TRUST IN AI MEDIATORS MAY CHANGE DELIBERATIVE OUTCOMES

Joshua Cohen[1] and Henrik D. Kugelberg[2]

eLetter responding to Michael Henry Tessler et al., "AI can help humans find common ground in democratic deliberation," Science 386, (2024). DOI:10.1126/science.adq2852

In their Research Article, "AI can help humans find common ground in democratic deliberation," M.H. Tessler *et al.* demonstrate impressive results from the Habermas Machine, an AI system designed to facilitate human deliberation. The Habermas Machine outperforms human mediators at reducing intra-group divisions, generating widely preferable group opinion statements, and fostering consensus.

These findings are very promising. But it is worth considering whether they are partly driven by participants' misperceptions of algorithmic objectivity. Previous studies have shown that people tend to trust algorithms because they see them as unbiased. In contrast, when people trust human decision-makers it is often because they are viewed as having a certain authority (Lee 2018). Because participants knew whether they were interacting with a Habermas Machine or a human, similar mechanisms may have contributed to the result. Perhaps participants thought better of the statements generated by the Habermas Machine because it was perceived as an objective mediator, and less well of those from human mediators because of their perceived lack of authority.

If so, some of the effect observed by M.H. Tessler *et al.* may stem from a lack of participant understanding of the design decisions that went into the Habermas Machine—for example, the use of the Schulze method for ranking statements, as well as fine-tuning decisions concerning the underlying model. It is well-known that different social choice procedures can generate very different collective choices from the same inputs (Arrow 1961). A Habermas Machine that, for instance, used the Borda method instead of the Schulze method would routinely select different winning statements, as the methods aggregate individual preferences differently. Similarly, changes to the model parameters would shift which statements are selected as most representative.

The striking findings may therefore be partly due to misconceptions about the objectivity of the Habermas Machine. If illusions about machine objectivity generate more trust in the results, this could both be beneficial and raise ethical and democratic concerns. It is desirable for deliberative procedures to generate agreement. However, it is concerning if agreement is founded on a misconception about the procedure and its outcomes. Future research could investigate if the results change when people do not know the source of the statements, or if they see how different machine designs yield different statements. This would give us better understanding of how, when, and why people trust algorithmic mediators.

K. J. Arrow. *Social Choice and Individual Values*, 2nd ed. Wiley, 1963.

M. K. Lee. "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management." *Big data & society* 5.1, 2018.

[1] Apple University and University of California Berkeley.
[2] University of Warwick.