

The Persuasion Hearing

Research and political reasoning



Political communication, AI persuasion and civic accountability

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 recent experiments examine a live question in political communication: how much AI persuasion depends on personal targeting, and what other features matter. We use the corrected version of Salvi and colleagues' article. The pair is a debate about emphasis and mechanisms, not a head-to-head replication in an identical setting.

Article A crosses opponent type and demographic-profile access in assigned debates and reports agreement odds. Article B varies many model, training, instruction and

personalization conditions and measures attitude-score movement and claim accuracy. Its personalization methods differ from A's demographic profiles. An uncertain direct estimate in A and a small positive added estimate in B can describe a world in which effects vary across interventions and settings; their units cannot be subtracted. Comparing the studies should change which component claim you make, which outcome you measure and which audit you defend—not merely add a second citation. Keep a comparison card for each claim: the two conditions, assigned feature, population, measured outcome and unit. Agreement odds, attitude-score points, supported-claim shares and participation counts belong to different cards.

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 On the conversational persuasiveness of GPT-4

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West (2025 · corrected 2026). Nature Human Behaviour. 9: 1645–1653 · DOI 10.1038/s41562-025-02194-6

Dates

Published 19 May 2025; corrected 3 September 2026. Experiment conducted December 2023–April 2024.

Access

Open-access updated publisher text and PDF. Read the 3 September 2026 correction: <https://doi.org/10.1038/s41562-026-02588-0>

Lens

Ask whether the comparison changes the opponent, access to personal information, or both.

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

Experimental design; Aggregate results; Discussion; Human sample; and the linked Author Correction.

Reading summary

The opponent comparison and the personalization comparison answer different questions. The correction narrows what can be claimed about personalization itself; it does not remove the finding relative to human opponents.

[Complete publisher text](#)

Research explanation

In everyday terms: this study asks whether an AI becomes more persuasive when it knows demographic information about the person it is debating. Comparing that AI with a human asks a different question from comparing the same AI with and without a profile.

Personalization means tailoring an argument using something known about its audience. A civic bot might use a stored profile; it might also respond to information a person

voluntarily supplies. The governance question is whether access to that information adds persuasive power beyond the conversation itself.

Consider a different-topic museum experiment. An audio guide with visitor information is compared with an unprofiled human guide. If recall differs, both speaker and information access changed. Comparing that same audio guide with and without visitor information instead asks about the added component. These two comparisons can yield different amounts of uncertainty without disagreeing.

The study randomized 900 US adults to short structured debates with human or GPT-4 opponents, with or without demographic information. It measured reported agreement before and after. Personalized GPT-4 exceeded the human baseline; the headline 81.2% is a relative increase in odds of higher agreement, not the proportion persuaded.

The 3 September 2026 correction changes the direct comparison of personalized and nonpersonalized GPT-4: the estimated increase in relative odds is +48.7%, with a 95% interval from -2.9% to +127.6% and corrected $P = .0678$. The interval includes no difference for that contrast; it does not establish exactly zero effect. The personalized-GPT-4 versus unpersonalized-human result remains +81.2% relative odds, interval +26.0% to +160.7%. These are reported agreement odds, not percentages of persuaded people or election votes.

Method

A fair test of a component changes that component while keeping the comparison intelligible. Random assignment reduces the chance that one condition simply receives people who were already easier to persuade. It does not make a timed experiment identical to an everyday conversation. For a binary event, odds are its probability divided by the probability of its complement; an odds ratio compares those odds across conditions. This paper uses an ordinal-regression analysis of ordered agreement responses. The optional binary-odds illustration below teaches the ratio idea only; it is not a explanation of that model or a conversion of its estimate into a head count.

Explain four opponent/information conditions: a human without the recipient's demographic profile; a human with it; GPT-4 without it; and GPT-4 with it. In the human-personalized debates access was asymmetric; the aggregate analysis uses the recipient without access to the opponent's profile. Identify which contrast keeps the opponent type fixed and which changes both opponent type and information access. Then say what outcome that contrast measured.

Findings

Personalized GPT-4 exceeded the unpersonalized-human baseline in the reported agreement outcome. That comparison establishes a difference between those conditions, not the added effect of demographic information alone.

The corrected direct GPT-4 comparison is compatible with several component effects, including no difference and a positive advantage. A privacy right need not depend on demonstrating a large persuasive benefit; that is a separate normative justification.

Limits

Paid adults defended assigned positions in a controlled task. The study did not observe election votes or establish an effect for every chatbot, population, or voluntary user.

A privacy rule can be justified by control over personal information even when its additional effect on persuasion is uncertain. Evidence about effectiveness and a judgment about rights answer related, different questions.

Technical details

The 600 debates involved 900 unique US Prolific adults; the four opponent/information combinations were crossed with three sets of topic strength. The aggregate ordinal-regression analysis uses 750 observations because the personalized human condition includes only recipients without access to opponents' profiles.

