RESEARCH ARTICLE SUMMARY

ARTIFICIAL INTELLIGENCE

AI can help humans find common ground in democratic deliberation

Michael Henry Tessler*†, Michiel A. Bakker*†, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick*, Christopher Summerfield*

INTRODUCTION: Democracy, at its best, rests upon the free and equal exchange of views among people with diverse perspectives. Collective deliberation can be effectively supported by structured events, such as citizens' assemblies, but such events are expensive, are difficult to scale, and can result in voices being heard unequally. This study investigates the potential of artificial intelligence (AI) to overcome these limitations, using AI mediation to help people find common ground on complex social and political issues.

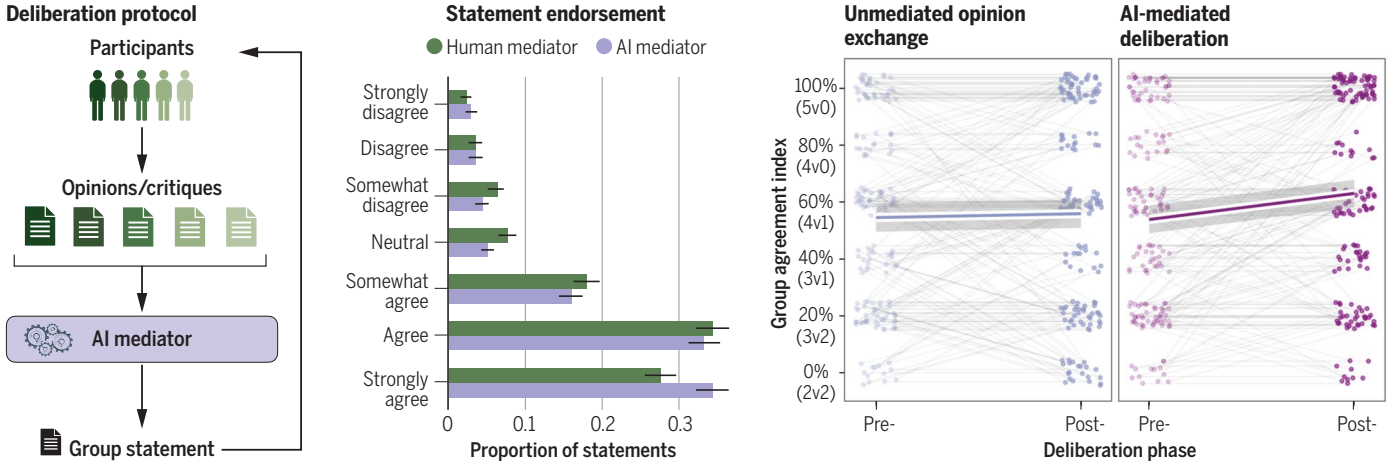
RATIONALE: We asked whether an AI system based on large language models (LLMs) could successfully capture the underlying shared perspectives of a group of human discussants by writing a "group statement" that the discussants would collectively endorse. Inspired by Jürgen Habermas's theory of communicative action, we designed the "Habermas Machine" to iteratively generate group statements that were based on the personal opinions and critiques from individual users, with the goal of maximizing group approval ratings. Through successive rounds of human data collection, we used supervised fine-tuning and reward modeling to progressively enhance the Habermas Machine's ability to capture shared perspectives. To evaluate the efficacy of AI-mediated deliberation, we conducted a series of experiments with over 5000 participants from the United Kingdom. These experiments investigated the impact of AI mediation on finding common ground, how the views of discussants changed across the process, the balance between minority and majority perspectives in group statements, and potential biases present in those statements. Lastly, we used the Habermas Machine for a virtual citizens' assembly, assessing its ability to support deliberation on controversial issues within a demographically representative sample of UK residents.

RESULTS: Group opinion statements generated by the Habermas Machine were consistently preferred by group members over those written by human mediators and received higher ratings from external judges for quality, clarity, informativeness, and perceived fairness. AI-mediated deliberation also reduced division within groups, with participants' reported stances converging toward a common position on the issue after deliberation; this result did not occur when discussants directly exchanged views, unmediated. Although support for the majority position increased after deliberation, the Habermas Machine demonstrably incorporated minority critiques into revised statements. We replicated these results in a virtual citizens' assembly, additionally finding that during AI-mediated deliberation, the views of groups of discussants tended to move in a similar direction on controversial issues. These shifts were not attributable to biases in the AI, suggesting that the deliberation process genuinely aided the emergence of shared perspectives on potentially polarizing social and political issues.

CONCLUSION: This research demonstrates the potential of AI to enhance collective deliberation by finding common ground among discussants with diverse views. The AI-mediated approach is time-efficient, fair, scalable, and outperforms human mediators on key dimensions. Rather than simply appealing to the majority, the Habermas Machine prominently incorporated dissenting voices into the group statements. AI-assisted deliberation is not without its risks, however; to ensure fair and inclusive debate, steps must be taken to ensure users are representative of the target population and are prepared to contribute in good faith. Under such conditions, AI may be leveraged to improve collective decision-making across various domains, from contract negotiations and conflict resolution to political discussions and citizens' assemblies. The Habermas Machine offers a promising tool for finding agreement and promoting collective action in an increasingly divided world. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: mhtessler@google.com (M.H.T.); miba@google.com (M.A.B.); botvinick@google.com (M.B.); christopher.summerfield@psy.ox.ac.uk (C.S.)
†These authors contributed equally to this work.
Cite this article as M. H. Tessler *et al.*, *Science* **386**, eadq2852 (2024). DOI: 10.1126/science.adq2852

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adq2852



AI helps people find common ground in collective deliberation. (Left) The AI mediator uses participants' opinions to generate group statements and iteratively refines those statements through participants' critiques. (Middle) Statements from the AI mediator (purple) garner stronger endorsement than those written by a human mediator (orange). (Right) AI mediation leaves groups less divided after deliberation, whereas simply sharing opinions with others does not.

RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

AI can help humans find common ground in democratic deliberation

Michael Henry Tessler[1]*†, Michiel A. Bakker[1]*†, Daniel Jarrett[1], Hannah Sheahan[1], Martin J. Chadwick[1], Raphael Koster[1], Georgina Evans[1], Lucy Campbell-Gillingham[1], Tantum Collins[1], David C. Parkes[1,2], Matthew Botvinick[1,3]*, Christopher Summerfield[1,4]*

Finding agreement through a free exchange of views is often difficult. Collective deliberation can be slow, difficult to scale, and unequally attentive to different voices. In this study, we trained an artificial intelligence (AI) to mediate human deliberation. Using participants' personal opinions and critiques, the AI mediator iteratively generates and refines statements that express common ground among the group on social or political issues. Participants (N = 5734) preferred AI-generated statements to those written by human mediators, rating them as more informative, clear, and unbiased. Discussants often updated their views after the deliberation, converging on a shared perspective. Text embeddings revealed that successful group statements incorporated dissenting voices while respecting the majority position. These findings were replicated in a virtual citizens' assembly involving a demographically representative sample of the UK population.

Human society is enriched by a plurality of legitimate viewpoints, but agreement is a prerequisite for people to act collectively (*1*). To find agreement, people typically gather in person (or online) for an unstructured or semistructured deliberation characterized by free exchange of opinions. Egalitarian and open-minded deliberation is a cornerstone of liberal democracy, often formally realized through citizens' assemblies, in which small but representative groups of unelected citizens are randomly selected to discuss controversial issues (*2*, *3*). To date, several nation-states, including France, Canada, and Iceland, have used large-scale citizens' assemblies to make key national policy decisions (*4*).

However, as a method for finding agreement, the free exchange of opinions has well-known limitations. Citizens' assemblies can be costly and time-consuming for organizers and participants alike, and deliberation is only possible among groups of limited size (*5*). Moreover, voices may be heard unequally during the debate, some discussants may strategically adopt extreme views to maximize their sway, and social desirability or group affiliation effects may lead beliefs to become entrenched (*6*, *7*). Correspondingly, research into the efficacy of citizens' assemblies has yielded mixed results, suggesting that they can produce either homogenizing or polarizing effects on opinions, either better or worse decisions, or either increases or decreases in future political participation (*8*).

