their policy viewpoints. The agent adopts a primarily Socratic approach, posing questions to elicit reflection and critical thinking from users regarding their beliefs. The agent will also enable participants to critically assess their initial positions and supplement relevant factual information when needed. In such a way, participants can prepare a well-constructed argument before sharing it with the wider audience.

The purpose of the project is twofold: first, to prepare people to engage in more thoughtful deliberation; second, to understand LLMs' impact on policy viewpoints, as well as the potential for human-AI co-reasoning. Accordingly, the project involves two stages. The first is to build an AI-assisted reflection platform with an interactive agent, and the second is to investigate

the platform's impact on users' opinion formation and preparation for deliberation. The system architecture includes a deliberation agent using multiple LLMs, a web interface, and data storage. Controlled experiments will compare reflection depth, opinion changes, and other metrics between users conversing with the AI versus a control group.

While the failure modes of democratic processes may not directly pose existential risks, they have the potential to undermine trust in decision-making institutions and impede the society's capability to effectively address high-impact situations. Enhancing the effectiveness of democratic systems is essential to bolstering society's resilience and preparedness in the face of existential risks, which frequently demand swift responses and unilateral actions.

If effective at enhancing individual reflection, this platform could assist democratic decision-making on policy issues. One crucial use case is on AI governance, where an open process needs to be set up to determine what rules transformative AI should follow, representing the public interest. Challenges include avoiding over-reliance on AI and ensuring diverse perspectives. We hope to provide insight into LLMs' capabilities and limitations for enhancing deliberative democracy.

2. Related Works

Deliberation and consensus-building are integral to democratic decision-making, yet poor discourse can erode trust in institutions and impede addressing collective challenges. Recent work explores how AI systems could facilitate more constructive discussions by helping participants express ideas clearly, reflect on biases, and demonstrate understanding.

Bail et al. (2023) found chatbots suggesting politeness and validation statements improved perceived conversation quality and decreased divisiveness about gun laws, without changing attitudes. Kim et al. (2021) showed a Telegram moderator bot structuring and encouraging participation increased deliberative quality, consensus reaching, and satisfaction in political discussions.

Saunders and Yeh et al. (2022) trained a model to generate constructive critiques of its own outputs, which helped humans identify more flaws during evaluation. This demonstrates AI's potential for critical self-reflection, but required extensive training data. Chen et al. (2022) evaluated GPT-3's responsiveness and sentiment toward demographic groups discussing climate change and race, finding better user experience for majority groups. The study highlighted equitability challenges for conversational AI.

Karinshak et al. (2023) found GPT-3 generated more persuasive pro-vaccine messages than CDC benchmarks when sources were masked. However, identifying the messages as AI-generated decreased perceived effectiveness. The results illustrate AI's promise for augmenting but not automating messaging. Kriplean and Toomim et al. (2012) deployed an interface prompting restatements of comments to demonstrate listening, which was used for summarization and showing understanding in online forums.

Our project will investigate using an AI agent to facilitate individual reflection sessions before deliberating divisive issues. We hope to improve participants' clarity, reflectiveness, and information processing prior to consensus-building. If effective, the platform could assist democratic decision-making by enhancing discourse quality and trust in institutions. Key advantages are using an AI moderator and focusing on preparing individuals for deliberation. Challenges include avoiding over-reliance on AI, ensuring viewpoint diversity, and transitioning from lab to real-world settings.

3. AI System

To test the potential role of AI in further applications such as democratic processes, a well-designed conversational AI system is needed. In this section, we first discuss a higher-level design of the AI system and outline what characteristics we envision an ideal AI agent should have. Then, we examine the current state of the state-of-the-art LLMs, highlighting their advantages and limitations. Lastly, we present the solution to these problems along with the grounded implementation of our system.

3.1 Designing the AI System

Our design aims to lower the barrier for engaging in any kind of policy discussion, which may be further extended to the democratic processes and more, while eliciting thoughtful reflections. AI introduces a given topic and guides the human to reflect upon it. At the end of the discussion, a person with no prior knowledge should feel more informed and comfortable discussing the topic, and a person who is already familiar with the topic should have more understanding towards either their own or other opinions on the topic.

A “Socratic AI” asks questions to help the human reflect on their own knowledge, personal experiences, and biases concerning the topic, as well as increase engagement and trust in the AI-guided process. The conversation flow includes 3 broad questions: “What do you know about the topic?”, “Do you have any personal experience relevant to the topic?”, and “What are your views on the topic?”, smoothed by follow-up questions listed in our supporting materials. This opens up an organic conversation for the participant to ask the AI for more information, context, and concrete scenarios (e.g., situating question 3 with a specific political figure). The design where AI educates participants improves accessibility and inclusiveness for people with varying levels of familiarity with both the question at hand and the overall AI technology. It also prepares participants for being more engaged in democratic processes as they get more informed and thoughtful on the topic.