The corrected personalized-versus-unpersonalized GPT-4 contrast is +48.7% relative odds, 95% CI -2.9% to +127.6%; corrected $P = .0678$, rounded .07. No correction to the headline human-baseline effect was reported. These estimates are optional details, not numbers to memorize.

Optional unrelated binary-odds illustration · entirely fictional parcel collection. In one 100-customer group, 25 collect a parcel by the next day and 75 do not: probability 25/100 = 25%; odds 25/75 = 1/3. In a second 100-customer group, 40 collect and 60 do not: probability 40%; odds 40/60 = 2/3. The odds ratio is $(2/3)/(1/3) = 2$, a 100% increase in odds, while the collection probability rises by 15 percentage points and 40% collect in the second group. These are different quantities. The made-up binary outcome is not Salvi's ordered agreement model; do not apply these counts or convert its reported relative-odds estimate into a percentage persuaded.

Evidence receipts

What changed in the comparison?

Check whether a comparison changes the speaker, personal information, or both. Write the two conditions in plain language before saying what caused the difference.

What does uncertainty allow?

An uncertain estimate can leave both an advantage and no advantage plausible. It supports caution about the specific contrast, not a declaration that the component never matters.

What does the headline measure?

Relative odds and the fraction of people changing their minds are different quantities. A headline about higher agreement cannot be read as a head count of converted voters.

Which judgment is still yours?

The experiment informs the capability question. Your rule also needs a stated purpose: privacy, factual reliability, transparency about advocacy, or some combination.

Contrast table

Article A · four opponent/information cells and two reported contrasts. All outcomes are reported agreement, not ballots. Population and unit cards remain attached to each comparison; optional exact estimates need not be memorized.

Cell or contrast	Opponent / information access	Question answered	Population / analysis frame	Measured outcome / unit
H0	Human without the recipient's demographic profile	Unpersonalized-human reference condition	US adults in controlled assigned debates; see the sample/analysis detail	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view
H1	Human with the recipient's demographic profile	Personalized-human condition; access asymmetric	US adults in controlled debates; aggregate recipient is the participant without access to the opponent's profile	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view
G0	GPT-4 without the recipient's demographic profile	Unpersonalized GPT-4 condition	US adults in controlled assigned debates; see the sample/analysis detail	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view
G1	GPT-4 with the recipient's demographic profile	Personalized GPT-4 condition	US adults in controlled assigned debates; see the sample/analysis detail	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view

Cell or contrast	Opponent / information access	Question answered	Population / analysis frame	Measured outcome / unit
G1 versus H0	Opponent and profile access both change	Package contrast: +81.2% relative odds; 95% CI [+26.0%, +160.7%]	US adults in controlled assigned debates; see the sample/analysis detail	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view
G1 versus G0	GPT-4 held fixed; profile access changes	Added-profile contrast: +48.7%; 95% CI [−2.9%, +127.6%]; corrected P=.0678	US adults in controlled assigned debates; see the sample/analysis detail	Ordered reported agreement; relative odds for reported contrasts, not a proportion changing view

Extra links

Read the named author correction · 3 September 2026

[Open the source](#)

Article B The levers of political persuasion with conversational artificial intelligence

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield (2025). Science. 390(6777): eaea3884 · DOI 10.1126/science.aea3884

Dates

Published 4 December 2025. Three experiments ran December 2024–May 2025.

Access

Author-hosted published full PDF. Methods and replication materials:
<https://github.com/kobihackenburg/scaling-conversational-AI>

Lens

Audit the system's objective and claims as well as its access to profiles.

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

How do models persuade?; How accurate is the information provided by the models?; Discussion; supplementary Experiment Methods.

Reading summary

More persuasive conditions can also produce more factual errors. This motivates an audit of objectives and claims; it does not identify errors as the cause of persuasion.

[Complete publisher text](#)

Research explanation

In everyday terms: this study asks which features make political AI conversations persuasive, and whether persuasive messages are accurate. More checkable claims do not automatically mean more true claims.

A chatbot has several features that can change together. It can know more about its user, draw on a different model, or be trained and instructed for a different purpose.

Post-training means adapting an already-built model. A system optimized for opinion change may prefer different replies from a system optimized for accurate understanding.

Information density here means the number of claims that could be checked, not a guarantee of truth. In a separate invented museum conversation, “the gallery opens at

ten,” “the sculpture is bronze,” and “an audio guide is available” are three checkable claims. They remain claims until a record verifies them. A long message containing opinions alone need not contain more checkable information.

Across 76,977 UK adults, 19 models, and 707 political issues, the researchers varied AI conditions and assessed opinion change. Personalization’s added effect was small and positive. Persuasion-oriented post-training and information-focused instructions produced larger gains. Conditions increasing persuasion often reduced claim accuracy.

