

The Civic Machine

Research and political reasoning



AI mediation, persuasion, agenda access and binding public decisions

Created by Yoav Duman

Start without advance reading

You can start here without reading a journal article beforehand. The two article explanations inside this assignment provide the research you need. Read them as part of the activity and return to them while answering. The original articles and further reading are available for optional follow-up.

Research words in everyday language

A claim is a conclusion you want to check. A hypothesis is a possible explanation. A mechanism is the sequence of steps through which something happens: who acts, what they do, and how that could produce a result.

A rival explanation is another plausible cause of that same result. A prediction says what you would expect to see if an explanation were right. Your job is to compare those predictions with evidence, not guess which opinion your instructor prefers.

An outcome is the result being measured. The population is the group you want to understand. The sample is the people actually studied. The unit of observation is who or what each record describes; a number's unit tells you whether it is a count, a percentage, hours, or score points.

The numerator is the number being counted. The denominator is the whole group used to calculate a percentage: 8 out of 20 means 8 divided by 20. A percentage-point difference subtracts one percentage from another.

An association means two things appear together in the data. It does not by itself tell you which caused which. Random assignment puts people into conditions by chance; a comparison then asks whether changing a feature changes the measured result.

A study's limits tell you what its evidence cannot establish. An uncertainty interval is a range describing how uncertain an estimate is under the analysis used. When an interval for a difference includes zero, it does not clearly distinguish a difference from no difference. That does not prove the true effect is exactly zero.

You do not need to memorize technical names or article statistics. Use the explanations, worked examples and records on screen. In your answer, make the reasoning visible: a specific detail, what it supports, and what remains uncertain.

The research question

These papers make a productive pairing because they study different jobs assigned to conversational AI. One investigates helping people formulate common ground; the other tests advocacy for an assigned candidate or ballot position. The question is how a public institution should evaluate a system whose output may be clear, persuasive, and widely approved. The studies do not directly test each other, and neither is an audit of our fictional city.

This is a tension between complementary findings, not a published dispute. A mediator tries to formulate something the participants can recognize together. An advocate tries to increase support for a specified outcome. Both can change views. The useful comparison concerns the assigned objective, the process, and the outcome measured—not whether every change of mind is either democratic or manipulative.

This packet contains the complete teaching explanations for the two core articles and links to the primary texts. It keeps article findings separate from invented local cases. Read the primary texts when checking a citation, design detail or contested inference.

Article A AI can help humans find common ground in democratic deliberation

Michael Henry Tessler, Michiel A. Bakker, and colleagues (2024). Science. 386(6719): eadq2852 · DOI 10.1126/science.adq2852

Dates

Published 18 October 2024. UK data collected in 2023; this is not a test of today's general chatbots.

Access

The complete Science publisher PDF supplied for this revision has been checked. Live publisher access may depend on an account or institution; the core explanation works without opening it.

Lens

Can people recognize shared priorities in a statement without surrendering every disagreement?

Read first

Start with the explanation on this screen. Then check what was compared, what was found, and what remains uncertain. The journal article link is optional.

Sections

Read the procedure and human comparison; RQ2 on viewpoint change; RQ3 on minority text and majority-aligned movement; then the Discussion's limitations.

Reading summary

A tested drafting/revision protocol. Compare endorsement, viewpoint movement and textual representation before using it to justify a public decision.

[Complete publisher text](#)

[Read the complete Science PDF](#)

Research explanation

In everyday terms: this study asks whether AI can help a group write a statement people recognize as common ground. Approving the statement is different from agreeing on a policy or giving someone the power to decide it.

Deliberation is an exchange of reasons about a shared decision. A mediator helps people understand positions and formulate something they can consider together. Common ground can concern a shared purpose, a conditional compromise or an accurate account of unresolved disagreement. It does not require every person to want the same policy.

Consider a separate, invented museum consultation. One visitor wants more late openings. Another wants quiet daytime access. Both might accept a statement asking the museum to widen access while protecting a quiet period, even though they disagree about the timetable. The statement contributes criteria for evaluating schedules; it does not itself choose one. To judge that contribution, compare the visitors' submissions with the actual words they endorsed.

A drafting procedure also has an objective. Predicting which text will win endorsement can help a group find language it recognizes, but endorsement is a particular criterion. It is not identical to factual accuracy, equal influence over a decision or protection against a severe cost. An institution must explain why it uses that criterion and what other requirements govern the eventual choice.

Tessler and colleagues' purpose-built system generated candidate statements from private opinions, selected using predicted preferences and revised after critiques. Participants preferred AI statements over statements from trained ordinary participants. The research also examined viewpoint movement and minority contributions in text. Those are related questions with different evidence: liking a statement, moving toward a stance and having a concern appear in wording are not interchangeable outcomes. Here, minority text weight means a model estimate of minority positions represented in the statement, not votes or power to nominate or adopt a policy.