Here, we asked whether newly available tools from artificial intelligence (AI) could be used to support collective deliberation (*9*). Our work is inspired by demonstrations showing that recent generations of large language models (LLMs) can effectively summarize perspectives on a public deliberation platform (*10*), generate statements that showcase the full range of a group's political views (*11*), temper partisan discussion by suggesting equanimous rewrites (*12*), and mitigate gender imbalance in debate participation by positively contributing to discussions about sensitive issues (*13*). Technologists and human-computer interaction researchers have prototyped various systems for assisting deliberation (*14*–*16*) but have not systematically evaluated their effectiveness, measured their impact on human beliefs, or studied the semantic or doxastic properties of the resulting deliberation. Nor do we know whether AI-based systems can help people find agreement in a virtual citizens' assembly.

We developed an approach to collective deliberation in which an AI system is trained to be a "caucus mediator." A caucus mediator meets privately with each discussant before making a proposal designed to be collectively acceptable (*17*). In our study, participants submitted their personal opinions on social and political issues to an LLM that had been trained to generate a "group statement" designed to maximize endorsement and thus to help them find common ground (*18*, *19*). We call this AI system the "Habermas Machine" (HM), after the theorist Jürgen Habermas, who proposed that when rational people deliberate under idealized conditions, agreement will emerge in the public sphere (*20*).

We focused on four research questions (RQs):

RQ1: Does AI-mediated deliberation help people find common ground?

RQ2: Does AI-mediated deliberation leave groups less divided?

RQ3: Does the AI mediator represent all viewpoints equally?

RQ4: Can AI mediation support deliberation in a citizens' assembly?

To answer these questions, we evaluated the impact of the HM on the deliberation process through a set of structured experiments involving human participants (total N = 5734). Our main hypothesis was that the HM would help people find "common ground" (RQ1). When a group finds common ground, they agree upon a pool of shared information that can be used to inform later proposals or outcomes (*19*). We also studied the extent to which the deliberative process using the HM reduced division within groups (RQ2), whether the HM was prone to suppress minority voices (RQ3), and how it could be used in practice (RQ4).

Method

Large language model assistant: The Habermas Machine

The HM is a system of LLMs with the task of finding common ground among a group of people discussing a social or political issue. It has two components, both of which were fine-tuned versions of a pretrained LLM [*Chinchilla* (*21*)]. A generative model that has undergone supervised fine-tuning proposes high-quality candidates for the "group statement," and a personalized reward model (PRM) scores the candidates according to the predicted preferences of each individual group member (*22*–*24*) (Fig. 1B). The HM decides which candidate to return by aggregating the PRM rankings using a social choice function with the "independence of clones" property, ensuring that very similar candidate statements do not split the vote (*25*). The HM also has the capacity to iteratively revise a group statement by incorporating written critiques from the individual group members through the same generation and selection process. It is thus designed to arbitrate among the various views by private interaction with each discussant, a process known as "caucus mediation" (*17*). Supplementary materials (SM) 3 provides further details on the HM.

Samples and participants

Participants were adult UK residents. Most participants were recruited by convenience sampling through an online crowdsourcing research platform. However, we also recruited a demographically representative sample of

[1]Google DeepMind, London, UK. [2]Department of Computer Science, Harvard University, Cambridge, MA, USA. [3]Yale Law School, New Haven, CT, USA. [4]Department of Experimental Psychology, University of Oxford, Oxford, UK.
*Corresponding author. Email: mhtessler@google.com (M.H.T.); miba@google.com (M.A.B.); botvinick@google.com (M.B.); christopher.summerfield@psy.ox.ac.uk (C.S.)
†These authors contributed equally to this work.

the UK population to participate in a virtual citizens' assembly, using a randomized, stratified sampling procedure (implemented by the Sortition Foundation). Deliberation questions covered potentially divisive issues in UK public policy (such as "should we lower the voting age to 16?" or "should the National Health Service be privatised?"). Candidate questions were generated by prompting an LLM with a small corpus of seed exemplars. After careful vetting to ensure questions minimized the risk of provoking offensive commentary, we arrived at a dataset of approximately 5500 distinct questions (SM, materials and methods, and SM 1).

Group deliberation procedure

Our experiments used a mediated collective deliberation procedure (Fig. 1A) in which small groups of people (typically sized at five) discussed three questions sequentially over the course of approximately 1 hour. For each question, participants were first given 5 min to write a short paragraph describing their personal opinion in private (10 to 200 words, average length 65 words). Participants typically provided both a description of their position and an accompanying justification. These opinions were passed to the HM, which generated a set of initial group statements. Participants then rated how strongly they endorsed each initial group statement and the quality of the argument (regardless of their opinion). They also ranked each statement according to the extent they agreed with each. The winning initial group statement (selected through a ranked choice election using the same social choice rule used in the HM) was returned to participants for them to critique in another short paragraph, again in private. Critiques were then sent to the HM, which generated a new set of revised group opinion statements. Participants again rated and ranked these statements. Participants were then asked to indicate which of the winning (top-ranked) initial and revised group opinion statements they preferred (see Fig. 1C for an example winning revised group statement). To conclude the deliberation round, participants completed a short survey, asking about whether and how their views might have been affected by the process. Immediately before and after the full set of deliberation rounds, participants completed a survey in which they rated their positions on the questions they deliberated over. Participants were made aware that the group statements were generated by an algorithm.

Results

RQ1: Does AI-mediated deliberation help people find common ground?

To test whether AI mediation helps people find common ground and to benchmark against human mediation, we recruited 75 six-person testing groups, with one participant in each group randomly assigned the role of "mediator" (n = 439 participants). Mediators were trained to write effective group statements and were incentivized by the promise of a generous financial bonus for each statement they wrote that was most preferred by the group (see SM 9.1.2 and 9.1.3 for full instructions and SM 7.2 for a broader discussion of mediator training). In each round of deliberation, participants saw one statement from the HM and one from the human mediator. (There was no critique phase with the human mediator.) The HM's statements were preferred by the participants 56% of the time versus 44% for the human mediators' statements [logistic mixed-effects model predicted probability $\hat{P}$ = 0.575 (0.535, 0.614), P < 0.001]. In addition, ordinal mixed-effects regression models confirmed that the HM statements were rated as higher quality and were endorsed more strongly than those written by the human mediators [$\beta_{endorsement}$ = 0.31, standard error (SE) = 0.07, odds ratio (OR) = 1.36, z score = 4.2, P < 0.001; $\beta_{quality}$ = 0.43, SE = 0.08, OR = 1.54, z score = 5.6, P < 0.001; Fig. 2A and SM 4.1.3]. This result demonstrates that the task the HM is performing is nontrivial and indeed challenging for humans.

The HM might have received higher levels of endorsement because it wrote longer statements than the human mediators [125.9 words in length,

