Generative Social Choice

Suhas Vundavilli
BTech 3rd Year
Indian Institute of Science
suhasv@iisc.ac.in

Abstract

The paper on **Generative Social Choice** explores the integration of generative AI with social choice theory to help enhance democratic decision-making. The concept of Generative Social Choice is introduced, which is a framework that uses LLMs to generate alternate scenarios dynamically and extrapolate popular preferences from free-text responses. The two main components of the framework are as follows: theoretical guarantees using idealized preference queries and empirical validation of LLM-based implementations. The application demonstrated is of statement selection, where free-text survey responses are summarized into a representative set of statements using a multi-winner voting rule called **Balanced Justified Representation** (BJR). The study highlights the potential of LLMs to facilitate more inclusive and representative decision-making while addressing challenges such as biases, computational scalability, and validation of generated outcomes.

1 Introduction

Social choice theory is the field of mathematics, economics, and political science that studies the aggregation of individual preferences towards collective decisions. The setting usually consists of a set of predetermined set of alternatives, to which participants are asked to express their preferences, often in the form of rankings. This model, however, fails when the when input is sought on nuanced decisions, such as those of climate change or constitutional reform. The rapid rise of generative AI, and LLMs in particular, has led to several organisations, the likes of which include Meta and OpenAI, experimenting with their use in democratic processes for AI value alignment.

These developments gave rise to the concept of *generative social choice*, combining the rigor of existing social choice theory with the power and flexibility of LLMs to facilitate the decision making process on complex issues in a principled manner.

LLMs help circumvent the two main obstacles of using classical social choice to answer open-ended questions:

- **Unforseen alternatives:** In classical social choice, the set of presented alternatives are explicitly specified and static. In contrast, LLMs are able to generate alternatives which were not initially anticipated, but lie in a common ground of all relevant outcomes for the problem at hand.
- **Extrapolating preferences:** In classical choice theory, agents specify their preferences in a rigid format, and evaluate each alternative independently or use a voting rule to aggregate preferences. This fails to capture the nuances of human preferences on unforseen alternatives. LLMs address this problem by extrapolating the preferences of the participants, usually by acting as a proxy and predicting their preferences over alternatives, whether forseen or newly generated.

The framework of generative social choice is built on two main components:

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

- **Theoretical guarantees:** The framework should be able to provide theoretical guarantees on the generated outcomes, and the process of generating them. This is usually done by using idealized preference queries, which are then used to generate the outcomes.
- **Empirical validation:** The generated outcomes should be empirically validated to ensure that they are representative of the preferences of the participants. This is usually done by comparing the generated outcomes with the preferences of the participants, and ensuring that they are in agreement.

The two components interact: The theory identifies queries that are useful for social choice and should hence be validated empirically. This idea is future-proof, since as LLMs continue to rapidly improve, so will their reliability in generating and answering queries, making LLM-based aggregation methods even more powerful.

2 Statement Selection Case Study

The generative social choice framework is demonstrated through a case study on statement selection — summarizing free-text survey responses into a representative set of statements. A multi-winner voting rule called *Balanced Justified Representation* (BJR) is used to ensure proportional representation of diverse opinions. The process consists of the following steps:

1. **Data collection:** Participants submit open-ended responses to a given topic, which form the free-text opinions.
2. **Theme extraction:** NLP methods (such as clustering) are used to extract major themes from the free-text responses.
3. **Statement generation:** LLMs are used to generate statements that capture the essence of the themes extracted.
4. **Application of voting rules:** A voting mechanism is used to select the set of most representative statements from the generated ones.
5. **Validation and feedback:** Additional participants rate the selected statements, and the feedback is used to refine the generated statements.

This method ensures that the final statements accurately represent the distribution of opinions within the participant pool.

3 Pilot Study on Chatbot Personalization

A pilot study was conducted to test the democratic process described in the paper in real-world conditions, by studying public opinions regarding chatbot personalization.

3.1 Study Design

The study was mainly composed of three phases, which are broadly as follows:

- **Phase 1:** Free-text opinions were collected from over 100 participants in the US about chatbot personalization.
- **Phase 2:** The framework was applied to get a democratic process which was aided by LLMs, and a representative set of 5 statements were obtained.
- **Phase 3:** A new group of 100 participants were given the above 5 statements, and asked to rate how representative these statements were.