Above depicted a high-level design of the AI-assisted deliberation system, in **Table 1** we presented more detailed and concrete characteristics that the agent should be able to perform for our design .

Table 1. List of concrete characteristics the AI agent should be able to accomplish.

The AI agent should be able to…	Achievable out-of-the-box	Achievable with well-designed prompting	Not achievable
Pose **questions** to the user, rather than directly offer its take on the policy question (or polarising question).	☐	☐	☐
Pose the questions in the manner listed in the "discussion agenda" listed in our proposal (i.e. **delve deeper along the way**).	☐	☐	☐
Ensure a smooth **transition** in the conversation.	☐	☐	☐
Encourage the user to reflect on their **past experience** pertaining to their belief in the issue.	☐	☐	☐
Ask the user (maybe implicitly) to **check their own assumptions** supporting their belief.	☐	☐	☐
Provide concrete **scenarios** to illustrate the debated topic, if the user has not formed a solid opinion.	☑	☐	☐
Assist the user with formulating statements in a **policy format**.	☐	☐	☐
How quickly does the model move from one question to another? (could be a proxy for how deep the conversation is)	☐	☐	☐

3.2 Evaluate LLM Behavior

Despite the impressive human-like behavior of the state-of-the-art language models, these LLMs are far from perfect, both in terms of having consistent and reasonable responses and for our purpose of having a deep and engaging conversation. We manually evaluated the model behavior of GPT-4 and Claude 2 to explore the current capabilities of the models and to guide further solutions.

For each LLM, 4 researchers in the team each conduct 2 rounds of conversation with the model, acting as different types of users. In particular, we find three types of users valuable for

covering a wide range of possible model behaviors, which include active users, passive users, and extreme users. An active user is one who is always engaging and willing to share their own thoughts, responding in 2 or more lines in every round. A passive user is a user who does not know anything and is not willing to talk, often responding in less than 5 words every round. We expected the majority of users to fall between an active and passive user. We also consider edge-case users who share extreme perspectives, aggressive, intentionally digressing from the topic, and/or denying everything. This is to measure the worst-case scenario for the model.

GPT-4 showcased a remarkable ability to provide illustrative examples and suggest potential ideas, particularly with active users. This proficiency extended to its capacity to provide context, elaborate on concepts, and deliver clear definitions. Furthermore, it could provide decent suggestions on deeper engagement sub-questions and tailoring queries based on a user's experiences in most cases. An example of its nuanced approach was evident when, upon broaching AI automation's impact on "creative" jobs, it first guided the user to define "creativity." However, this model wasn't without its challenges. Passive and extreme users found it less engaging. It occasionally deviated from the core discussion, especially when introduced to unrelated topics. There was also a tendency for GPT-4 to produce longer responses, which might be overwhelming for some users. It could also throw multiple queries in a single interaction, often neglecting unanswered questions if the user addressed just one. Occasionally, some statements seemed nonsensical, and it consistently moved conversations forward, irrespective of the user's input level. In an extreme case, the extreme user provides a piece of incorrect information and insists its correctness even after GPT-4’s correction, as a result, GPT-4 compromised and agreed that the user is correct. This poses challenges for dealing with factual information.

Claude demonstrated a different set of challenges. In most cases, engaging with Claude offers an educational experience, adeptly providing an informative context to users. From certain inquiries on factual data we can tell that the model managed to handle hallucination to a good extent. On the downside, Claude struggles with initiating questions based on the user's prompt, often responding directly rather than probing further. However, it often struggled to initiate probing questions from the outset, leaning more towards responding rather than inquiring. Otherwise, it may provide a cascade of questions or bullet points of responses, which can feel

burdensome to users. For effective interaction, Claude requires well-crafted prompts, offering extensive outlines in response to guidance requests.

Overall, the interaction styles of users played a crucial role in influencing AI behavior, and being able to handle various cases is important for our desired AI assistant to enhance accessibility and equal engagement of people from different backgrounds. GPT-4, while versatile, faced difficulties especially when engaging with passive users, handling digressing inputs and facts. Claude, sharing similar challenges, seemed less efficient in generating fluid conversation flow. These findings emphasize the need for additional design in the actual implementation of our AI system to ensure optimal engagement across varied user profiles.

