Beyond argumentation: AI-powered Socratic dialogue and political moderation

José Ramón Enríquez*
Stanford University

March 25, 2026

Abstract

I examine whether AI-guided reasoning reduces issue polarization in two preregistered experiments conducted on `deliberation.io`, a purpose-built platform for deliberation research, with 5,000 participants across contentious U.S. policy issues. Relative to reflective writing, AI-emotional regulation, and AI-grammar correction, AI-Socratic dialogue —which prompts users to articulate supporting arguments— significantly moderates extreme positions on mental health-based gun regulation. A second experiment comparing AI-Socratic dialogue to AI-grammar correction across abortion, handgun regulation, and voter ID requirements finds that extreme participants moderate their positions and exhibit behavioral changes, including increased cross-partisan donations and reduced endorsement of extreme comments. Moderation effects are largest for abortion, where baseline cross-partisan agreement was highest, and concentrate among participants with low initial confidence, high susceptibility to elite cues, and limited cross-party interactions. Text analysis of conversational transcripts reveals that moderation operates through increased consideration of alternative perspectives rather than improved argument quality, challenging standard deliberative theory. One-month follow-up displays mixed persistence and minimal spillovers to untreated issues, consistent with issue-specific effects. Perceived agency remains equivalent across conditions, suggesting preserved autonomy. Overall, results indicate that AI conversational agents can facilitate compromise through previously under-explored mechanisms, offering promising scalable applications.

JEL Codes: D72, D83, L86, O33

Keywords: Deliberation, Large Language Models, Generative Artificial Intelligence, Polarization

*Postdoctoral Fellow at the Digital Economy Lab, Stanford Institute for Human-Centered AI, and the Golub Capital Social Impact Lab, Stanford Graduate School of Business. I thank Susan Athey, Camila Blanes, Erik Brynjolfsson, Horacio Larreguy, Jiaxin Pei, Alex Pentland, Julio Solís, Lily Tsai and seminar participants at Google, Stanford, and Yale for comments and suggestions. Guinness Chen, Umar Patel, and Chinmaya Pimpalkhare provided excellent research assistance. I acknowledge Project Liberty Institute for generous support. The experiments are preregistered with the AEA RCT Registry (AEARCTR-0015461). This research was approved by Stanford University Human Subjects Research Compliance Office (IRB-75616) and MIT Committee on the Use of Humans as Experimental Subjects (IRB-5798). The platform enabling this intervention can be directly accessed at `deliberation.io`. All errors and omissions are my own. Email: jrenriquez@stanford.edu.

1 Introduction

Political polarization has intensified across developed democracies, with citizens increasingly sorted into ideological echo chambers that limit cross-cutting exposure and undermine governance (e.g., Gentzkow and Shapiro, 2011; Boxell et al., 2024). This polarization manifests in reduced cross-partisan cooperation (McCarty et al., 2016; McConnell et al., 2018), increased affective hostility between partisan groups (Lelkes et al., 2017), and growing geographic, economic, and social segregation along political lines (Chinoy and Koenen, 2024; Kim, 2024). While extensive research documents the causes and consequences of polarization, scalable interventions to reduce political divides remain limited (e.g., Voelkel et al., 2023; Allcott et al., 2024; Enríquez et al., 2024; Voelkel et al., 2024; Beknazar-Yuzbashev et al., 2025).

Digital platforms offer particular promise for hosting inclusive political conversations, potentially reaching citizens who cannot or do not wish to participate in traditional face-to-face forums (Cohen et al., 2021; Landemore, 2021). However, online conversations often fall short of deliberative ideals, namely the open exchange of reasons and evidence among citizens for a common good (Habermas, 1985, 2015; Cohen, 2007). While theory suggests that fostering dialogue among individuals can increase mutual understanding, particularly through the exchange of high-quality arguments that help participants engage with evidence and logic (Benson, 2021, 2023; Chambers, 2018), most online platforms incentivize quick, emotive responses over thoughtful reasoning (Brady et al., 2017; Pennycook et al., 2021; Tucker et al., 2018). The resulting interactions tend toward polarization rather than compromise (Bail et al., 2018), with participants reinforcing existing beliefs rather than updating views based on new information. Recent advances in generative artificial intelligence offer new possibilities to address these challenges by augmenting civil interactions and promoting mutual understanding at scale (Tsai et al., 2024; Pentland and Tsai, 2024). However, rigorous empirical evidence on their effectiveness for reducing polarization remains scarce.

This paper studies whether AI-guided Socratic dialogue —where an AI conversational agent prompts users to articulate arguments supporting their positions— can improve the quality of online political conversations and foster compromise on polarizing issues. To address this question, I conduct two randomized controlled trials using `deliberation.io`, a purpose-built open-source platform designed for rigorous research on deliberation (Pei et al., 2025). Unlike existing social media platforms or commercial deliberation tools, this custom platform enables precise manipulation of conversational features while maintaining ecological validity through authentic political discussions among citizens. The platform allows researchers to randomly assign participants to different AI-powered modules and compare them against non-AI control conditions, while tracking detailed attitudinal and behavioral outcomes including belief changes, comment endorsements, and donation decisions.