3.2 Results

The results of the pilot study were as follows:

- 75% of the participants felt the statements were perfectly representative, and an additonal 18% rated them to be mostly representative.

- The final set of statements can be broken down to 3 major themes : Privacy and Data Security, User Control, and Truthfulness.
- The statements obtained, on validation, were found to be proportionally representative, and no major group was left out.

4 Benefits and Challenges of Generative Social Choice

4.1 Benefits

- **Democracy:** The framework ensures that decisions incorporate a diverse spectrum of perspectives by summarizing a large range of public opinions.
- **Efficiency:** The summarization and representation of large scale free-text input is automated, reducing the time and effort required to process the data.
- **Scalability:** The framework can be applied over a wide variety of fields, such as public policies, AI ethics, and corporate governance.
- **Transparency:** The framework uses structured algorithms to generate outcomes, ensuring fair representation of all opinions while reducing bias.

4.2 Challenges

- **LLM biases:** It is a known facts are heavily biased towards training data. Thus, elimination of this bias to ensure fair representation of all opinions is a challenge.
- **Computational scalability:** The time and computational resources required to process lage scale free-text datasets are very high, which leads to scalability issues.
- **Validation complexity:** the validation of generated outcomes is a complex process, and requires a large number of participants to ensure that the generated outcomes are representative of the preferences of the participants alongwith rigorous empirical testing.
- **Interpretability:** The generated outcomes are often difficult to interpret, and require a high level of domain-specific expertise to understand.

5 Plan of Action

I plan on using the framework of generative social choice to develop an AI-assisted policy drafting tool. First, I will **create a dataset** of policy proposals by collecting a diverse range of public opinions through structured surveys and open-ended questions. Next, I will **extract the major themes** from the given responses by the use of clustering methods (such as k-means clustering or hierarchical clustering). The **statement generation** phase will involve the use of LLMs to generate statements that capture the essence of the themes extracted. The generated statements will be reviewed to ensure factual accuracy and fair representation. Subsequently, I shall make use of the balanced justified representation (BJR) or similar multi-winner voting techniques to **choose a set of the most representative statements** from the generated ones. Finally, I will **validate the selected statements** by presenting them to additional participants for rating and feedback. The feedback will be used to refine the generated statements and ensure that they accurately represent the distribution of opinions within the participant pool. If required, I will iterate over the process to refine the generated statements further, and ensure that the final set of statements accurately represent the diverse opinions of the participants.

6 Conclusion

The Generative Social Choice framework outlined in the paper provides an exciting way of integrating powerful AI tools into the process of collective decision-making. It enhances the ability to generate diverse outcomes and extrapolate preferences by leveraging the power of LLMs, and is shown to have applications in various fields such as policy-making and ethics. Even though challenges such as bias and complex validation exist, the approach provides a transparent, efficient and scalable solution to existing decision-making systems.

References

[1] S. Fish, P. Gölz, D. C. Parkes, A. D. Procaccia, G. Rusak, and M. Wüthrich. Generative Social Choice. *arXiv:2309.01291*, 2023.

[2] H. Aziz, M. Brill, V. Conitzer, E. Elkind, R. Freeman, and T. Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 42(2):461-485, 2017.

[3] M. Bakker, M. Chadwick, H. Sheahan, M. Tessler, L. Campbell-Gillingham, J. Balaguer, N. McAleese, A. Glaese, J. Aslanides, M. Botvinick, and C. Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[4] M. Brill and J. Peters. Robust and verifiable proportionality axioms for multiwinner voting. In *Proceedings of the 14th ACM Conference on Economics and Computation (EC)*, 2023.

[5] B. Flanigan, P. Gölz, A. Gupta, B. Hennig, and A. D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548-552, 2021.

[6] N.Clegg. BringingpeopletogethertoinformdecisionmakingongenerativeAI. Blogpost,2023. Available at https://about.fb.com/news/2023/06/generative-ai-community-forum/.

[7] J. Hartmann, J. Schwenzow, and M. Witte. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv:2301.01768*, 2023.

[8] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 166-172, 2019.

[9] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, R. Noothigattu, D. See, S. Lee, C.-A. Psomas, and A. D. Procaccia. WeBuildAI: Participatory framework for fair and efficient algorithmic governance. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, article 181, 2019.

[10] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAI Conference on Artificial Intelligence (AAAI)*, pages 1587-1594, 2018.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730-27744, 2022.

[12] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *arXiv:2307.03172*, 2023.