The politically difficult finding is that minority contributions could remain visible while attitudes shifted more often toward the initial majority overall. The text-position analysis found proportionate minority weight in first statements and greater weight after revision. That measure concerns textual positions; it does not demonstrate that minorities had equal agenda power or that anyone was protected from a material loss. Learning from new reasons and accommodating the group can be rival accounts of an attitude shift. The reported results do not settle every mechanism.

Return to the museum example. A summary could faithfully include quiet access while only one schedule is allowed onto the decision agenda. A visitor may change their preference and still regard the rule as unfair, or lose the timetable vote while accepting the procedure. Ask separately about representation, influence, consequences and authority. The study can inform the design of a drafting step; the institution must defend its broader rule for a binding decision. In a separate invented example, a visitor may prefer Sunday opening yet accept a previously announced selection rule after a weekday schedule wins. Ask about the preferred schedule and acceptance of the rule separately.

Method

UK adults submitted opinions, rated candidate common-ground statements and provided critiques for revision. The system optimized predicted endorsement. The human comparison used trained, incentivized ordinary participants, not professional mediators.

Selection occurred at two levels. Inside the HM, predicted personal rankings selected a generated output. Participants then ranked the displayed outputs themselves to select the statement they would critique. The human-versus-AI comparison omitted the critique phase; separate experiments tested critique and revision.

Findings

AI statements won 56% of preferences against 44% for human statements. Revised texts were favored. Attitude movement leaned toward initial majorities overall; text-position analysis found minority contributions and increased minority weight after critique.

Limits

These tested drafting and deliberation outcomes do not establish factual verification, equal decision influence, universal fairness or municipal authority. Textual representation and attitude movement require separate interpretation.

Technical details

The research-program total was 5,734, not each experiment's sample; the human-mediator comparison involved 439 participants. Deliberation involved up to five discussants. In the human comparison, six-person groups assigned one member to mediate the other five. A representative virtual assembly recruited 200 people, who deliberated in small groups.

RQ3's text-position model estimated initial minority weight at 0.29, matching mean minority size, and revised weight at 0.36. This is an indirect aggregate text measure, not a guarantee for each concern or person. For an unrelated fictional coding example, counting quiet-hours and late-opening positions in a museum draft describes its text; it does not count which visitor can nominate a timetable or authorize its adoption.

The 56% versus 44% result concerns observed statement preferences in the human-mediator comparison, not 56% of all 5,734 participants. Do not read it as the effect of the complete iterative protocol. Separate critique-exclusion and revision experiments bear on the value of incorporating critiques.

Independent raters found differences favoring HM statements on clarity/informativeness, logic, majority capture and expression of a collective opinion. They did not find statistically significant differences on fair summary, minority inclusion or polarization. A detected difference on one criterion does not establish superiority on all criteria; an undetected difference does not establish equivalence.

The representative virtual assembly met in three weekly sessions in April and May 2023. It replicated the deliberation phase without the usual fact-finding and expert-testimony phases. The tested small groups do not establish the performance of a complete citizens' assembly or one meeting of all 200 participants.

Evidence receipts

Unrelated museum example · object of endorsement

Two visitors can accept shared access criteria while preferring different schedules. Identify the exact words accepted before inferring a policy choice.

Unrelated museum example · two measures

A concern appearing in a summary and a visitor changing their preferred schedule need different records. Neither measure substitutes for the other.

Unrelated museum example · affected interests

An equal vote among visitors and protection of an accessibility requirement are different decision criteria. State which rule applies and why.

Unrelated museum example · representation and authority

Visitors may amend a description without controlling the eligible schedules. Trace who drafts, nominates, decides and can appeal.

Article B Persuading voters using human-artificial intelligence dialogues

Hause Lin, Gabriela Czarnek, Benjamin Lewis, and colleagues (2025). Nature. 648: 394-401 · DOI 10.1038/s41586-025-09771-9

Dates

Published online 4 December 2025; issue 11 December 2025. Experiments involved elections and a ballot measure in 2024-25.

Access

Publisher article page; a complete free publisher PDF is available on a coauthor's MIT site.

Lens

A respectful conversation can still have an assigned directional objective. Inspect its instructions and the outcome it is meant to change.

Read first

Start with the explanation on this screen. Then check what was compared, what was found, and what remains uncertain. The journal article link is optional.

Sections

Read Effects of the human-AI dialogue; Testing mechanisms in Canada and Poland; and Discussion, including the field-deployment limits.

Reading summary

An experimental study of advocacy rather than consultation summarizing. Use it to ask whether the system's objective could change judgments under controlled comparisons.

[Complete publisher text](#)

[Free coauthor-hosted PDF](#)

Research explanation

In everyday terms: this study asks whether an AI conversation can move people toward a candidate or position it was told to support. A change in a survey answer is different from an observed vote.