Theoretically, I develop a framework grounded in correlation neglect in belief formation (Ortoleva and Snowberg, 2015; Enke and Zimmermann, 2019; Eyster et al., 2019). I model political positions as arising from Bayesian updating where individuals aggregate signals from multiple sources —personal experiences, social networks, media consumption— that are fundamentally correlated

through shared factors like geographic sorting and partisan information ecosystems. When individuals fail to recognize this correlation structure, they overstate the amount of independent information they possess, inflating confidence in their beliefs and generating political extremism (Ortoleva and Snowberg, 2015). The Socratic method reduces polarization by increasing correlation awareness: systematic questioning prompts individuals to examine the origins of their beliefs, revealing that seemingly independent sources actually stem from common partisan origins or that citizens possess fewer independent arguments than they actually thought. This recognition mechanically reduces perceived precision and increases weight on alternative perspectives without providing new substantive information. The framework generates testable predictions that distinguish this epistemic openness mechanism from alternative pathways emphasizing argument quality or evidence-based reasoning. While the experiment does not directly measure correlation perception, the observed pattern —reduced confidence and increased perspective-taking without improved argument quality, concentrated among individuals in homogeneous information environments— provides indirect evidence consistent with the correlation-awareness pathway. This behavioral foundation helps explain both why traditional interventions emphasizing factual correction often fail and why AI-facilitated reflection focusing on metacognitive awareness succeeds in reducing polarization.

Empirically, the first experiment examines gun purchase regulations for people with mental health conditions with approximately 1,500 crowdsourced participants residing in the United States. After providing an initial response comment, position on a Likert scale, and confidence level, participants were randomly assigned to one of four active treatment conditions or a waiting control group. The AI-Socratic dialogue condition engaged users in a five-round conversation where a large language model (LLM) asked participants to explicitly articulate the arguments supporting their positions. The AI conversational agent was prompt-tuned to function uniquely as an interrogator, asking clarifying questions, without introducing external information or taking substantive positions on the issues. The remaining active conditions were designed to require equivalent participant engagement through writing tasks that prompt self-reflection or conversations with LLM-guided interventions focused on emotional regulation or grammar correction. I evaluate treatment effectiveness using a within- and between-subject design, collecting baseline and endline surveys to measure positioning, confidence, comment endorsements on seeded content, and actual donation decisions to a cross-partisan advocacy group. I assess argument quality through a pre-defined rubric evaluated by both human annotators and fine-tuned LLMs.

Based on preregistered analysis, I find that extreme participants —those holding strong positions either favoring or opposing the policy premise— in the AI-Socratic dialogue condition exhibit significantly greater movement toward the center of the opinion distribution compared to extreme participants in other treatment conditions. Specifically, extreme participants in the AI-Socratic condition are 9.4 percentage points less likely to maintain extreme positions compared to extreme participants in the waiting control (SE = 3.98), representing approximately 20% of the control standard deviation. They are also 4.6 percentage points more likely to move toward the topic pre-treatment median (SE = 1.89), nearly doubling the probability of centrist movement. Other

treatment conditions show substantially smaller and statistically insignificant effects, suggesting that the specific questioning approach of Socratic dialogue drives the observed moderation.

Post-treatment comments in the AI-Socratic dialogue condition are not more thoughtful compared to the waiting control, with estimates pointing toward precisely estimated zeros on a comprehensive argument quality rubric. However, AI-Socratic dialogue increases perspective diversity by 0.07 standard deviations (SE = 0.031) among extreme participants. This finding suggests that the intervention promotes consideration of alternative viewpoints without improving overall logical rigor, challenging core assumptions about the mechanisms underlying political attitude change.

To examine whether these findings hold across different contentious topics, I conducted a second experiment with approximately 3,500 online participants examining three topics with varying prior distributions and confidence levels: gestational limits for abortion, handgun possession bans, and voter ID requirements. Using a similar design, participants were randomized to either an AI-Socratic dialogue or AI-Grammar correction condition to control for AI chatbot interaction effects.

Consistent with the first experiment, extreme users in the AI-Socratic dialogue condition moved toward more moderate positions, with moderation effects concentrated on abortion where cross-partisan overlap is greatest. Extreme participants in the AI-Socratic condition are 3.1 percentage points less likely to maintain extreme positions when pooling across topics (SE = 1.6), with the strongest effects for abortion policy (4.6 percentage points, SE = 2.3). Two behavioral outcomes reinforce these findings: extreme users in the AI-Socratic dialogue condition are 12 percentage points more likely to donate a $1 USD bonus to cross-partisan advocacy groups (SE = 4.74) and 6 percentage points less likely to endorse extreme political comments (SE = 4.34).

Follow-up analysis conducted one month later provides evidence for partial persistence of treatment effects on the same topics, with minimal spillover to untreated issues. Differential attrition poses minimal threats to internal validity: main survey effects for the follow-up subsample closely match those for the full sample (3.3 percentage points vs. 3.0 percentage points reduction in extreme positioning), and follow-up participation rates show no systematic relationship with treatment assignment and observables. When pooling across all three topics, extreme participants who received AI-Socratic dialogue are 3.7 percentage points less likely to maintain extreme positions one month later, representing complete persistence relative to the 3.3 percentage point main survey effect for this subsample, though the estimate is not statistically significant due to reduced statistical power ($p = 0.185$). Topic-specific analysis reveals suggestive evidence of lasting moderation for abortion (9.0 percentage points) and voter ID requirements (7.8 percentage points). Spillover effects to an entirely new political topic are limited and statistically insignificant (1.6 percentage points overall), suggesting that the intervention's benefits operate through engagement with specific policy content rather than developing generalizable deliberative skills.

