Conversational Agents for a Deliberative Age

Michaël Grauwde, Mark Neerincx, Olya Kudina
Delft University of Technology: Technische Universiteit Delft, Delft, The Netherlands

Abstract

The discourse around conversational agents has become very prominent in 2023, with ChatGPT and LLMs taking the world by storm. However, the question remains of how to make value-aligned conversational agents. This paper takes a domain-specific examination, looking at value alignment for a democratic discourse. We analyze conversational agents, their use for deliberation, and how to design a conversational agent for deliberative purposes. The paper delves into a literature review while comparing theories like deliberative democracy and agonistic pluralism to gain insights into the properties of deliberation. Lastly, it presents some theoretical and technical of designing such conversational agents for deliberative purposes.

Introduction

Our aim is to build a value-aligned conversational agent for deliberation. To do so, we need to first analyze the domain of democratic discourse from where most theory on deliberation is written and from where we can base the conversational agent's design properties. As such, the paper will give a short introduction to conversational agents before analyzing deliberative democracy and its criticisms. We will then look more closely at the properties of deliberation and their role for the conversational agent. The paper will end by looking at some theoretical and technical challenges with creating a conversational agent for value explication.

Language models have taken over the imagination of what AI is and have stimulated interest in ethical questions about what an AI-supported future could bring. Ethical conundrums with language models and conversational agents have prompted thinking on aligning conversational agents with human values (Weidinger et al. 2021; Kasirzadeh and Gabriel 2023). Recent research by Kasirzadeh and Gabriel (2023) has looked at the way language models may require different norms for different discursive ideals to be successful in different domains. One domain analyzed is the roles of conversational agents in democratic discourse. Here, the authors speculate how a conversational agent can assist in deliberation between members of the public and the role of the conversational agent in this scenario (Kasirzadeh and Gabriel 2023). We build on this scenario in our research. Traditionally, deliberation is aimed toward consensus, with understanding between participants being a means to this end (Estlund and Landemore 2018). Our goal is slightly different as it is the explication of participants' values. Our goal is to build a conversational agent that assists participants in deliberation to reflect or explicate their values.

Our research questions are below:

RQ1. How can we use models of deliberation to build a more effective conversational agent for deliberative purposes?

RQ2. How can current conversational agent architecture help us in answering the above research question?

Conversational Agents

Conversational agents are programs that are designed to communicate with users using natural language (Jurafsky and Martin 2023). These are traditionally separated into two groups, task-oriented dialogue systems and chatbots (Jurafsky and Martin 2023). Task-oriented dialogue systems are systems like Siri and Alexa, and customer service applications, while chatbots are systems like ChatGPT and Bard from Google. The former use conversations with their users to help them complete tasks, while the latter mimic conversations between humans (Jurafsky and Martin 2023). The latter are often used for entertainment purposes but can be used for making task-oriented agents sound more natural (Jurafsky and Martin 2023). Recently, generative models have come to the fore with applications such as ChatGPT and LaMDA that are trained by transformer models which have a larger architecture and training data (Gozalo-Brizuela and Garrido-Merchán 2023). Conversational agents function in many domains, from personal daily assistants to conversational agents that assist in healthcare (Wahde and Virgolin 2022).

Deliberation

Deliberation can be defined in many ways but the description that we will use here is from Mansbridge et al. 2010, who define it as "communication that induces reflection on preferences, values, and interests in a non-coercive fashion" (p. 65). Our goal is to build a conversational agent that can aid in deliberative purposes for the value explication of deliberative participants. To build such an agent, we must analyze deliberation. We start with deliberative democracy.

Presented at the Works-in-Progress and Demonstrations track, AAAI HCOMP 2023.

Deliberative democracy is focused on engaging citizens in democratic discussion rather than in isolated responses of individuals to survey questions (Dryzek et al. 2019). It looks at the way a collective of individuals can come together to reason and make decisions on policies that will affect them directly (Cohen 2007; Bächtiger et al. 2018). Early deliberative democratic theorists, such as Jurgen Habermas, Joshua Cohen, and John Rawls portrayed it as an alternative to existing aggregative liberal democracy (Bächtiger et al. 2018). However, the field has received criticism from scholars such as Chantal Mouffe who have argued for a more pluralistic and open model of democracy. Mouffe argues that deliberative democracy through its focus on rationality, prefers certain individuals over others thus reducing the amount of plurality (Mouffe 2000). She argues that the theory also sees emotion as a negative element, while emotion is inherent to politics. Also, it focuses on consensus which contrasts with deliberation. Consensus attempts to close the discussion, while deliberation is inherently open-ended (Jezierska 2019). While deliberative democracy avoids conflict Mouffe argues that conflict is essential and insists that democracy should strive to use conflicts as a manner to develop new democratic designs (Mouffe 1999). Deliberative democracy has since early proponents seen new attempts to deal with its shortfalls.