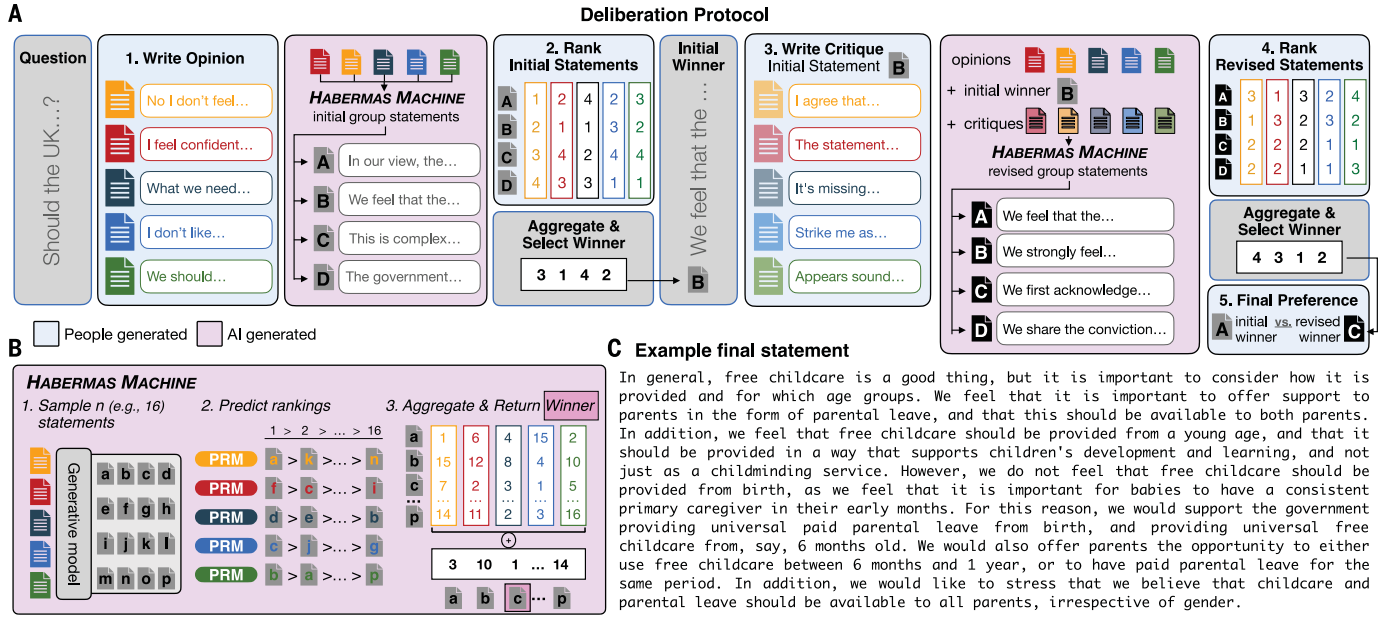


Fig. 1. Overview of methods. (**A**) Mediated deliberation procedure. **1.** Participants, organized into small groups, privately wrote an opinion statement in response to a question. The Habermas Machine (HM) generated candidate initial group statements from the group's individual opinions. **2.** Participants ranked these initial statements. The top-ranked statement, on the basis of aggregated rankings, was returned to the group. **3.** Participants privately wrote critiques of the initial winner. The HM generated revised group statements from the group's critiques (along with the initial opinions and initial group winner). **4.** Participants ranked these revised statements, and the winner was again selected through aggregated rankings. **5.** Participants made a final preference judgement between the initial and revised winning statements. A deliberation round for a single question lasted approximately 15 min. (**B**) The HM produces a group statement through a simulated election. **1.** A generative model samples many candidate group statements. **2**. A personalized reward model produces predicted rankings for each person in the group. **3.** The top-ranked statement, on the basis of aggregated rankings, is returned. (**C**) Example top-ranked revised group opinion statement, from the virtual citizens' assembly (see SM 6 for full example, including the opinions and critiques).

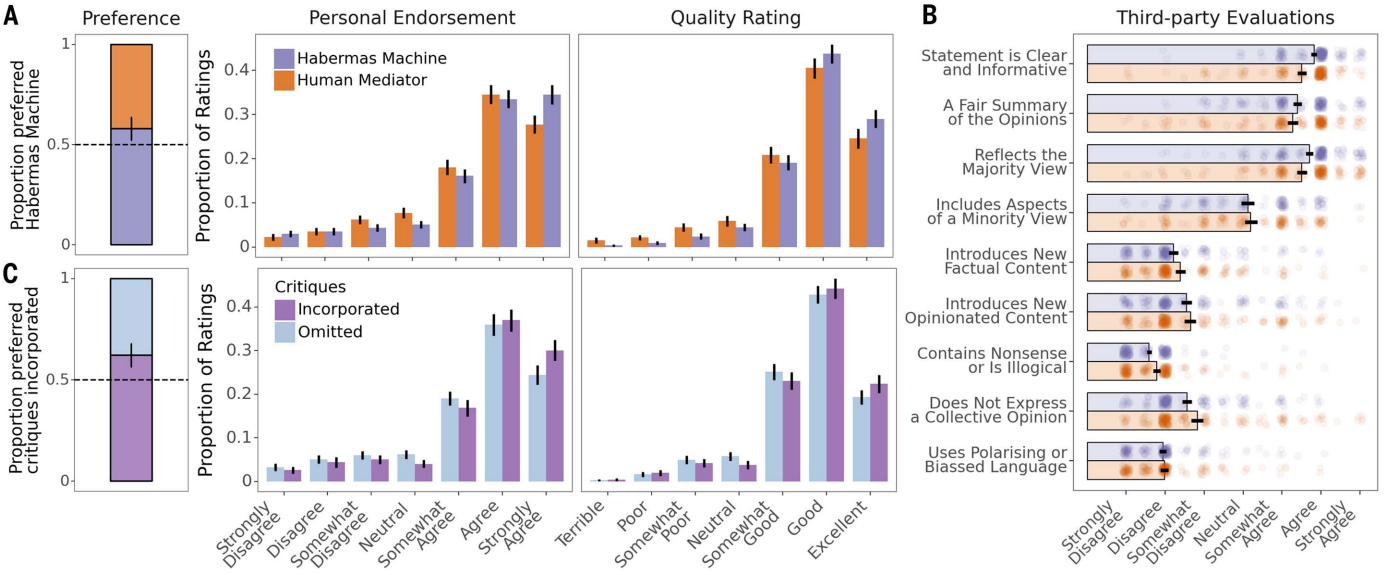


Fig. 2. The Habermas Machine generates high-quality group opinion statements that are preferred to human-written group statements, and critiquing provides further improvements. (**A**) (Left) Proportion of rankings that prefer the Habermas Machine (HM) statement over a human-written group statement. (Right) Endorsement and quality ratings of the HM- and human-written group statements in terms of relative frequencies of each Likert rating. (**B**) Third-party ratings of nine dimensions of quality in terms of average Likert ratings. Points indicate median ratings for a single statement, averaged over five (third-party) raters. (**C**) (Left) Proportion of rankings that prefer a revised statement that incorporates the critiques (purple) over an initial statement in which critiques were omitted (light blue). (Right) Endorsement and quality ratings, from the postcritique phase for the critique exclusion cohort. All error-bars denote bootstrapped 95% CIs.

95% confidence interval (CI) (121.4, 130.1), compared with 105 (100.2, 108.5) words] (*26*), although we refined our methods for encouraging longer human-mediated statements over several pilot tests and conducted additional tests to verify compliance (SM 4.3.2). Participants' statement rankings as well as their endorsement and quality ratings were all impacted by length; the endorsement ratings ("do you personally agree with this statement?"), however, still showed that the HM's statements were preferred to those of the humans, controlling for length (SM 4.3.6). In addition, a separate cohort of third-party raters (n = 155) judged HM-generated group statements to be clearer and more informative (β = 0.449, SE = 0.14, z score = 3.26, P = 0.001), less illogical (β = −0.26, SE = 0.11, z score = −2.28, P = 0.022), better at capturing the majority perspective (β = 0.225, SE = 0.11, z score = 2.08, P = 0.037), and less likely to represent the individual opinion of the group statement writer (β = −0.226, SE = 0.10, z score = 2.19, P = 0.028) than the human-written statements (Fig. 2B). (No statistical differences were observed between the HM- and human-generated statements on whether they provided a fair summary of the opinions, whether they included the minority view, or whether they were polarizing; full results provided in SM 4.3.7.) The HM thus was able to generate statements that garnered high levels of endorsements after controlling for length and that were viewed as higher quality by external judges on a number of dimensions.

Next, we asked which features of the AI model and deliberative process were important for securing positive endorsement of the group statements. We collected three additional cohorts of data (main task cohorts 1 to 3; n = 1692 participants in 349 groups completing 1047 deliberation rounds) to perform ablation analyses (to measure the importance of the PRM and supervised fine tuning) and to verify that our results generalized to new, out-of-distribution discussion topics that had not been included in the training data (SM 4.4). These datasets also allowed us to measure whether participants' critiques and subsequent revisions by the HM led participants to prefer the revised statements over the initial statements (final preference trial; Fig. 1A, step 5). Across all three cohorts, we found a consistent preference in the final preference trial for revised group opinion statements over initial statements [cohort 1: paired preference = 68%; 95% CI (66, 71), P < 0.001; cohort 2: 64% (62, 66), P < 0.001; cohort 3: 66% (64, 69), P < 0.001; we found a significant preference controlling for length; see SM 4.4.5 for details].

One risk in this evaluation is that revised statements might be preferred simply because they were encountered later and after a slight additional investment of effort, driving an effort justification bias (*27*). To rule out this alternative explanation, we conducted a further experiment with a new set of participants (critique exclusion cohort 4; n = 245 participants in 50 groups). These participants performed a comparable experiment, except that, after writing critiques, they rated and ranked group statements that were generated either by an HM that did not observe the critiques or by one that did. Revised statements were once again ranked more favorably by the groups, a result that remained significant after controlling for statement length (Fig. 2C, β = 0.34; SE = 0.12; z score = 2.8, P < 0.0039), indicating that the mediated deliberation process was helpful for finding common ground (see SM 4.5 for more details).

RQ2: Does AI-mediated deliberation leave groups less divided?