Preregistered heterogeneous effect analysis indicates that moderation effects concentrate among participants with low baseline cross-party interactions, low baseline confidence, high susceptibility to elite cues, and high reputation concerns about changing positions. Data-driven heterogeneity analysis using causal forests confirms these patterns: participants in the highest treatment-effect

quartile exhibit markedly lower belief confidence (0.44 vs. 0.56 scaled deviations from the mean in the lowest quartile), higher social conformity concerns (0.27 vs. 0.17), and greater elite cue susceptibility compared to participants in the lowest quartile. These patterns suggest AI interventions may be particularly effective for politically disengaged citizens who rely primarily on ideologically aligned information sources.

Text analysis of the AI assistant and participant conversational transcripts reveals that the AI-Socratic agent uses substantially more epistemic openness language (0.156 SDs), perspective-taking language (0.516 SDs), and hedging language (0.152 SDs) compared to the AI-grammar correction, with participants responding similarly across all three dimensions. Notably, argument quality shows no improvement and actually decreases markedly in AI messages (-0.990 SDs) while remaining unchanged in participant responses, confirming that the intervention operates through epistemic openness rather than enhanced reasoning quality. Falsification checks demonstrate that the Socratic AI uses more topic-specific vocabulary in its questions while participants maintain similar usage regardless of condition, and placebo terms from unrelated policy domains show no differences across treatment arms, validating the measurement approach.

Additional results based on the one-month follow-up sample indicate that extreme participants in the AI-Socratic condition report 8.6 percentage points higher openness to others' perspectives ($p < 0.10$) and suggestive increases in seeking information about other polarized political topics (8.4 percentage points). Crucially, participants report similar levels of perceived agency and neutrality across treatment conditions, suggesting the intervention preserves user autonomy while encouraging moderation.

Across both experiments, I find limited evidence that argumentative quality improvements drive the observed moderation effects. While AI-Socratic dialogue significantly enhances perspective diversity by 0.07–0.08 standard deviations in both experiments, it produces precisely estimated null effects on overall thoughtfulness, challenging core assumptions that high-quality reasoning is necessary for meaningful political compromise. Instead, the results suggest that fostering openness to alternative viewpoints may be more relevant, pointing toward new mechanisms for reducing polarization in digital environments.

This study makes three main contributions to the literature on political economy, the economic and political implications of artificial intelligence, and online platforms design. First, it provides novel causal evidence that scalable interventions leveraging AI conversational agents can reduce political polarization without undermining user agency. While most of the political economy literature documents the role of partisan media, affective and ideological sorting, and social networks in amplifying divides (DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2011; Boxell et al., 2017; Martin and Yurukoglu, 2017; Levy, 2021; Chinoy and Koenen, 2024), depolarizing interventions that can operate at scale remain limited (e.g., Levy, 2021; Akbiyik et al., 2025; Voelkel et al., 2023; Allcott et al., 2024; Enríquez et al., 2024; Voelkel et al., 2024; Greene et al., 2025; Beknazar-Yuzbashev et al., 2025). By showing that generative AI can induce both attitudinal and downstream behavioral changes across multiple contentious domains, my findings suggest new

technological avenues to improve cross-partisan cooperative exchanges, complementing recent work on algorithmic tools for collective intelligence and online deliberation (Wright and Street, 2007; Argyle et al., 2023; Goldberg et al., 2024; Tessler et al., 2024; Tsai et al., 2024).

Second, the study challenges conventional assumptions about the mechanisms underlying political attitude change and compromise formation. While canonical work highlights argument quality and evidence-based discourse as drivers of opinion change and consensus (Cohen, 2007; Habermas, 2015; Fishkin, 2018; Chambers, 2018), my findings indicate that political moderation can occur independently of improvements in argumentative rigor. Interventions that enhance psychological openness and epistemic humility appear more consequential than those enhancing logical reasoning skills. This finding shifts the analytical focus from purely cognitive mechanisms to social-psychological constraints on attitude and behavioral change, with important implications for understanding how societies can bridge political divides. The results also help reconcile mixed evidence that exposure to counter-attitudinal content can backfire (e.g., Bail et al., 2018; Mousa et al., 2025; Enríquez et al., 2024) with designs that reduce conflict by reshaping conversational dynamics rather than information content per se (Argyle et al., 2023; Tessler et al., 2024).

Third, I contribute methodologically by introducing `deliberation.io`, an experimental platform that enables causal identification in real-time political discussions with generative-AI assistance. Unlike observational analyses of digital discourse, the design delivers experimental variation in conversational form and content while preserving ecological validity. This innovation speaks directly to an emerging agenda in economics and governance exploring how AI systems can be harnessed to promote civility, participation, and compromise while producing data-driven evidence and acknowledging potential limitations (Schiff et al., 2024; Velez, 2024; Hackenburg et al., 2025; Jungherr and Rauchfleisch, 2025).

Together, the results suggest that interventions to foster compromise in large-scale participatory processes may be most effective when they target openness and social-epistemic frictions, not just analytical reasoning skills. This perspective challenges civic-education approaches centered exclusively on critical-thinking while supporting interventions that address emotional, social, and epistemic barriers to attitude and behavioral change (e.g., Moehring et al., 2023; Pinzón Hernández et al., 2024; Athey et al., 2025a; Enríquez et al., 2024). Furthermore, the present experimental design addresses potential limitations through randomization across multiple contentious topics, behavioral outcome measures, and the recruitment of participants across ethnic, age, gender, and partisan affiliation groups. Ultimately, however, external validity depends on successful scaling beyond controlled environments. Future applications span participatory budgeting, deliberative polling, citizens' assemblies, and internal organizational consultations, where AI could play a critical role in lowering barriers to engagement and sustaining civility in polarized societies.