Recently, more scholars have begun conceptualizing deliberative democracy without consensus (Chambers 2003). A focus on consensus can downplay pluralism in what matters to people and so the values of other stakeholders. The risks here are that if a discussion aims at reaching a consensus, the focus can become one of conformity and participants can be excluded. So, consensus risks erasing the plurality of values and matters of concern. Scholar I.M. Young states that traditional deliberation preferences impassioned, logical, and reasoned communication which often privileges male voices and voices of privileged groups (Young 1996). This goes against the goals of deliberation, namely equality and inclusion (Jezierska 2019). Mouffe states that with true pluralism of values, there will be an increase in conflict but that this is inherently better for understanding your fellow stakeholders and where they are coming from (Mouffe 1999). Conflict can engender awareness of certain problems so it's vital for stakeholder dialogue (Brand, Blok, and Verweij 2020). Agreements will only be temporary, so conflict is necessary as new situations will produce new problems which will require new responses. Conflict can also be valuable in fighting power imbalances. Traditionally, deliberative democrats will argue against revealing the self-interest of the stakeholders involved, however, in stakeholder dialogue, revealing the interest and position of the stakeholder can be informative and have an influence on the different positions that the stakeholders take (Brand, Blok, and Verweij 2020). Self-interest is vital in deliberation because, without it, mutual understanding and respect between parties can be prevented (Mansbridge 2006). Agonistic deliberation allows for pluralism, emotion, and passion to enter the deliberative sphere. This is the version of deliberation that is conducive to the goal of stakeholders' value reflection, and the one we aim to clarify further.

Properties of Deliberation

While above we have looked at deliberation from its different theoretical positions, in this section we analyze some properties that lead to better and more successful deliberation. While critics of Habermasian deliberation realize the necessity for updated theories, some norms such as civility, legitimacy of opponents in the deliberation, and justification of all the views being deliberated amongst the participants remain vital (Dryzek et al. 2019). Mutual justification of the views of the participants can be argued as a goal in deliberation (Brand, Blok, and Verweij 2020). In this way, mutual justification can reveal participants' views without necessitating that the participants change their minds based on the views of their fellow participants. Advocates for agonistic deliberation also argue that this type of deliberation can play a vital role in developing values, without marginalizing certain groups due to power imbalances (Dahlberg 2007).

In conducting an inductive study of deliberation, Mansbridge et al. (2006) found that the free flow of speech which is honest and not restricted due to fear of retaliation is a vital property of deliberation (Mansbridge et al. 2006). Free flow of speech allows for space to challenge one another and for spaces of conflict. This brings together values of freedom, respect, dissent, and can provide dialogue that is more understandable due to the free flow of speech. Another property is equality, which is determined by the inclusive nature of participation in discussion, self-facilitation and group control, and fair representation of views. And lastly, an important property is reason and emotional input (Mansbridge et al. 2006). Facilitators of deliberations found that emotion-layered deliberation leads to better elicitation of ideas rather than solely relying on facts (Mansbridge et al. 2006). A combination of the two motivates participants to work better together on tasks. From there, we can get a glimpse into the norms that we want to promote to reach our goal.

Related Work

The research done on conversational agents' use in deliberation is sparse. Nevertheless, chatbots have been used as moderators in deliberative discussions. Kim et al. (2021) used a chatbot to moderate discussions and found that the chatbot stimulated discussion and resulted in higher discussion quality. This chatbot, however, like most deliberations was aimed towards consensus. The same is true for Shin et

al. (2022) who designed a chatbot for consensus-building between various stakeholders. However, in the research, the stakeholders used the chatbot as a mediator through which to build consensus for co-design. There was no direct interaction between the stakeholders in the discussion.

Hadfi et al. (2021) developed an argumentative agent within an online deliberative space to analyze the way the agent could influence discussions among participants. The conversational agent was seen to increase the responsiveness of the participants and the ability to generate solutions to issues that were previously raised (Hadfi et al. 2021). The above examples present conversational agents that are working in deliberative settings; however, they work with the goal of consensus building.