Common ground is built by mutual acceptance of a common pool of information (*19*), but perspective shifts may be a precursor to finding consensus (*20*). We thus asked whether interacting with the HM prompted participants' reported stance on the issues to move in a consistent direction. We calculated a metric called "group agreement," which indexed the relative proportion of participants who agreed or disagreed with the position statement (collected before and after the full set of deliberation rounds; SM, materials and methods). For example, if 80% of participants "agree" and 20% "disagree" (to any extent) with the position statement (e.g., "Taxes are too high"), then the group agreement is the difference of these proportions, or 60 percentage points (if a group is perfectly split down the middle, its group agreement is 0). We asked whether group agreement

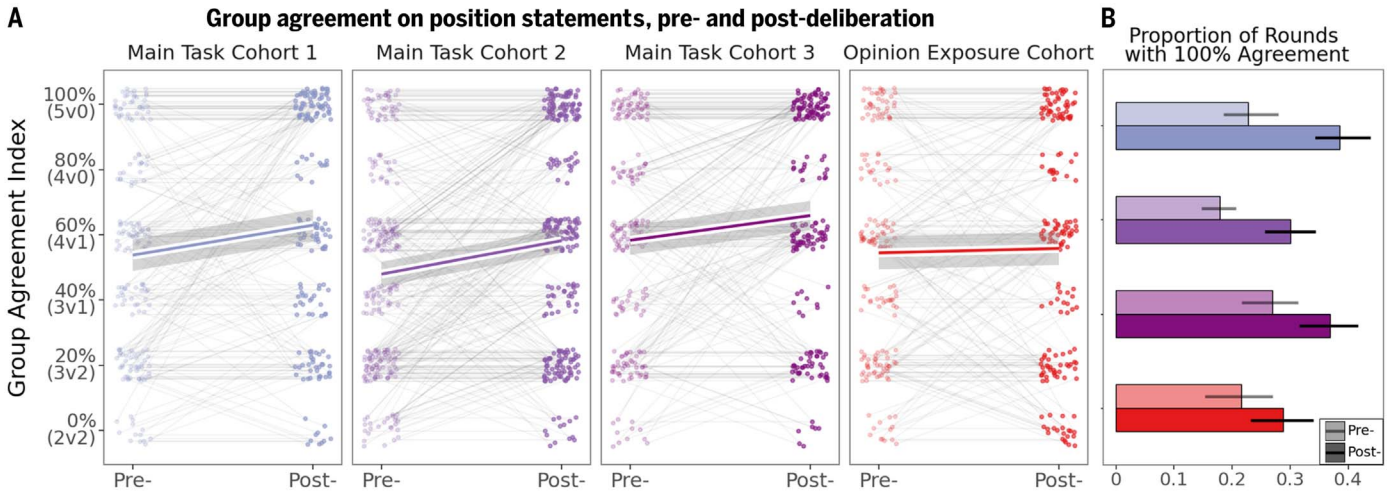


Fig. 3. AI-mediated deliberation leaves groups less divided. (**A**) Group agreement on the position statements increased from pre- to postdeliberation when interacting with the Habermas Machine (main task cohorts; blue to purple) but not when participants viewed each other's opinions unmediated (opinion exposure cohort; red). The *y* axis shows percentage-point difference between those who endorsed one versus the opposite position on an issue, as well as the canonical instantiation of that disagreement in terms of the numbers of participants [e.g., three versus two (3v2)]. Points represent individual groups discussing one question. Lines are fitted linear regression lines with standard errors. (**B**) The proportion of groups that achieved unanimous agreement increases from pre- to postdeliberation. In both panels, data are shown only for question rounds that had all five groups' members present pre- and postdeliberation. Error bars denote bootstrapped 95% CIs.

increased, decreased, or stayed the same from before to after interaction with the HM.

In all three main task cohorts, group agreement increased on average by about eight percentage points, indicating that participants converged on a common position on the issue within their group after AI-mediated deliberation [cohort 1: $\beta = 0.086$, SE = 0.019, $t(286) = 4.5$, $P < 0.001$; cohort 2: $\beta = 0.093$, SE = 0.015, $t(499) = 6.2$, $P < 0.001$; cohort 3: $\beta = 0.059$, SE = 0.019, $t(305) = 3.1$, $P = 0.0023$] (Fig. 3A). The number of groups that achieved unanimous agreement (i.e., all members supported the same side of the argument; group agreement = 100) increased from 22.8 to 38.6% in main task cohort 1, with similar shifts in the other two cohorts (Fig. 3B). We additionally found that the more a participant endorsed the final group statement, the more that participant's postdeliberation position ratings were influenced by the predeliberation ratings of the others in their group (interaction between group statement endorsement and influence of others' initial position ratings; main task cohort 1: $\beta = 0.24$, SE = 0.04, z score = 5.0, $P < 0.001$, other cohorts similar; main effect of own predeliberation position rating also significant; SM 4.4.7), suggesting that finding a good statement of common ground is what allowed people to shift their perspective. Consistent with this effect, on the surveys taken at the end of each deliberation round, participants reported that their view had moved "closer to the group opinion" on 30.1% of the deliberation rounds (60.3% reported no change, and 4.6% claimed it had moved in "some other way"), and on 66.6% of rounds, they reported that the best expression of their current perspective was the "final group opinion" (versus 33.3% who reported it was their "own original opinion"; full results provided in SM 4.4.6). In summary, although the HM was trained only to produce group opinion statements that maximized endorsement, the deliberation process reduced group division at least partly through convergent changes in participants' perspectives.

HM group statements will include the perspectives of other participants, raising the possibility that the reduction in division results from exposure to others' opinions rather than from successful mediation per se. To distinguish these possibilities, we recruited an additional cohort (opinion exposure cohort, $n = 372$ participants in 75 groups) who performed a comparable experiment; but rather than rating group statements for each issue, they rated each other's opinions. This provided a baseline condition that involved collective viewpoint exchange but no mediation. Here, unlike for the main task cohorts, we found that group agreement did not change during the deliberation process [$\beta = -0.005$, SE = 0.01, $t(230) = -0.46$, $P = 0.64$]. Moreover, across cohorts, there was a significant interaction between group agreement change and exposure type [individual opinions versus HM group statements; $\beta = -0.038$, SE = 0.014, $t(1491) = -2.76$, $P = 0.0058$], indicating that AI mediation was more successful at reducing disagreements than was mere exposure to the views of others (Fig. 3A). An exploratory analysis found similar viewpoint convergence resulting from the human mediator on rounds in which they beat the HM (in the human mediator task), suggesting that convergence may result from the caucus mediation process per se rather than from distinct properties of the AI mediator (SM 4.3 and Discussion).

RQ3: Does the Habermas Machine represent all viewpoints equally?

An important issue for any mediator is how to arbitrate between majority and minority views. Roughly 77% of deliberation rounds (803/1047) in our main task cohorts exhibited initial disagreements (e.g., only three out of five group members "agreed" with the position statement), allowing us to independently quantify change in support for the majority and minority positions. We found that support increased (from pre- to postdeliberation) for majority positions on some issues and for minority positions on others, implying that group statements did not simply flatter the majority but also allowed minority positions to flourish. The movement in position, however, was not symmetric: The size of the initial majority increased in 29 to 32% of rounds, whereas the minority increased in 20 to 26% of rounds. Moreover, participants' "majority-aligned" position ratings (position ratings that were recoded in terms of endorsement of the majority position; SM 4.1.2.1) increased from pre- to postdeliberation (main task cohort 1: $\beta = 0.52$, SE = 0.08, z score = 6.5, $P < 0.001$; other cohorts were similar). This implies that viewpoints were more likely overall to shift toward a majority view, even while minority voices continued to be heard. This movement, which is distinct from viewpoint convergence, is sometimes called "parallel belief updating" (*28*).

Which properties of the HM allowed for both minority and majority views to be expressed?