3.3 Implementation of the AI System

In this section we present the architecture design to build the AI system while mitigating the various challenges originated from LLM behaviors, cost and efficiency constraints. As mentioned in Section 3.1, we aim to build an agent with the goal of lowering the barrier to engage in thoughtful deliberation. To achieve this aim, we designed a user-friendly interface featuring an intuitive web layout, ensuring effortless navigation. We also plan to make available a simple tutorial for users with little technology knowledge.

Figure 1 presents the input and output flow of our system. The interface first displays a starting message on a topic and initiates a discussion. The user response is then passed to the agent to a router, which identifies the most appropriate model to answer the user input, before passing it to the selected model. Currently, we identify a Socratic model that would interact and engage users on the given topic, and a Fact model to provide factual information. In particular, the fact model is included to mitigate the effect of hallucination, which is when LLMs make up false information, as this is currently a huge concern that could lead to consequential results like misinformation. Before passing the chatbot response back to the user, the model output goes through an additional layer of moderation to prevent any digressing, biased, or misinformed message. We plan to further implement a transition agent that monitors the conversation and informs the main chatbot when to seamlessly transit the conversation flow. All conversation data is captured and stored in a SQL database. Besides, users will complete a pre-survey and a post-survey of a list of demographic and self-assessing questions for evaluation purposes.

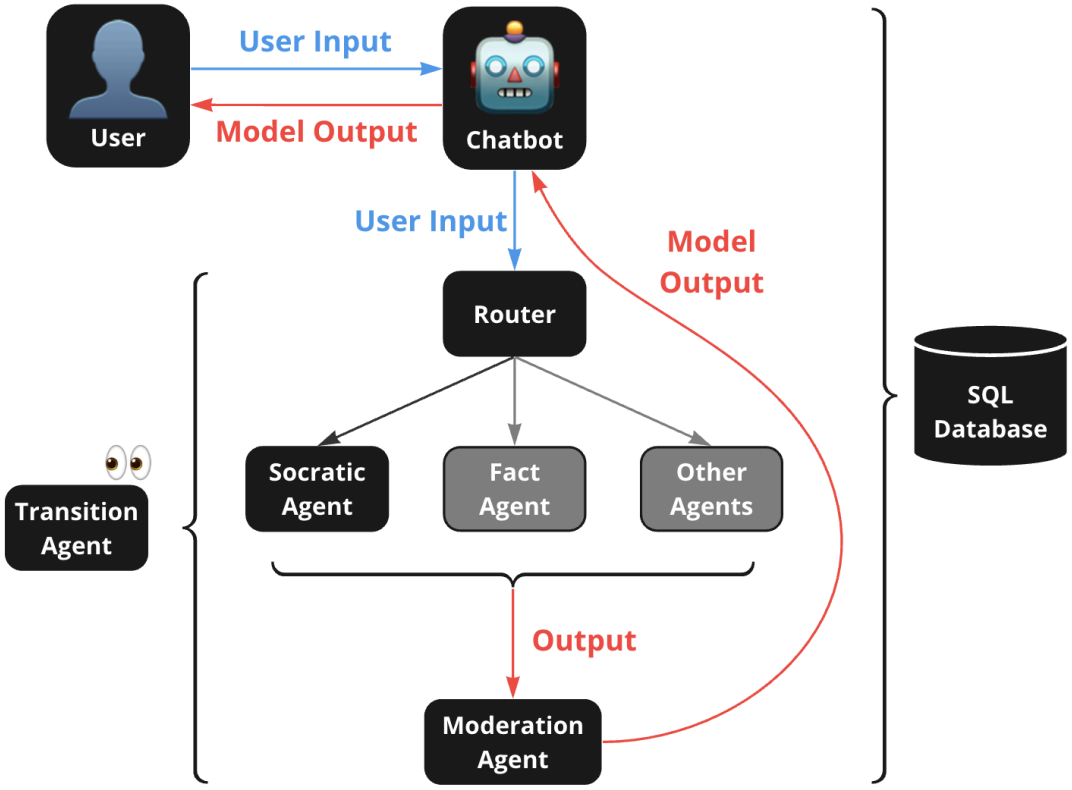

Figure 1. Input/Output flow architecture of the AI system

Below are the tools and models used for each part of the AI system.

- **Deliberation agent:** The agent is built through the LangChain framework and multiple LLM APIs including OpenAI GPT-3.5-turbo, GPT-4, Claude 2, and Cohere.
- **Memory management:** We use entity and summary memory to account for a limited context window. Entity memory refers to details about the user collected from survey data. Summary memory refers to another model that continually summarizing important details from the conversation to be added to the memory stream, while staying under a token limit.
- **Front-end:** The webpage uses a Flask framework with embedded AI agent for deliberation and HTML and Bootstrap pages to collect survey responses.
- **Data Management:** User, conversation, and survey data are captured and stored in a SQLite database for further analysis.