Theoretical Challenges

As we design a conversational agent to assist in deliberative purposes that does not intend to pursue consensus and welcomes conflict and emotions; it is essential to recognize what the conversational agent will be doing in terms of assisting. The conversational agent can replace or complement the facilitator. Our conversational agent is tasked with assisting participants in value explication/reflection. This differs from the agents that have been discussed above which are geared towards consensus building as a goal. However, the agent could also support deliberation through reflection with each participant to explicate their values prior to deliberation which can then be adjusted during the deliberative process. There could also be an agent that could assist each participant in the deliberation. Another example could be the facilitation of the deliberation after a human facilitator has provided values to be deliberated.

As deliberation will take part among numerous differing stakeholders, the position of agonistic deliberation can be vital as conflict can better elicit values in a structured manner (Mouffe 1999). It is necessary to create a conversational agent that can function while considering the properties of deliberative such as civility, free-flowing information, emotion, mutual understanding, and equality. If the instrumental goal of deliberation is making progress on the task while the normative goal is one of non-domination, the conversational agent should be designed in such a manner that it allows these goals to thrive. Furthermore, it should also be designed to account for other norms, such as civility in the discussion and the promotion of free-flowing information. It can aim to do so by helping participants express their emotions so there may be mutual understanding. It is only from this position that we can design a conversational agent that will help our participants during deliberative processes to explicate values that are important to them. The agent will also be designed in a value-sensitive design manner. In this way, we can take in broad stakeholder perspectives, look at the stakeholders impacted, while also getting a broad analysis of the values that are vital throughout the interaction with the agent (Friedman, Kahn Jr., and Borning 2006).

Technical Challenges

Based on the analysis, there are several insights relevant to developing a conversational agent that promotes value reflection. Like conversational agents such as Siri, this conversational agent will be a task-oriented agent functioning in a closed domain (Jurafsky and Martin 2023). The goal of the agent is to assist users in the explication of their values during deliberative processes. Furthermore, as foreseen interaction may be lengthy, the agent should be able to store information for extended periods of time. A long-short-term memory network (LSTM) can assist the agent to remember for the length of the deliberation and to pull information between distant periods in time (Hussain et al. 2019). A bidirectional long-short term memory classifier can extract data from the deliberation related to values and can observe the overall content of the deliberation (Suzuki et al. 2020).

One important feature of deliberation is free-flowing speech. So, an agent should know when to jump into a conversation so they can process the recent utterances and respond. This is endpoint detection which helps the agent recognize when the user is done speaking (Jurafsky and Martin 2023). The agent should also solicit the views of individuals that are quieter than others in the deliberation as this can allow for more equality and diversity in the deliberation (Kim et al. 2021). To stimulate more emotionally laden deliberation, the agent should have the ability to nudge certain users to be more expressive in presenting their values. As a person has responded, the agent recognizes a lack of emotional content, it can nudge the participant to delve deeper in their answer, asking, “Why does [X name] feel this way?” (Kim et al. 2021). While this deliberative agent takes a different approach than traditional deliberations and other agents, through insights from prior agents, we hope to build an agent that can attain the goal of value explication.

Conclusion

In this paper, we analyzed the way to design a conversational agent for value explication. We proposed two research questions that form the basis of our research on this topic. We examined the existing literature on conversational agents, deliberation, and deliberative conversational agents; before presenting some challenges that we may face with designing our own agent. We hope this research can provide insights for facilitating further research into deliberative conversational agents.

References

Bächtiger, A.; Dryzek, J. S.; Mansbridge, J.; and Warren, M. 2018. Deliberative Democracy. In *The Oxford handbook of deliberative democracy*, edited by A. Bächtiger, J. S. Dryzek, J. Mansbridge, M. E. Warren 1-34. Oxford. University Press.

Brand, T.; Blok, V.; and Verweij, M. 2020. Stakeholder Dialogue as Agonistic Deliberation: Exploring the Role of Conflict and Self-Interest in Business-NGO Interaction. *Business Ethics Quarterly*, *30* (1), 3-30. doi:10.1017/beq.2019.21.

Chambers, S. 2003. Deliberative Democratic Theory. Annual Review of Political Science, 6 (1), 307-326. doi.org/10.1146/annurev.polisci.6.121901.085538.

Cohen, J. 2007. Deliberative Democracy. In *Deliberation, Participation and Democracy,* edited by: Rosenberg, S.W., 219-236. Palgrave Macmillan.

Dahlberg, L. 2007. The Internet and Discursive Exclusion: From Deliberative to Agonistic Public Sphere Theory. In *Radical Democracy and the Internet: Interrogating Theory and Practice,* 128 – 147. London: Palgrave Macmillan UK. doi.org/10.1057/9780230592469_8.