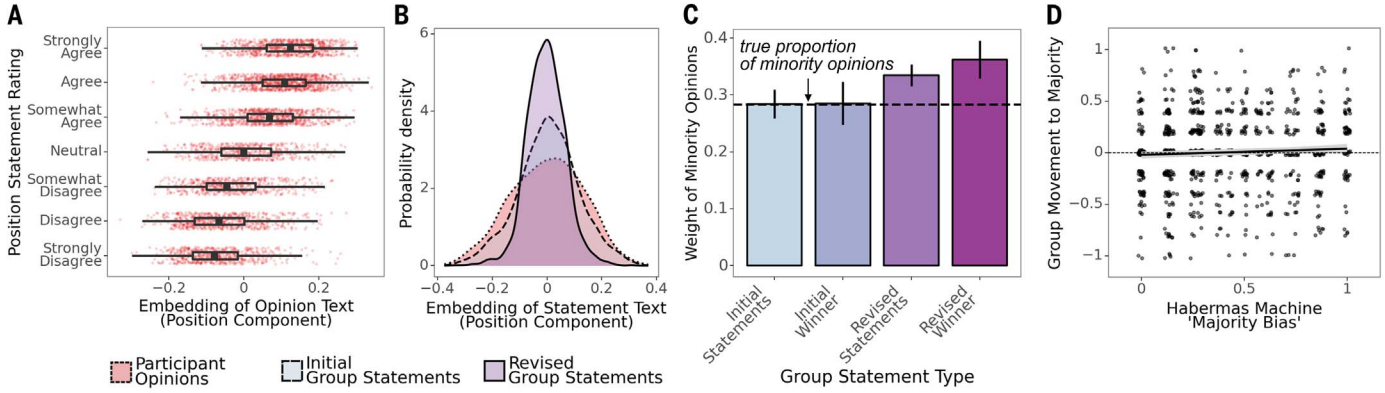


Fig. 4. The Habermas Machine's statements are fair compromises and avoid the "tyranny of the majority." (**A**) The position embeddings of participants' opinion text correlate strongly with participants' predeliberation position statement ratings, indicating that the dimension in text space relates to participants' positions on the issues. Each point represents an individual participant's position Likert rating and their opinion embedding in response to a question. (**B**) Distributional shifts in the position embeddings suggest that as the process proceeds from opinions (red) to initial group statements (blue) to revised group statements (purple), the group statements increasingly represent a compromise between opinions along the position axis. Distributions indicate marginal kernel density estimates of opinion and group statement embeddings along the position axis. (**C**) Initial group statements proportionately represent minority opinions, and revised statements overweight them relative to the true proportion of minority opinions (dotted line). Error bars indicate ± 1 SE for the estimated regression coefficients. (**D**) There is no evidence for an association between the tendency for the Habermas Machine to produce majority-leaning group statements and groups moving in the direction of the majority. Individual points represent a single group discussing a single question.

Using a T5 sentence encoder (*29*) (a machine learning tool for embedding text as a high-dimensional vector), we can study the relative weight that the HM gave to minority and majority opinions in the group statement. To do this, we derived a "position component score" by rotating the 768-dimensional embedding vectors (for individual opinions and group statements) onto the axis that connected the "agree" and "disagree" stances for each question (SM, materials and methods). Reassuringly, position component scores for participants' individual opinions were strongly correlated with corresponding predeliberation position statement ratings [correlation coefficient (r) = 0.64, coefficient of determination (r^2) = 0.41; Fig. 4A], indicating that the approach is meaningfully capturing information about the stance expressed in a paragraph of text. In this embedding dimension, the group statements represented a compromise among the opinions of the group, with 96% of scores for group statements falling within the range spanned by the corresponding individual opinions (Fig. 4B).

Do the group statements fairly represent the views of group members in the minority and majority? Modeling the group statement scores as convex combinations of majority and minority opinion scores within each group, we found that the initial group statements weighed minority opinions exactly in proportion to their empirical frequency (average size of minority = 29%; empirical average minority weight = 0.29). The revised group statements, however, tended to overweight the minority (empirical minority weight for revised statements $\beta = 0.36$, SE = 0.03, t = 2.64; Fig. 4C). This pattern suggests that the HM's initial statement appeals to the majority, which prompts critiques from the minority, resulting in revised statements that align more closely with the minority. However, rather than oscillating between majority and minority views, the earlier results (Fig. 2C) showing a preference for the revised statements over the initial ones suggest that the HM is finding meaningful common ground.

Lastly, discussants might have gravitated toward the majority view simply because they were asked to judge several group statements that supported that position. To test this, we measured the relationship between participants' viewpoint change toward the majority position (from pre- to postdeliberation position ratings) and the fraction of group statements whose position component scores fell on the majority side of the median opinion ("majority bias," SM, materials and methods). We found no relationship between fractional exposure to majority views and subsequent change in viewpoint toward the majority ($\beta = 0.058$, SE = 0.07, z score = 0.9, $P = 0.37$; Fig. 4D). This implies that any change toward the majority was due to the contents of the group statement and not to mere exposure to that view.

RQ4: Can AI-mediation support deliberation in a citizens' assembly?

The experiments reported so far involved a convenience sample of UK residents. This raises the question of whether AI-mediated deliberation would also be effective in a real-world sample that more closely reflects the diversity in the UK population. To address this, we partnered with the Sortition Foundation, a nonprofit organization experienced in the running of citizens' assemblies, to recruit a sample of 200 participants who were representative of the UK population across the demographic dimensions of age, gender identity, ethnicity, socioeconomic status, and geographic region (demographic summaries for all cohorts can be found in SM 4.8). Participants took part in a virtual citizens' assembly that occurred over three weekly 1-hour sessions in April and May 2023; the virtual citizens' assembly primarily mimicked the "deliberation phase" of a citizens' assembly, without the typical fact-finding or expert testimony phases. During each assembly, participants deliberated over questions in the same procedure as the main task cohorts. In comparison to the cohorts with crowd-sourced participants, we found that levels of endorsement for the group opinions were similarly high, and we found the same significant increase in group agreement from pre- to postdeliberation (Fig. 5, A and B, and SM 4.7.2).

We designed the virtual citizens' assembly such that all groups deliberated over the same nine questions (three per session), covering the potentially divisive issues of immigration, the retirement age, the prison population, Brexit, climate change, universal childcare, the minimum wage, national pride, and the value of the internet (these were identified as topics that participants rated as highly important in the training set; SM 4.7.1). This design offers a singular opportunity to examine whether AI-mediated deliberation led to convergent shifts in opinion across groups; i.e., whether, after debating freely, participants would tend to arrive at a common stance (as Habermas originally

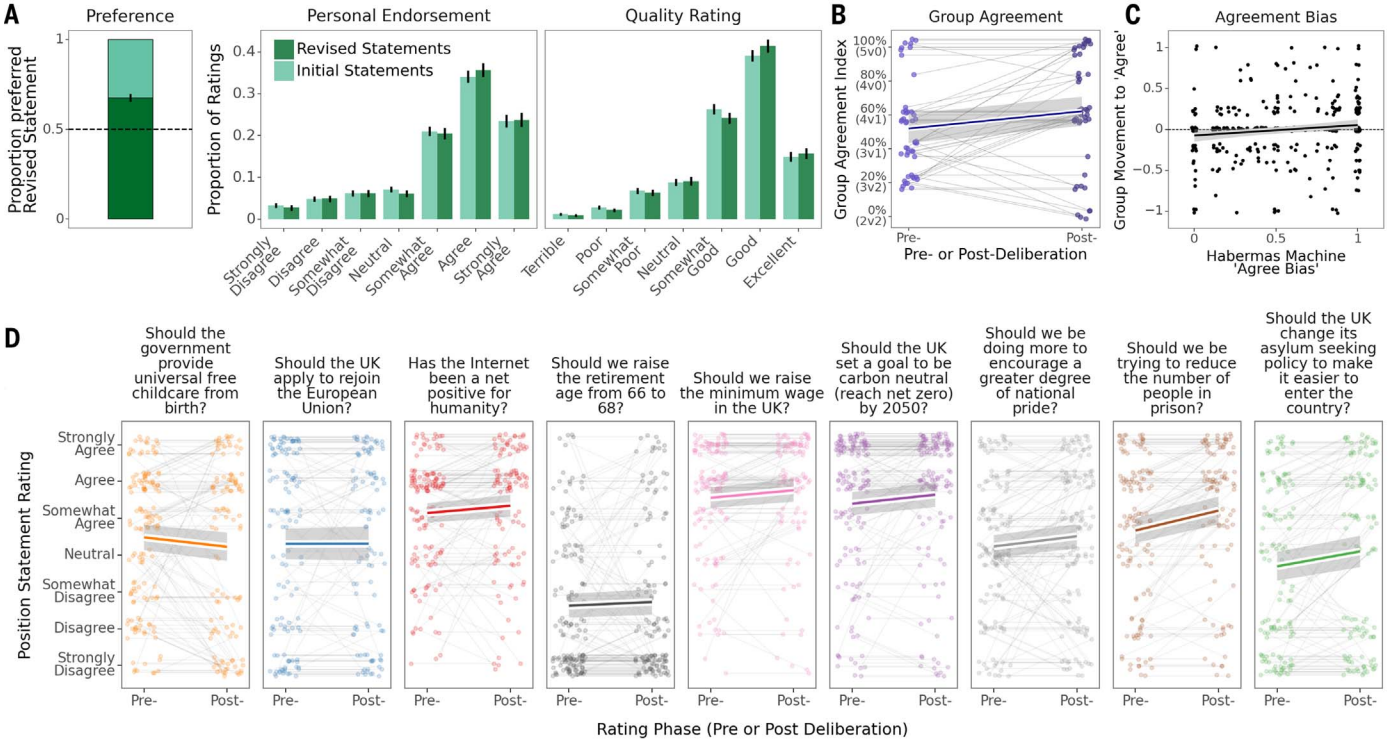


Fig. 5. The Habermas Machine helps a real-world virtual citizens' assembly find common ground on potentially contentious issues. (**A**) Revised statements (postcritique) were preferred over the initial statements in the final preference judgement, and group statements were positively endorsed and high quality. Error bars denote bootstrapped 95% CIs. (**B**) Group agreement index increased from before to after the deliberation. Points represent individual groups discussing a single question. (**C**) There is no evidence for an association between groups moving in the direction of endorsing the affirming position statement and the tendency for the Habermas Machine to produce group statements more on the "affirming side" of the median-position embedding of the group's opinions. Individual points represent a single group discussing a question. (**D**) Individual position statement ratings reveal that several questions showed movement in a consistent direction, from pre to postdeliberation. Individual points represent a single participant rating a question; lines are fitted linear regression lines with standard errors.