The remainder of the paper proceeds as follows. Section 2 reviews relevant literature on polarization, AI-facilitated dialogue, and democratic deliberation. Section 4 presents the political contexts and topic selection. Section 5 describes the experimental design and platform architecture. Section 6 presents results from both experiments, including underlying mechanisms and additional

analysis. Section 7 discusses implications for political polarization and the design of deliberative platforms. Section 8 concludes.

2 Related literature

This section examines related literature in light of the study's three primary contributions: providing causal evidence on scalable polarization reduction, challenging conventional assumptions about political attitude change mechanisms, and introducing a methodological innovation for studying AI-facilitated political conversations. First, I discuss foundational theoretical work on deliberative democracy, which emphasizes reasoned argumentation as the pathway to political compromise. Then, I examine empirical work on online political discourse and emerging research on AI as a tool for improving political conversations

2.1 Finding compromise through argumentation

The theoretical foundations of deliberative democracy rest on the premise that legitimate political decisions emerge from inclusive public discourse where citizens exchange reasons and justify their positions to one another (Habermas, 1985; Cohen, 2007; Gutmann and Thompson, 2004). Central to this framework is the argument that high-quality deliberation, characterized by reasoned argumentation, evidence-based discourse, and mutual respect, can transform individual preferences and generate consensus around the common good (Chambers, 2018; Fishkin, 2018). This "better argument" thesis suggests that exposure to superior reasoning naturally leads to attitude convergence and political compromise.

However, more recent scholarship has challenged the exclusive focus on rational argumentation in deliberative theory. Mansbridge (1996) argues that deliberative processes must accommodate both self-interest articulation and emotional expression, noting that "appeals for the common good require an emotional and cognitive probing of one's own feelings of empathy, admiration, revulsion, or horror." While recognizing that appeals to emotion can be dangerous, this perspective acknowledges that purely logical reasoning may play a secondary role in representative political conversations (Gastil, 2000; Sanders, 1997).

My results suggest that fostering psychological openness proves more consequential than enhancing logical rigor, aligning with broader conceptions of communication that can bridge political differences (Young, 2002; Mansbridge et al., 2017). This finding connects to a growing body of work in political and behavioral economics examining the psychological and social foundations of political belief formation (Lodge and Taber, 2013; Alesina et al., 2020; Enke and Graeber, 2023; Bursztyn et al., 2020; Peterson and Iyengar, 2021). For instance, research on motivated reasoning demonstrates that individuals often use their cognitive abilities to rationalize pre-existing beliefs rather than update them based on new evidence (Klayman and Ha, 1987; Nyhan and Reifler, 2010; Thaler, 2024). Studies of partisan bias show that even highly educated individuals exhibit strong tendencies to interpret information in ways that confirm their prior political commitments (Taber

and Lodge, 2006; Schwardmann et al., 2022). Taken together, this body of work suggests that political compromise may require interventions that address psychological barriers to attitude change rather than simply providing better arguments or more information.

2.2 The promise and pitfalls of online deliberation

The rise of political polarization has generated extensive research on its causes and consequences for democratic governance (see Cruz and Larreguy 2025 and Larreguy and Raffler 2025 for a review). Studies document how partisan media consumption (DellaVigna and Kaplan, 2007; Braghieri et al., 2024; Djourelova et al., 2025), ideological sorting (Gentzkow and Shapiro, 2011; Eady et al., 2025; Hobolt et al., 2024), and social network effects (Levy, 2021) contribute to growing political divides. This polarization manifests in reduced cross-partisan cooperation (McConnell et al., 2018), increased affective hostility between partisan groups (Lelkes et al., 2017), and social and geographic segregation that limits cross-cutting political exposure (Chinoy and Koenen, 2024; Kim, 2024).

Digital platforms initially offered promise for bridging these divides by enabling large-scale political conversations across geographic and social boundaries (Wright and Street, 2007; Akbiyik et al., 2025; Levy, 2021; Rossiter and Carlson, 2024; Braghieri et al., 2025; Carney, 2025). However, empirical studies reveal systematic failures to achieve these goals. The architecture of most digital platforms incentivizes rapid, emotional responses over thoughtful reasoning (Brady et al., 2017; Tucker et al., 2018), with documented negative effects on political discourse quality and user well-being (Allcott et al., 2020; Braghieri et al., 2022). In fact, Bail et al. (2018) demonstrate that exposure to opposing political content on social media can actually increase polarization, with participants becoming more entrenched in their initial positions rather than moderating their views. This polarized behavior can escalate to violence, as documented by multiple studies of hate crimes against refugees in Germany (Müller and Schwarz, 2021; Jiménez Durán et al., 2024).

Recent interventions have attempted to address these challenges through various platform design modifications. Kriplean et al. (2012) developed ConsiderIt, a platform that structures political discussions around pro and con arguments, finding modest improvements in cross-partisan argument consideration. Munger (2017) shows that automated messages promoting civility can reduce harassment on Twitter. However, these interventions often face trade-offs between effectiveness and scalability, with human moderation proving effective but resource-intensive. Despite extensive research documenting polarization's causes and effects, scalable, durable interventions to reduce political divides remain limited (Voelkel et al., 2023; Allcott et al., 2024; Enríquez et al., 2024; Voelkel et al., 2024; Cruz and Larreguy, 2025).

This study addresses the scalability challenge by leveraging AI to provide consistent facilitation that would be prohibitively expensive with human moderators. Unlike previous interventions that focus primarily on platform design or content moderation, I test whether AI-guided conversational interventions can directly influence the cognitive and psychological processes underlying political attitude change.