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.; and Warren, M.E. 2019. The Crisis of Democracy and the Science of Deliberation. *Science*, 363, 1144 – 1146. doi.org/10.1126/science.aaw2694.

Estlund, D.; and Landemore, H. 2018. The Epistemic Value of Democratic Deliberation. In *The Oxford handbook of deliberative democracy,* edited by Bächtiger, A.; Dryzek, J. S.; Mansbridge, J. and Warren, M. E. 113-131. Oxford University Press.

Friedman, B.; Kahn, P. H.; and Borning, A. 2006 (2013). Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the Laboratory*, edited by Doorn, N.; Schuurbiers, D.; van de Poel, I.; and Gorman, M.E., 55-95. Springer International Publishing. doi.org/10.1007/978-94-007-7844-3_4.

Gozalo-Brizuela, R.; and Garrido-Merchan, E. C. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. arXiv preprint arXiv:2301.04655. Ithaca, NY: Cornell University Library.

Hafdi, R.; Haqbeen, J.; Sahab, S.; and Ito, T. 2021. Argumentative Conversational Agents for Online Discussions. *Journal of Systems Science and Systems Engineering,* 30, 450 – 464. doi.org/10.1007/s11518-021-5497-1.

Hussain, S.; Ameri Sianaki, O.; and Ababneh, N. 2019. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. Web, Artificial Intelligence and Network Applications. In Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33 (pp. 946-956). Springer International Publishing. doi.org/10.1007/978-3-030-15035-8_93.

Jezierska, K. 2019. With Habermas against Habermas. Deliberation without Consensus. *Journal of Public Deliberation* 15(1). doi.org/10.16997/jdd.326.

Jurafsky, D.; and Martin, J. 2023. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd Edition draft.

Kasirzadeh, A.; and Gabriel, I. 2023. In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology,* 36 (2), 1 – 24. doi.org/10.1007/s13347-023-00606-x.

Kim, S.; Eun, J.; Seering, J.; and Lee, J. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. In Proceedings of the ACM on Human-Computer Interaction,.1-26. doi.org/10.1145/3449161.

Mansbridge, J. 2006. Conflict and Self-Interest in Deliberation. In *Deliberative democracy and its discontents*, edited by Jose Luis Marti, Samantha Besson 107-32. Routledge.

Mansbridge, J., Bohman, J., Chambers, S., Estlund, D., Føllesdal, A., Fung, A., Lafont, C., Manin, B. and Martí, J.L. 2010. The Place of Self-Interest and the Role of Power in Deliberative Democracy. Journal of PoliticalPhilosophy,18:64-100. doi.org/10.1111/j.1467-9760.2009.00344.x

Mansbridge, J.; Hartz-Karp, J.; Amengual, M.; and Gastil, J. 2017. (2006). Norms of Deliberation: An Inductive Study. In *Multi-Party Dispute Resolution, Democracy and Decision-Making*, edited by Carrie Menkel-Meadow, 139-185. Routledge.

Mouffe, C. 1999. Deliberative Democracy or Agonistic Pluralism?. In *Social research*, 745-758.

Mouffe, C. 2000. For an agonistic model of democracy. In Political Theory in Transition, Edited by N. O'Sullivan London: Routledge.

Shin, J.; Hedderich, M. A.; Lucero, A.; and Oulasvirta, A. 2022. Chatbots Facilitating Consensus-Building in Asynchronous Co-Design. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (pp. 1 -13). doi.org/10.1145/3526113.3545671

Suzuki, S.; Yamaguchi, N.; Nishida, T.; Moustafa, A.; Shibata, D.; Yoshino, K.; Hiraishi, K. and Ito, T. 2020. Extraction of Online Discussion Structures for Automated Facilitation Agent. In Advances in Artificial Intelligence: Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence (JSAI 2019) 33 150-161. Springer International Publishing. doi.org/10.1007/978-3-030-39878-1_14.

Wahde, M. and Virgolin, M. 2022. Conversational Agents: Theory and Applications. In *Handbook on Computer Learning and Intelligence: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation* edited by P. P. Angelov 497-544. World Scientific Publishing. doi.org/10.1142/9789811247323_0012.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S.; Hawkins, W.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and Social Risks of Harm from Language Models. arXiv preprint arXiv:2112.04359. Ithaca, NY: Cornell University Library.

Young, I. M. 1996. Communication and the Other: Beyond Deliberative Democracy. In *Democracy and Difference: Contesting the Boundaries of the Political*, edited by S. Benhabib 120 – 136. Princeton University Press. doi.org/10.2307/j.ctv1nxcvsv.9