Persuasion is an attempt to change judgment through reasons or appeals. It may be disclosed and legitimate. Directional advocacy has a specified target, such as support for one candidate. A common-ground drafting system seeks an acceptable group statement.

Both jobs can alter views; a view change alone does not tell us whether a process was truthful, representative or democratically authorized.

In the unrelated museum example, compare an assistant asked to represent visitors' reasons with one asked to increase support for late openings. Both might acknowledge concerns respectfully. The tone cannot reveal the objective. Inspect the commissioning instruction, material selected, alternatives discussed and response measured. A statement's accuracy and the visitor's preferred schedule remain separate questions.

Lin and colleagues used assigned advocacy in survey dialogues concerning candidates and a ballot measure. Their national experiments compared opposing advocates; they were not simply an advocate versus no conversation. The measured changes concerned attitudes and reported intentions. An attitude-scale movement is not observed voting, a vote-share change or an election result.

A before-and-after change can also reflect repeated questioning, unrelated information or selection into completing a conversation. A controlled comparison holds the defined outcome and relevant conditions comparable while varying the assigned feature. Random allocation helps address prior differences under its assumptions; it does not make every later generalization automatic.

The study's randomized strategy changes are especially useful for mechanism reasoning. Removing facts and evidence weakened persuasion in Canada and Poland; removing personalization in Poland did not clearly reduce it. An observed link between a strategy and attitude change is weaker causal evidence than deliberately varying that strategy. Some claims were inaccurate, but this is not a demonstration that falsehood caused the influence.

Use this research to pose a precise institutional question. Who chose the system's objective, who benefits from movement toward the target, and which outcome should a reviewer evaluate? Knowing that a system is AI or barring one kind of personal data does not settle those questions. Neither article chooses a legitimate museum timetable—or a legitimate rule for any other institution—for us.

Method

Lin and colleagues randomly assigned survey participants to candidate-advocating AI dialogues in the US, Canada, and Poland, and to conditions concerning a Massachusetts ballot measure. They measured attitudes and reported voting intentions.

Findings

Assigned advocacy changed candidate preferences. Removing facts and evidence reduced persuasion in Canada and Poland; removing personalization in Poland did not significantly reduce it. Some AI claims were inaccurate. Part of the US attitude change remained at follow-up.

Limits

These experiments establish influence under their survey conditions, not observed ballot changes or election outcomes. They do not show that all persuasion depends on personal data, that falsehood caused the effects, or that a common-ground mediator has the same objective.

Technical details

Completed initial samples: US 2,306; Massachusetts 501; Canada 1,530; Poland 2,118. Figure captions sometimes list starters instead. US follow-up: 1,936 participants after a mean of about 41 days.

Post-hoc strategy correlations do not identify their causal effects. The later randomized strategy changes are stronger tests. Candidate-opposition comparisons are not the same as a no-conversation control.

Evidence receipts

Unrelated museum example · assigned objective

Representing reasons and increasing late-opening support are distinct assignments. Inspect the instruction and actual behavior rather than infer the job from courtesy.

Unrelated museum example · matched comparison

Keep invitations, information, question wording and timing comparable when varying the objective. Otherwise more than one difference could account for the result.

Unrelated museum example · outcome

Ask separately about accurate representation, schedule preference and intention to attend. Changing one does not establish a change in all three.

Unrelated museum example · source boundary

Candidate-advocacy research can motivate an objective comparison. It does not estimate the effects of an invented museum assistant or settle its governing rule.

Further reading

Can AI Mediation Improve Democratic Deliberation?

Michael Henry Tessler, Georgina Evans, Michiel A. Bakker, and colleagues (2026). arXiv · author perspective, not a new experimental result. arXiv:2601.05904 · 9 January 2026

Free author-posted full text and PDF · CC BY 4.0.

The mediation authors discuss trade-offs among broad participation, thoughtful discussion, and equal contribution. They examine preference prediction, representative input, human oversight, and what a digital process can miss. This perspective develops design questions around the earlier experiment; it is not evidence that every proposed safeguard already works.

Which safeguard would test faithful representation of a resident's view, and which different safeguard would address who never entered the consultation?

[Complete publisher text](#)

Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale

Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, and colleagues (2023). Proceedings of the National Academy of Sciences. 120(41): e2311627120 · DOI 10.1073/pnas.2311627120

Free full text on PMC: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10576030/>.

In an online gun-policy experiment, an AI assistant suggested rephrasings that participants could accept, edit, or reject. Some conversational-quality and reciprocity measures improved, especially for the partners of assisted participants. The assistant did not produce additional policy-attitude change relative to control conversations. This separates improving how people hear each other from making them endorse the same policy; the effects were not clearly persistent at a later follow-up.

If participants disagree as much as before but understand each other better, which outcome has improved—and which claim should a public report avoid?

[Complete publisher text](#)