2.3 AI as a deliberative facilitator

The emergence of large language models has opened new possibilities for scalable interventions in political discourse. Early experiments demonstrate AI's potential to facilitate productive political conversations through various mechanisms. Argyle et al. (2023) find that AI-generated evidence-based suggestions improved participants' quality perceptions in gun policy discussions, though without effects on policy positions. Tessler et al. (2024) show that AI can help groups find common ground statements without marginalizing minority perspectives, suggesting mechanisms beyond traditional argumentation. Konya et al. (2025) demonstrate that AI can facilitate compromise in peace-building discussions by helping participants identify shared values and visions.

This nascent literature suggests that AI systems may be particularly effective at addressing the psychological and social barriers that prevent productive political conversations. Unlike human facilitators, AI can provide patient, non-judgmental interaction that may be less threatening to participants' self-concept (Tsai et al., 2024). The scalability of AI facilitation means that such interventions could reach populations unlikely to participate in traditional deliberative forums while maintaining consistency across interactions.

However, important concerns have been raised about AI manipulation and user agency in political contexts (Kirk et al., 2024; Sorensen et al., 2024; Hackenburg et al., 2025). Recent work emphasizes the importance of designing AI systems that preserve human autonomy while promoting constructive dialogue (Pentland and Tsai, 2024; Goldberg et al., 2024). This study contributes to this literature by testing a specific mechanism—AI-guided Socratic questioning—that restricts the AI's role to asking questions rather than providing substantive content, ensuring that attitude changes emerge from participants' own reasoning processes rather than AI persuasion.

The Socratic approach, characterized by systematic questioning to expose assumptions and encourage critical thinking, has demonstrated effectiveness in educational contexts (Paul, 1993). Recent work shows that AI systems can successfully implement Socratic questioning (Ding et al., 2024), but research has focused primarily on testing AI functionality rather than examining downstream effects on user attitudes and behavior. This study fills this gap by providing rigorous experimental evidence on whether AI-guided Socratic dialogue can reduce issue polarization through enhanced epistemic openness rather than improved argumentation.

3 Theoretical framework

This section develops a theoretical framework to explain how AI-guided Socratic dialogue can reduce political polarization. I distinguish between two candidate mechanisms: an *argument-quality hypothesis*, which emphasizes cognitive processing of superior reasoning, and an *epistemic openness hypothesis*, which focuses on reducing overconfidence stemming from correlation neglect in belief formation. The framework generates testable predictions regarding when and why AI-guided reflection interventions might succeed, which I evaluate against the empirical findings.

3.1 Ideology as Bayesian updating with correlation neglect

Consider individual i forming a policy position $\theta_i \in \Theta$ on a particular political issue. Following recent work on correlation neglect in belief formation (Ortoleva and Snowberg, 2015; Enke and Zimmermann, 2019), I model ideology as having a behavioral basis rooted in how citizens aggregate information from multiple sources.

The individual receives N signals $\{s_1, s_2, ..., s_N\}$ about the optimal policy $\theta^* \in \Theta$ from various seemingly distinct sources—personal experiences, social network conversations, media consumption. Each signal takes the form $s_n = \theta^* + \epsilon_n$, where $\text{Corr}(\epsilon_n, \epsilon_m) = \rho$ for $n \neq m$. These signals are fundamentally correlated through shared underlying factors: geographic sorting into like-minded communities, partisan information ecosystems that filter news through similar lenses, and homophilous social networks where friends share common backgrounds and experiences.

Under correlation neglect, the individual fails to recognize this correlation structure, treating signals as independent despite $\rho > 0$. This behavioral friction —documented in recent economic literature (Ortoleva and Snowberg, 2015; Enke and Zimmermann, 2019; Eyster et al., 2019)— leads individuals to overstate the amount of independent information they possess. The degree of neglect can be parameterized by correlation awareness $\omega_i \in [0, 1]$, where $\omega_i = 0$ represents complete neglect (treating all signals as independent) and $\omega_i = 1$ represents full rational recognition of correlation.

This friction directly affects perceived precision of beliefs. Let τ_s denote individual signal precision and τ_0 denote prior precision. The individual's perceived precision is:

$$\hat{\tau}_i(\omega_i) = \tau_0 + \frac{N\tau_s}{1 + (N-1)\rho\omega_i} \tag{1}$$

When $\omega_i = 0$ (complete neglect), perceived precision becomes $\hat{\tau}_i = \tau_0 + N\tau_s$, treating each signal as providing independent information. When $\omega_i = 1$ (full awareness), perceived precision correctly accounts for correlation: $\hat{\tau}_i = \tau_0 + \frac{N\tau_s}{1+(N-1)\rho}$. Since $\frac{\partial \hat{\tau}_i}{\partial \omega_i} < 0$, increased correlation awareness reduces perceived precision toward its rational benchmark.

The individual's posterior policy position combines a common prior θ_0 (reflecting moderate or centrist positions prevalent in the broader population) with a subjective mean belief μ_i (incorporating personal experiences, preferences, and partisan bias), weighted by their relative precision:

$$\theta_i = \frac{\tau_0\theta_0 + \hat{\tau}_i(\omega_i)\mu_i}{\tau_0 + \hat{\tau}_i(\omega_i)} \tag{2}$$

Equation (2) reveals how correlation neglect generates political extremism. When ω_i is low, perceived precision $\hat{\tau}_i$ is inflated relative to the rational benchmark, causing θ_i to place excessive weight on the idiosyncratic subjective mean μ_i and relatively low weight on the common prior θ_0. Individuals thus adopt more extreme positions than warranted by their true information, with the degree of extremism depending on the extent of neglect, realized evidence, and political bias (Ortoleva and Snowberg, 2015). This increases dispersion in the distribution of policy positions, placing more mass on the distributional tails—precisely the observed empirical pattern of political

polarization.