proposed). In exploratory analyses, we found some evidence for convergent shifts for five of the nine questions (Fig. 5D). Specifically, after deliberation, participants tended to shift their ratings toward stronger agreement that "we should be trying to reduce the number of people in prison" [$\beta = 0.58$, SE = 0.13, $t(883) = 4.1$, $P < 0.001$], with the fraction of participants rating on the "agree side" of the scale increasing from 60 to 75%, and that "the UK should change its asylum seeking policy to make it easier to enter the country" [$\beta = 0.51$, SE = 0.13, $t(886) = 3.8$, $P < 0.001$], with the fraction supporting increasing from 39 to 51%. We found more marginal evidence for a shift toward endorsing position statements in favor of encouraging national pride (increased support from 42 to 57%), raising the minimum wage (77 to 85%), and reaching net zero (75 to 80%) (P values for rating changes < 0.05). For issues where one would expect more entrenchment (such as Brexit), we saw no change at all, and one issue—universal free childcare from birth—showed a shift toward the opposing position, with opposition increasing from 33 to 41%, although the rating movement was not statistically significant [$\beta = -0.19$, SE = 0.13, $t(882) = -1.42$, $P < 0.156$] (see SM 4.7.7 for full results).

We also tested whether convergent viewpoint shifts across groups might be due to biases in the LLM that arose during pretraining or fine-tuning (*30*). Building on the approach described in RQ3 (Fig. 4D), we computed a bias score representing the fraction of group statements whose position component scores fell on the "agree" side of the median opinion ("agree bias" SM, materials and methods). We saw no evidence that the HM's tendency to produce statements overly supporting a position predicted the proportion of people in a group who changed their view to that position [$\beta = 0.11$, SE = 0.07, $t(124) = 1.44$, $P = 0.152$; Fig. 5C]. This argues against the claim that LLM bias drives the consistent change in belief observed at the group level in our citizens' assembly.

Discussion

This research project evaluates a new, AI-based approach to human collective deliberation, which involves using an AI system as a "caucus mediator." We show that group statements produced by an AI mediator [the Habermas Machine (HM)] won broad-based agreement from participants—a key step for finding common ground—and were preferred to those written by human mediators. After interacting with the HM, groups were often less divided, converging to a common stance on social and political issues. Using a language model allowed us to quantify the features of successful group statements. We found that the AI learned to respect the majority stance but also to upweight dissenting views. We replicated these findings in a virtual citizens' assembly, which involved a demographically representative sample of participants from across the UK rather than those conveniently found on a crowd-sourcing platform. These results considerably extend earlier work using more hand-engineered artificial systems that can debate with humans (*31*), summarize opinionated arguments in natural language (*32*), or support deliberation but without any motivation of finding common ground (*14*–*16*); but to our knowledge, they constitute the first demonstration that AI can successfully be used to mediate human collective deliberation at scale.

AI-mediated group statements were endorsed at higher rates than those written by incentivized

but nonprofessional human mediators. However, the key translational opportunity provided by the HM is not its potentially "superhuman" mediation but rather its ability to facilitate collective deliberation that is time-efficient, fair, and scalable (*5*). The AI system produced high-quality group statements within seconds, in contrast to human mediators, who required several minutes (time efficiency). The HM selects a group statement from a set of candidates by simulating a ranked-choice election in which each participant's vote has equal weight (fairness). Further, empirically, we found that the group statement incorporated the critiques of the minority (while respecting the majority view), thus avoiding the "tyranny" of the majority (*33*). Lastly, although we only tested groups of up to five participants, our methods in theory scale to groups in which hundreds of people deliberate collectively, when combined with more modern, longer-context LLMs [e.g., Gemini 1.5 (*34*) could fit a thousand opinions into its context window] (scalability). When we compared our smaller (Chinchilla-based) fine-tuned model with a prompted version of the larger and more modern Gemini 1.5 Pro, we found evidence that the former was more performant, suggesting that combining bespoke fine-tuning with a larger model would improve our results yet further (SM 3.5). The HM thus offers a new approach to collective deliberation that circumvents some of the limitations of in-person deliberation, including its cost, limited scale, the potential for mediator bias, and proneness to social desirability effects or inequality of contribution (*6*, *7*). Nevertheless, the caucus mediation approach may miss out on other advantages that arise from in-person discussion, including nonverbal cues and the opportunity to build interpersonal relationships with other discussants.

A potential limitation of our study is the reliance on UK participants discussing nationally relevant political and social issues. We do not know whether our findings would generalize to other groups, but we have no reason to believe that they stem from distinctive features of the UK context. Further, although many of our participants were sampled by convenience, we obtained comparable results in a virtual citizens' assembly (conducted with the Sortition Foundation) that used a demographically representative sample of the UK population, and we found similar results using a propensity-weighting approach to account for potential demographic imbalance in our convenience-sampled cohorts (SM 4.3.5). We thus think it is likely that our approach could generalize to different groups and contexts.

One may question whether finding agreement is a desirable objective. The group statements that the HM produces need not express a singular view but can reflect a wide distribution of opinions discussed. Users may endorse the group statement because they believe that it represents the outcome of a fair deliberation process (e.g., reflecting a form of "settled disagreement") rather than because it aligns entirely with their view. That is, users may be finding "common ground" by agreeing to shared information that can be used to inform later proposals or outcomes (*19*). Even after deliberation, a plurality of views may persist (*35*, *36*), highlighting the nuanced nature of consensus in deliberative processes (see SM 7 for further discussion).

It is important to acknowledge that the HM, in its current form, is limited in its capacity to handle certain aspects of real-world deliberation. For example, the HM does not have the mediation-relevant capacities of fact-checking, staying on topic, or moderating the discourse. If the human opinions are ill-informed or harmful, then the HM may generate an ill-informed or harmful output. This feature is shared with other democratic processes and mirrors the principle on which citizens' assemblies are based: that legitimate agreement can emerge from free and fair deliberation among citizens. Still, deliberative democratic processes involve other mitigations, such as incorporating expert testimony in a manner that does not predetermine the outcome of the debate (*37*, *38*). In this project, we offered one such mitigation, using questions designed to reduce the risk of harmful or unproductive deliberations (SM 1). Nevertheless, if used in the real world, the HM ought to be embedded in a larger deliberative process, including careful selection of participants to ensure that a balanced and diverse community of stakeholders is represented in the debate.

Our analyses did verify that the group statements were fair reflections of the majority view and not politically slanted by putative biases in the AI model. The HM did not appear to "tyrannically" ignore the views of the minority (*33*) but reliably upweighted dissenting voices in the AI-generated group statements. Still, vigilance is needed to ensure that any AI-assisted deliberative process is fair and legitimate (*39*). Moreover, there is room for healthy debate over the role that algorithms should play in the political process, and some people may have a principled aversion to the idea that political or social ideas can be generated by a computer (*40*, *41*), which may temper enthusiasm for our approach.

There are considerable benefits to a technology that helps people find agreement in a time-efficient, fair, and scalable manner. Many real-world situations require groups of people to agree over the content of a written statement. These include, but are not limited to, contract agreement, conflict resolution, jury deliberation, diplomatic negotiations, constitutional conventions, artistic co-creation, and political or legislative discussions, as well as formal citizens' assemblies. More generally, finding common ground is a precursor to collective action, in which people work together for the common good. Thus, there is considerable potential for the HM to be translated for societal benefit.