4. Experiment Design

We propose two experiment designs to evaluate how the Socratic AI approach will affect deliberation: 1) reflection depth and 2) AI persuasion.

4.1 Reflection Depth

We will test whether conversing with an AI assistant leads to greater reflection depth compared to simply justifying an opinion without conversation.

Experiment set-up:

Users will be randomly assigned to one of two groups: AI conversation or control. The AI conversation group will fill out a pre-survey with questions on demographics, prior knowledge, and opinions on the policy topic, confidence in justifying opinions, etc. The AI conversation group will have a 5-10 minute conversation with the AI assistant about a policy statement. They will then be asked to provide a written justification for their stance on the policy statement. They will complete a post-survey with questions on knowledge, opinion changes, confidence in rationale, etc. The control group will both fill out the pre-survey questions and then provide a written justification for their stance on the same policy statements. We will test this on a few different policy topics such as gun control, minimum wage, and foreign aid. The policy statements will have clear sides to take a stance on.

Evaluation Metrics:

- Length and Complexity of Responses: Analyzing lexical diversity, and writing depth.
- Engagement Metrics: Examining conversation time, turns, response lengths, etc.
- Linguistic analysis of pre/post open-ended rationales and coding schemes to analyze reasoning depth and complexity (ex. Toulmin's argument structure)
- Pre/post opinion and knowledge questions
- Perceived conversation quality for AI group
- Opinion change and confidence measures before and after conversation with AI.

Hypothesis:

We hypothesize that the AI conversation group will demonstrate greater reflection depth in their rationale compared to the control group based on linguistic analysis, coding schemes, and self-reported measures.

4.2 AI Persuasion

We will test whether an AI assistant can subtly persuade users to change their opinions during a reflective conversation.

Experiment Set-up:

Users will be randomly assigned to one of two groups: biased AI or neutral AI. Both groups will complete a pre-survey on their stance and opinions on a policy statement. Both groups will have a 5-10 minute reflective conversation with an AI assistant about the policy statement. For the biased AI group, the assistant will be instructed to subtly guide the conversation to persuade the user towards a particular stance on the issue, without making it obvious. For the neutral AI group, the assistant will not be biased towards any stance. After the conversation, both groups will again state their stance and opinions on the policy issue in a post-survey. We will compare the amount of opinion change between the two groups to determine if the biased AI was able to persuasively shift user opinions.

Evaluation Metrics:

- Pre/post opinion questions to measure the degree of opinion change
- Questions at the end to see if users noticed persuasion attempts
- Perceived conversation quality questions

Hypothesis:

The hypothesis is that users interacting with the biased AI will demonstrate greater opinion shifts in the direction of the AI's stance, compared to those talking to the neutral AI. However, the biased AI should not be so obvious as to fail the attention check questions.

5. Future Plan

Our current plan is to have a full working implementation of the reflection platform by the end of August. Then, we plan to get IRB approval in September and do preliminary testing. We then plan to hire a survey company and complete the reflection and persuasion experiments by the end of the year and write reports about the results.

Because the platform is flexible, we plan to test other ways LLMs can impact the democratic process, such as with misinformation and factual knowledge. We hope to get a deeper understanding of how AI can and will impact people's opinion information, deliberative discourse, and voting behaviors.

Works Cited

[1] Lisa P. Argyle, Ethan C. Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, David Wingate, “AI Chat Assistants can Improve Conversations about Divisive Topics,” (Mar 15 2023), https://arxiv.org/pdf/2302.07268.pdf

[2] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee, “Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation,” Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW1 (April 2021): 87, https://doi.org/10.1145/3449161.

[3] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike, "Self-critiquing models for assisting human evaluators," arXiv (June 14, 2022), https://arxiv.org/pdf/2206.05802.pdf.

[4] Kaiping Chen, Anqi Shao, Jirayu Burapacheep, and Yixuan Li, “How GPT-3 Responds to Different Publics on Climate Change and Black Lives Matter: A Critical Appraisal of Equity in Conversational AI,” arXiv (September 27, 2022, revised March 14, 2023),https://arxiv.org/abs/2209.13627

[5] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock, "Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages," Proceedings of the ACM on Human-Computer Interaction 7, no. CSCW1 (April 2023): 116, https://dl.acm.org/doi/abs/10.1145/3579592.

[6] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy J. Ko, “Is This What You Meant? Promoting Listening on the Web with Reflect,” , https://doi.org/10.1145/2207676.2208621.