3.2 Two pathways to political moderation

I now contrast two distinct mechanisms through which AI-guided Socratic dialogue might reduce this polarization, drawing theoretical distinctions that generate testable empirical predictions.

Argument-quality hypothesis. Classical deliberative theory posits that political attitude change occurs primarily through exposure to high-quality arguments and evidence-based reasoning (Cohen, 2007; Habermas, 2015). In the Bayesian framework, this mechanism operates by providing new high-quality signals s^* with precision $\tau^* > \tau_s$ that shift the subjective mean μ_i toward positions better supported by higher-quality evidence. Alternatively, interventions may improve individuals' ability to process existing signals, effectively increasing signal precision τ_s in equation (1).

Under this hypothesis, exposure to high-quality deliberation leads to belief convergence. Individuals update their subjective means toward evidence-supported positions: $\mu_i^1 \to \theta^*$, which mechanically moderates positions through equation (2).

Epistemic openness hypothesis. An alternative pathway focuses on reducing overconfidence stemming from correlation neglect rather than improving reasoning quality or providing new substantive information. This mechanism operates by increasing correlation awareness ω_i, thereby reducing inflated perceived precision $\hat{\tau}_i$ without necessarily changing the subjective mean μ_i.

The Socratic method is particularly well-suited to revealing correlation among information sources. Systematic questioning prompts individuals to trace how many independent arguments they know that validate their position. Through this examination, participants may discover that multiple seemingly independent sources all derive from partisan media ecosystems, that personal experiences generalize poorly beyond specific contexts, or that friends' opinions reflect shared demographics rather than independent evidence evaluation. This revelation increases correlation awareness: $\omega_i^0 \to \omega_i^1$ where $\omega_i^1 > \omega_i^0$.

Critically, this recognition can occur without participants receiving new information about policy substance. Individuals discover correlation among their information sources through self-reflection, avoiding defensive responses triggered when views are directly challenged (Bénabou and Tirole, 2011). The Socratic method positions participants as agents of their own epistemic recalibration rather than targets of correction.

3.3 Theoretical predictions and empirical tests

The framework generates predictions on when AI-guided reflection interventions would be most effective. I organize these predictions around the main predictions on moderation and confidence, followed by three key dimensions that moderate the gap between perceived and true precision.

3.3.1 Main predictions

Moderation and reduction in confidence. From equation (1), increased correlation awareness mechanically reduces perceived precision: $\hat{\tau}_i(\omega_i^1) < \hat{\tau}_i(\omega_i^0)$. This reduction in overconfidence has two key consequences that jointly distinguish the epistemic openness hypothesis from the argument-quality hypothesis.

First, from equation (2), lower perceived precision automatically increases the weight placed on the common prior θ_0 relative to the idiosyncratic subjective mean μ_i. Since $\hat{\tau}_i(\omega_i^1) < \hat{\tau}_i(\omega_i^0)$, the posterior position θ_i^1 moves closer to the moderate common prior compared to the initial position θ_i^0. The weight on others' perspectives is:

$$w_i^{\text{others}} = \frac{\tau_0}{\tau_0 + \hat{\tau}_i(\omega_i)}, \tag{3}$$

where $\frac{\partial w_i^{\text{others}}}{\partial \omega_i} > 0$. Recognizing that one's information sources are correlated mechanically increases the weight on alternative viewpoints when forming policy positions.

Second, lower confidence reflects increased epistemic humility: acknowledging that one's information sources provide less independent evidence than previously believed. This manifests as reduced self-reported confidence $c_i = h(\hat{\tau}_i(\omega_i))$ where $\frac{\partial c_i}{\partial \hat{\tau}_i} > 0$, generating $c_i^1 < c_i^0$ as correlation awareness increases.

Together, these consequences formalize epistemic openness: individuals simultaneously become less confident in their own positions and more receptive to alternative perspectives. The epistemic openness hypothesis thus predicts attitude moderation can occur independently of argument quality improvements; the intervention directly reduces confidence in political positions; and participants increase weight on others' perspectives when forming opinions.

3.3.2 AI-guided reflection effectiveness

High correlation, low awareness environments. The treatment effects are largest when individuals have high underlying correlation ρ among information sources but low baseline awareness ω_i^0 of this correlation. Three observable characteristics jointly proxy for this condition.

First, participants with *limited cross-partisan exposure* likely experience high correlation among information sources. Low cross-partisan exposure E_i means personal experiences, social network conversations, and media consumption all occur within homogeneous partisan environments, creating highly correlated signals (ρ high) that individuals may mistakenly treat as independent (ω_i^0 low). The intervention substitutes for the natural correlation-revealing function of cross-partisan interaction:

$$\frac{\partial^2 \Pr(\theta_i^1 \neq \theta_i^0)}{\partial T_i \partial E_i} < 0, \tag{4}$$

where $T_i \in \{0, 1\}$ indicates treatment assignment.