The researchers compare experimental packages. An information-focused instruction can change the number of claims, length, style and other reply features together. Its effect does not by itself isolate claim count as the step connecting the cause to the result (causal mediator). Here mediator means a step through which a cause might affect an outcome, rather than a person or system facilitating a discussion. The separate fabrication condition added inaccuracies without a statistically clear persuasive gain. Errors and persuasion can rise together without errors being the demonstrated cause of persuasion.

Method

The experiment randomized conditions such as model instructions and personalization; its control groups did not receive the same political advocacy. Claim counts and accuracy were assessed afterwards, with automated checks validated against professional human fact-checkers.

Randomizing an instruction can show an effect of that instruction. It does not automatically isolate every feature it changes: claims, length, tone, and replies may move together. To test the cause of a fictional pilot difference, propose a comparison holding the rival features steady.

Findings

Personalization had a small positive added effect in this study; persuasion-oriented training and information-focused instruction packages produced larger gains. Some persuasive conditions also reduced claim accuracy.

This is a reason to measure objectives, replies, knowledge and support separately. It is not proof that every advocate is deceptive, that all opinion movement is harmful, or that the number of claims is the only operative process.

Limits

The participants were a convenience sample of adults in a study. Observed attitude effects do not by themselves establish changes in votes or public opinion after voluntary deployment.

An accurate argument may legitimately change a mind. Your task is to distinguish that possibility from unsupported claims and concealed advocacy, then explain which rules address which concern.

Technical details

The published results report a pooled added-personalization effect of +0.43 points on the study's 0–100 attitude metric, 95% CI .22–.64, compared with an 8.5-point unpersonalized persuasive effect. Small is not the same as zero.

Methods in the author repository distinguish randomized instructions from later associations between claim counts and effects. UK convenience sampling, voluntary engagement outside a paid study, and behavioral outcomes remain important limits. Exact counts and training procedures are optional.

Evidence receipts

What is counted?

Count claims that could be checked separately from words, sources, or opinions. A claim count tells you what to verify; it does not verify it.

Which comparisons answer the cause question?

A comparison with profiles disabled can show that profiles are not required for an effect. To isolate another component, keep the model, audience, and rival features comparable.

Does influence establish reliability?

Ask whether a changed opinion rests on sound information. Measure accuracy separately instead of using agreement with the system as a substitute.

Does an error establish deception?

A false claim is a reliability failure. Evidence about the system's objective, instructions, or decisions is needed for a claim that the operator intended deception.

Contrast table

Article B · personal-information intervention, assigned packages and measured outcomes. The treatments and units differ from Article A.

Comparison	What changes	Outcome and boundary	Population / evidence frame	Outcome / unit card
Personalized versus unpersonalized persuasive conditions	Personalization methods across experiments; not Article A's identical demographic-profile intervention	Added effect +0.43 attitude-score points, 95% CI [.22,.64]; small positive effect	UK adult convenience sample under the assigned study conditions	Attitude movement on the 0-100 score; score-point differences, not agreement odds
Persuasion-oriented training or information-focused instructions versus stated comparison	Assigned training/instruction package; multiple reply features can change	Attitude movement; does not uniquely isolate claims, length or tone	UK adult convenience sample under the assigned study conditions	Attitude movement on the 0-100 score; score-point differences, not agreement odds
Claim-accuracy audit across dialogues	Audit of produced checkable claims after conversations	Accuracy is a separate outcome; an association with movement is not proof of mediation	Produced claims audited after dialogues; adults in the UK study sample	Claim accuracy, separately assessed from attitude movement
Fabrication condition versus comparison	Instructions increase inaccurate claims	No statistically clear persuasive gain from that condition; not proof accuracy is politically irrelevant	UK adult convenience sample under the assigned study conditions	Attitude movement on the 0-100 score; score-point differences, not agreement odds

Further reading

Evaluating the persuasive influence of political microtargeting with large language models

Kobi Hackenburg and Helen Margetts (2024). Proceedings of the National Academy of Sciences. 121(24): e2403116121 · DOI 10.1073/pnas.2403116121

Verified author-hosted published full PDF; also openly available in PubMed Central.

This earlier experiment examines personalized versus generic GPT-4 political messages. Both moved attitudes relative to control; their average persuasive difference was not statistically clear. Its static-message comparison should not be silently substituted for a live-dialogue comparison.

What would have to remain the same before you could attribute different personalization findings to dialogue rather than another difference between studies?

[Complete publisher text](#)

Persuading voters using human-artificial intelligence dialogues

Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, and David G. Rand (2025). Nature. 648: 394–401 · DOI 10.1038/s41586-025-09771-9

Author-hosted published full PDF; publisher record:
<https://www.nature.com/articles/s41586-025-09771-9>

Experiments around elections in the US, Canada, and Poland randomly assigned candidate advocacy. Candidate-preference effects and inaccurate claims raise questions about public influence and information quality. Reported preference or vote intention is not an observed ballot.

Which additional evidence would connect a survey preference change with an effect on an actual election?

[Complete publisher text](#)