Materials and methods summary

Questions

Participants discussed questions of social or political significance in the UK. We constructed this dataset by prompting the Chinchilla 70B model to generate 10,000 questions, using examples drawn from 175 author-written seed questions (SM 10). Questions were reviewed for safety, ethics, clarity, and potential to stimulate diverse discussion (SM 8; see SM 1 for more details).

Human data

For the virtual citizens' assembly, participants were recruited with randomized stratified sampling (through the Sortition Foundation) to create a cohort that was demographically representative of the UK population with respect to age, gender, income, ethnicity, and region (SM 4.8. provides demographic summaries). Other participants were adult UK residents recruited by convenience through a crowdsourcing platform. Both training (n = 2473) and evaluation data (n = 2673) were collected between January and August 2023. Participants were compensated above the living wage, and the study was approved by our internal Human Behavioral Research Ethics Committee (*42*) (SM 2.1). Additional bonuses were offered in the human mediator task for the human mediator, up to a roughly 50% increase in base pay (SM 4.3).

We ran seven evaluation experiments, including

- Main task/ablation experiments (1 to 3): Assessing the importance of different HM components.
- Critique exclusion experiment: Evaluating the impact of the critique phase.
- Opinion exposure experiment: Examining the impact of mediation on belief change.
- Human mediator experiment: Comparing HM performance to human mediators.
- Virtual citizen's assembly: Testing external validity with a demographically representative sample.

Each study and analysis is described in SM 4. Sample sizes, exclusion criteria, and primary analyses for all evaluation experiments were preregistered at osf.io/uy4z3.

Human task

After providing informed consent and reading instructions, participants were organized into groups of up to five individuals. The task consisted of three phases:

1. Predeliberation: Initial position statements and question importance were rated.

2. Deliberation blocks (1 question per block, $n = 3$ blocks). Each involved eight stages: writing opinions, rating and ranking group statements, critiquing the top-ranked statement, rating ranking revised group statements, expressing final preferences, and completing an end-of-round survey (Fig. 1A).

3. Postdeliberation: Position statements and question importance were rated again.

The experiment concluded with a survey collecting free-form comments and a demographic questionnaire (see SM 2.2 for more details). The human mediator and opinion exposure tasks excluded the critique phase. The human mediator task involved four deliberation rounds and included third-party raters to assess transcript quality.

HM details

The generative model is a Chinchilla 70B language model (*21*) fine-tuned on a dataset of opinions (and critiques) from previous rounds with target group statements that were rated as high-quality, covering 578 groups (2473 participants) collectively discussing 1630 deliberation questions across diverse topics. Questions were clustered into 105 clusters, using k-means applied to the vector embeddings derived using the Universal Sentence Encoder (*43*). Out-of-distribution generalization was tested using 12 topics held out during fine-tuning (SM 4.4).

The reward model is a smaller, 1.4 billion–parameter version of Chinchilla with an additional linear layer for scalar reward prediction. We train this model on a dataset in which participants' rankings over the statements are converted to pairwise preferences, fitting a Bradley-Terry model yielding rewards proportional to how much each participant is expected to agree with a candidate group statement (*22*, *44*, *45*). The reward model training also includes an auxiliary classification loss based on binarized quality ratings (see SM 3.1 for prompts and SM 3.3 for further details).

The HM uses a simulated election process to select an output statement. Sixteen candidate statements are generated, and the reward model predicts each participant's most likely ranking of the statements based on the participant's opinion (and critique). We aggregate these rankings using the Schulze method to determine the winning group statement (*22*, *24*). The group winner is returned as a sample from the HM. In the main evaluation experiments, participants see multiple samples from the HM, which involves repeating this procedure multiple times (SM 3.4).

Data preprocessing and analysis

Data analysis followed preregistered procedures for data exclusion (SM 4.1 provides data analysis details). Inferential statistics were based on generalized linear mixed-effects models, implemented in *R* (*46*, *47*), adopting the maximal random-effects structure that our data permits in each case (*48*). We fit ordinal models for analyses relating the Likert measurements (*49*). All P values reported are from two-sided inferential tests (SM 4.1.3).

Embedding geometry

We embed opinions, group statements, and affirming and negating position statements for each question (e.g., "Yes, I agree. We should raise taxes." and "No, I disagree. We should not raise taxes.") using the Sentence T5 model (*29*). The line segment from the negating to affirming embeddings defines a "position axis." The projection onto this position axis defines, for any piece of text (i.e., an opinion or group statement), a decomposition into a "position" embedding component and "residual" embedding components. Alternative embedding models and definitions for position axes are possible [(*43*, *50*); SM 5.1].

Minority (versus nonminority) participants are identified by which side of neutral their position statement ratings fall on. Position embeddings for group statements are regressed on convex combinations of the constituent opinions. Regression coefficients of minority opinions are summed to produce a total minority weight. This process is repeated separately for the level of division of groups (e.g., five total participants with one in minority, five total participants with two in minority, etc.), and we report the final average weight of minority opinions (see SM 5.5 for more details).

The HM's "majority" bias is measured by determining whether its generated statements align with the majority or minority side of the median opinion for each question. The bias score is the proportion of statements leaning toward the majority view. Similarly, the HM's "agree" bias in the virtual citizen's assembly is assessed by comparing the statement's position to the "agree" and "disagree" sides relative to the median opinion. Further analyses and results relating to compromise, fairness, and novelty can be found in SM 5.

REFERENCES AND NOTES

1. R. Huckfeldt, P. E. Johnson, J. Sprague, *Political Disagreement: The Survival of Diverse Opinions Within Communication Networks* (Cambridge Univ. Press, 2004). doi: 10.1017/CBO9780511617102
2. J. S. Fishkin, *Democracy and Deliberation: New Directions for Democratic Reform* (Yale Univ. Press, 1993).
3. H. Landemore, *Open Democracy: Reinventing Popular Rule for the Twenty-First Century* (Princeton Univ. Press, 2020).
4. M. Reuchamps, J. Vrydagh, Y. Welp, Eds. *De Gruyter Handbook of Citizens' Assemblies* (De Gruyter, 2023). doi: 10.1515/9783110758269
5. C. Lafont, *Democracy Without Shortcuts: A Participatory Conception of Deliberative Democracy* (Oxford Univ. Press, 2020).
6. D. M. Kahan, Ideology, motivated reasoning, and cognitive reflection. *Judgm. Decis. Mak.* **8**, 407–424 (2013). doi: 10.1017/S1930297500005271
7. T. Sharot, M. Rollwage, C. R. Sunstein, S. M. Fleming, Why and When Beliefs Change. *Perspect. Psychol. Sci.* **18**, 142–151 (2023). doi: 10.1177/17456916221082967; pmid: 35939828
8. H. Mercier, H. Landemore, Reasoning is for arguing: Understanding the successes and failures of deliberation. *Polit. Psychol.* **33**, 243–258 (2012). doi: 10.1111/j.1467-9221.2012.00873.x
9. H. Landemore, "Can AI bring deliberative democracy to the masses?" in *Conversations in Philosophy, Law, and Politics*, R. Chang and A. Srinivasan, Eds. (Oxford Univ. Press, 2023), pp. 39–69.
10. C. T. Small *et al.*, Opportunities and Risks for LLMs for Scalable Deliberation with Polis. arxiv.2306.11932 [cs.SI] (2023).
11. S. Fish *et al.*, Generative Social Choice. arXiv.2309.01291 [cs.GT] (2023).
12. L. P. Argyle *et al.*, Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2311627120 (2023). doi: 10.1073/pnas.2311627120; pmid: 37788311
13. R. Hadfi *et al.*, Conversational agents enhance women's contribution in online debates. *Sci. Rep.* **13**, 14534 (2023). doi: 10.1038/s41598-023-41703-3; pmid: 37666917
14. S. Kim, J. Eun, C. Oh, B. Suh, J. Lee, "Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot," *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, 25 to 30 April 2020 (Association for Computing Machinery, 2020), pp. 1–13. doi: 10.1145/3313831.3376785
15. J. Shin, M. A. Hedderich, A. Lucero, A. Oulasvirta, "Chatbots Facilitating Consensus-Building in Asynchronous Co-Design," *UIST '22: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, Bend, OR, 29 October to 2 November 2022 (ACM, 2022), pp. 1–13.
16. S. Ma *et al.*, Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. arXiv.2403.16812 [cs.HC] (2024). doi: 10.2139/ssrn.4772689
17. C. W. Moore, The caucus: Private meetings that promote settlement. *Mediation Q.* **1987**, 87–101 (1987). doi: 10.1002/crq.39019871611
18. R. Stalnaker, Common ground. *Linguist. Philos.* **25**, 701–721 (2002). doi: 10.1023/A:1020867916902
19. L. Morrissey, J. Boswell, Finding Common Ground. *Eur. J. Polit. Theory* **22**, 141–160 (2023). doi: 10.1177/1474885120969920
20. J. Habermas, *Theory of Communicative Action, Volume One: Reason and the Rationalization of Society* (Beacon Press, 1981).
21. J. Hoffmann *et al.*, Training Compute-Optimal Large Language Models. arXiv.2203.15556 [cs.CL] (2022).
22. P. F. Christiano *et al.*, "Deep reinforcement learning from human preferences". *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* Long Beach, CA, 4 to 9 December 2017 (NeurIPS Foundation, 2017), 4299–4307.
23. A. Askell *et al.*, A general language assistant as a laboratory for alignment. arXiv:2112.00861 [cs.CL] (2021).
24. R. Thoppilan *et al.*, LaMDA: Language Models for Dialog Applications. arXiv.2201.08239 [cs.CL] (2022).
25. M. Schulze, A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Soc. Choice Welfare* **36**, 267–303 (2011). doi: 10.1007/s00355-010-0475-4
26. P. Singhal, T. Goyal, J. Xu, G. Durrett, A Long Way to Go: Investigating Length Correlations in RLHF. arXiv.2310.03716 [cs.CL] (2023).
27. E. Aronson, J. Mills, The effect of severity of initiation on liking for a group. *J. Abnorm. Soc. Psychol.* **59**, 177–181 (1959). doi: 10.1037/h0047195
28. A. Jern, K. M. K. Chang, C. Kemp, Belief polarization is not always irrational. *Psychol. Rev.* **121**, 206–224 (2014). doi: 10.1037/a0035941; pmid: 24730598
29. J. Ni *et al.*, Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. arXiv.2108.08877 [cs.CL] (2021).
30. S. Feng, C. Y. Park, Y. Liu, Y. Tsvetkov, "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 9 to 14 July 2023 (Association for Computational Linguistics, 2023), pp. 11737–11762.
31. N. Slonim *et al.*, An autonomous debating system. *Nature* **591**, 379–384 (2021). doi: 10.1038/s41586-021-03215-w; pmid: 33731946