Second, participants who *rely heavily on elite cues* likely treat highly correlated signals as independent. Let $\alpha_i \in [0, 1]$ denote susceptibility to partisan elite guidance. High α_i implies that

personal experiences, social network views, and media consumption all stem from common partisan origins filtered through elite messaging, generating high correlation that individuals fail to recognize. When prompted to articulate independent reasoning, these individuals discover correlation:

$$\frac{\partial^2 \Pr(\theta_i^1 \neq \theta_i^0)}{\partial T_i \partial \alpha_i} > 0 \tag{5}$$

Third, participants with *high reputational concerns* about changing positions face social costs that can prevent belief updating even when individuals recognize overconfidence. Let utility from updating be:

$$U_i(\theta_i^1) = V(\theta_i^1 \mid \omega_i, \mu_i) - S_i \cdot \mathbb{1}[\theta_i^1 \neq \theta_{\text{group}}], \tag{6}$$

where $V(\cdot)$ represents epistemic utility and S_i measures social cost of deviating from the reference group. High S_i makes individuals particularly responsive to interventions providing private, non-threatening environments for recognizing information correlation. The anonymous AI dialogue reduces effective social cost ($\tilde{S}_i < S_i$), enabling belief updating:

$$\frac{\partial^2 \Pr(\theta_i^1 \neq \theta_i^0)}{\partial T_i \partial S_i} > 0 \tag{7}$$

Moderate baseline confidence. The framework predicts nonlinear effects with respect to baseline confidence, reflecting both mechanical and behavioral forces.

Mechanically, from equation (2), individuals with moderate baseline confidence have the largest potential for position change because they place intermediate weight on both their subjective mean μ_i and the common prior θ_0. Reducing $\hat{\tau}_i$ shifts this balance meaningfully toward θ_0. Conversely, individuals with very low initial confidence already place substantial weight on common priors, limiting scope for additional moderation.

Behaviorally, highly confident individuals may have political identities built on perceived expertise, making acknowledgment that information sources are less independent psychologically costly (Bénabou and Tirole, 2011). The social cost term in equation (6) captures this constraint: even when individuals could recognize overconfidence, identity concerns prevent position change. This yields:

$$\frac{\partial^2 \Pr(\theta_i^1 \neq \theta_i^0)}{\partial T_i \partial (c_i^0)^2} < 0, \tag{8}$$

where the quadratic term captures the inverted-U relationship between baseline confidence c_i^0 and treatment responsiveness.

Overlap of common priors across issues. The framework predicts that intervention effectiveness varies systematically across political issues based on the distribution of subjective means in the population. Critically, treatment effects would be largest when different groups' subjective means μ_i cluster near common values despite apparent polarization.

Let $\sigma^2_{\mu,j}$ represent the variance in subjective means across individuals on issue j. When this

variance is small —indicating substantial overlap in prior distributions across partisan groups— individuals hold similar underlying preferences but correlation neglect inflates confidence, causing them to adopt more extreme positions than their true preferences warrant. On such issues, reducing perceived precision through increased correlation awareness allows positions to converge toward the overlapping common prior. Conversely, when $\sigma^2_{\mu,j}$ is large, fundamental disagreements in subjective means limit the scope for moderation even when overconfidence is reduced. The treatment effect on issue j is:

$$\tau_j = g(\sigma^2_{\mu,j}, \bar{\rho}_j, \bar{\omega}^0_j), \tag{9}$$

where $\bar{\rho}_j$ represents average correlation among information sources on issue j, $\bar{\omega}^0_j$ represents average baseline correlation awareness, and $\frac{\partial g}{\partial \sigma^2_{\mu,j}} < 0$. Greater overlap in subjective means (lower $\sigma^2_{\mu,j}$) amplifies the depolarizing effect of correlation-awareness interventions, as reducing overconfidence reveals the underlying consensus that inflated precision concealed. Issues with high prior overlap would thus exhibit the largest treatment effects because correlation neglect creates artificial polarization around a shared center, which the intervention can effectively eliminate.

Empirical tests. The experimental design provides direct tests distinguishing the two theoretical mechanisms through randomized assignment to AI-guided Socratic dialogue compared to appropriate control conditions.

The argument-quality hypothesis predicts joint changes. Namely, the AI-Socratic dialogue would simultaneously improve argument thoughtfulness and moderate attitudes. Formally, let q_i measure argument quality. The hypothesis predicts $\mathbb{E}[q_i \mid T_i = 1] > \mathbb{E}[q_i \mid T_i = 0]$ and $\mathbb{E}[|\theta^1_i - \theta_0| \mid T_i = 1] < \mathbb{E}[|\theta^1_i - \theta_0| \mid T_i = 0]$, where these effects operate through enhanced reasoning quality improving substantive beliefs ($\Delta\mu_i > 0$). The argument-quality mechanism makes no clear predictions about metacognitive changes in confidence or perspective-taking, as these concern awareness of epistemic limitations rather than reasoning quality.

The epistemic openness hypothesis makes distinct predictions. First, attitude moderation can occur independently of argument quality improvements: $\mathbb{E}[|\theta^1_i - \theta_0| \mid T_i = 1] < \mathbb{E}[|\theta^1_i - \theta_0| \mid T_i = 0]$ even when $\mathbb{E}[q_i \mid T_i = 1] \approx \mathbb{E}[q_i \mid T_i = 0]$. The mechanism operates through reduced perceived precision ($\Delta\omega_i > 0 \Rightarrow \Delta\hat{\tau}_i < 0$) rather than enhanced reasoning quality. Second, from equations (1) and (3), the intervention would directly reduce confidence in political positions and increase weight on others' perspectives: $\mathbb{E}[c_i \mid T_i = 1] < \mathbb{E}[c_i \mid T_i = 0]$ and $\mathbb{E}[w^{\text{others}}_i \mid T_i = 1] > \mathbb{E}[w^{\text{others}}_i \mid T_i = 0]$, where reduced confidence reflects recognition that information sources are correlated rather than learning that one's substantive position is incorrect, and increased weight on others' perspectives manifests as greater consideration of alternative viewpoints and diverse arguments.