32. Y. Suhara, X. Wang, S. Angelidis, W.-C. Tan, "OpinionDigest: A Simple Framework for Opinion Summarization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5 to 10 July 2020 (ACL, 2020), pp. 5789–5798. doi: 10.18653/v1/2020.acl-main.513
33. A. de Tocqueville, *Democracy in America*, H. Reeve, Transl. (Saunders and Otley, 1835).
34. Gemini Team Google, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv.2403.05530 [cs.CL] (2024).
35. C. Mouffe, Deliberative democracy or agonistic pluralism. *Soc. Res.* **66**, 745–758 (1999).
36. I. Shapiro, Collusion in Restraint of Democracy: Against Political Deliberation. *Daedalus* **146**, 77–84 (2017). doi: 10.1162/DAED_a_00448
37. K. Hobson, S. Niemeyer, "What sceptics believe": The effects of information and deliberation on climate change scepticism. *Public Underst. Sci.* **22**, 396–412 (2013). doi: 10.1177/0963662511430459; pmid: 23833106
38. S. Müller, G. Kennedy, T. Maher, Reactions to experts in deliberative democracy: The 2016-2018 Irish Citizens' Assembly. *Ir. Polit. Stud.* **38**, 467–488 (2023). doi: 10.1080/07907184.2023.2211014
39. A. Birhane *et al*., "The forgotten margins of AI ethics," *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 21 to 24 June 2022 (ACM, 2022), pp. 948–958. doi: 10.1145/3531146.3533177
40. R. M. Dawes, The robust beauty of improper linear models in decision making. *Am. Psychol.* **34**, 571–582 (1979). doi: 10.1037/0003-066X.34.7.571
41. B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015). doi: 10.1037/xge0000033; pmid: 25401381
42. A. Paterson, W. Hawkins, "Best practices for data enrichment" (2022); https://deepmind.google/discover/blog/best-practices-for-data-enrichment.
43. D. Cer *et al*., Universal sentence encoder. arXiv.1803.11175 [cs.CL] (2018).
44. R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952). doi: 10.1093/biomet/39.3-4.324
45. D. M. Ziegler *et al*., Fine-Tuning Language Models from Human Preferences. arXiv.1803.11175 [cs.CL] (2019).
46. D. M. Bates, "lme4: Mixed-effects modeling with R" (2010); https://people.math.ethz.ch/~maechler/MEMo-pages/lMMwR.pdf.
47. R. H. B. Christensen, "ordinal: Regression models for ordinal data," R package version 12-4.1 (2023); https://cran.r-project.org/web/packages/ordinal/index.html.
48. D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013). doi: 10.1016/j.jml.2012.11.001; pmid: 24403724
49. P.-C. Bürkner, M. Vuorre, Ordinal regression models in psychology: A tutorial. *Adv. Methods Pract. Psychol. Sci.* **2**, 77–101 (2019). doi: 10.1177/2515245918823199
50. Z. Yang, Y. Yang, D. Cer, J. Law, E. Darve, "Universal sentence representation learning with conditional masked language model," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7 to 11 November 2021 (ACL, 2021), pp. 6216–6228.
51. M. H. Tessler, Fine-tuning LMs for consensus: Evaluation of fine-tuning and iteration #1, OSF (2024); https://doi.org/10.17605/OSF.IO/QUSHN.
52. M. H. Tessler, Fine-tuning LMs for consensus (Model Comparison 2), OSF (2024); https://doi.org/10.17605/OSF.IO/QH6YR.
53. M. H. Tessler, Fine-tuning LMs for consensus (Model comparison - OOD generalization), OSF (2024); https://doi.org/10.17605/OSF.IO/YQABX.
54. M. H. Tessler, Fine-tuning LMs for consensus (Human Consensus Writer eval), OSF (2024); https://doi.org/10.17605/OSF.IO/9UYZ8.
55. M. H. Tessler, Fine-tuning LMs for consensus (Critique Exclusion), OSF (2024); https://doi.org/10.17605/OSF.IO/4ZJEU.
56. M. H. Tessler *et al*., google-deepmind/habermas_machine: Habermas Machine v0, Zenodo (2024); https://www.doi.org/10.5281/zenodo.13821139.

ACKNOWLEDGMENTS

The authors thank the following individuals for helpful conversations, constructive feedback, and technical assistance throughout the course of this project: I. Gabriel, S. Bridgers, W. Hawkins, S. El-Sayed, S. Brown, L. A. Hendricks, K. McKee, A. Friend, R. Rippin, N. Gill, B. Hennig, J. Balaguer, D. Williams, J. Sanchez Elias, F. Fischer, Y. Doron, A. Stjerngren, J. Aslanides, M. Glaese, N. McAleese, M. Staib, R. Faulkner, D. Kasenberg, Y. Jiao, J. Godwin, N. Fernando, A. Ahmad, A. Belias, S. Choudhry, R. Patel, S. Onike, P. Kirk, A. Paterson, N. Marchal, B. Wu, Z. Darme, E. Grau, W. Isaac, H. Law, L. Weidinger, J. Uesato, T. Liechty, N. Shalbak, T. Terwiliger, K. Tuyls, V. Reiser, D. Hassabis, and A. Dragan. **Author contributions:** Conceptualization: M.H.T., M.A.B., H.S., M.J.C., M.B., and C.S. Methodology: M.H.T., M.A.B., H.S., D.J., M.J.C., R.K., G.E., D.C.P., T.C., and C.S. Investigation: M.H.T., M.A.B., and C.S. Visualization: M.H.T., M.A.B., and D.J. Funding acquisition: n/a Project administration: M.H.T., M.A.B., M.J.C., L.C.-G, and C.S. Supervision: C.S. and M.B. Writing – original draft: M.H.T., M.A.B, D.J., G.E., M.B., and C.S. Writing – review and editing: M.H.T., M.A.B., G.E., M.B., and C.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Sample sizes, exclusion criteria, and primary analyses for all evaluation experiments were preregistered at osf.io/uy4z3/registrations (*51*–*55*). The dataset of questions and the human feedback data used for training and evaluating the Habermas Machine are publicly available on GitHub (https://github.com/google-deepmind/habermas_machine) and Zenodo (*56*). The full set of seed questions can be found in the supplementary materials, materials and methods.

SUPPLEMENTARY MATERIALS

Submitted 6 May 2024; accepted 23 September 2024
10.1126/science.adq2852