Third, the heterogeneous treatment effect predictions from equations (4)–(9) provide further tests. The epistemic openness mechanism predicts effects concentrate among individuals in high-correlation, low-awareness environments (low cross-partisan exposure E_i, high elite cue susceptibility α_i, high reputation concerns S_i), moderate-confidence individuals (inverted-U relationship with baseline c^0_i), and on issues with greater heterogeneity in underlying priors (higher $\sigma^2_{\mu,j}$). These pat-

terns would be difficult to explain under the argument-quality hypothesis, which predicts stronger effects for high-ability individuals and makes no clear predictions about information environment characteristics or prior distributions. Importantly, while the experiment does not directly measure correlation perception, the conjunction of reduced confidence and increased perspective-taking without corresponding improvements in argument quality provides indirect evidence consistent with the correlation-awareness pathway rather than enhanced reasoning.

3.4 Implications for political polarization

Evidence favoring the epistemic openness hypothesis grounded in correlation neglect carries important implications for understanding political polarization and designing interventions to reduce it.

First, the framework suggests that interventions prioritize revealing the correlated structure of individuals' information sources rather than focusing on improving reasoning skills or providing new factual information. This perspective offers a behavioral foundation for political overconfidence that goes beyond generic miscalibration, identifying correlation neglect as a specific cognitive friction that structured reflection can address. Simply providing fact-checks or high-quality information adds new signals without correcting the fundamental error of treating correlated signals as independent. If individuals already overweight existing information due to correlation neglect, additional signals may be discounted or assimilated in ways that preserve existing positions. In contrast, interventions increasing correlation awareness can reduce overconfidence without providing new substantive content, creating psychological space for moderation.

Second, the framework helps explain why traditional interventions often fail to reduce polarization. Exposure to diverse viewpoints can backfire when individuals with high perceived precision encounter counter-attitudinal information. From equation (2), individuals with inflated $\hat{\tau}_i$ place excessive weight on their subjective mean μ_i, making them resistant to alternative perspectives they perceive as biased or uninformed. Reducing overconfidence first —by revealing correlation among existing sources through Socratic questioning— may be necessary before cross-partisan exposure can productively influence beliefs. The framework thus suggests a two-stage approach: epistemic interventions to reduce overconfidence, followed by informational interventions to update substantive beliefs.

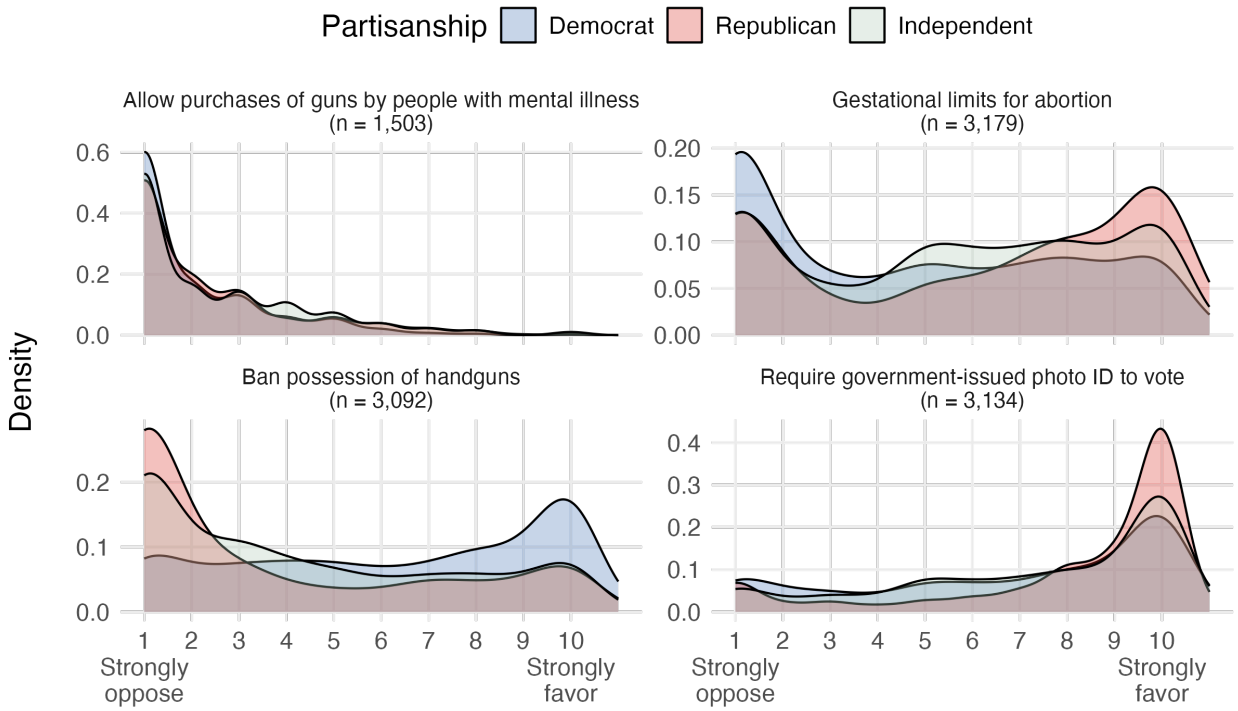
Third, the AI-facilitated Socratic method appears particularly well-suited to increasing correlation awareness because it can systematically prompt individuals to examine the origins and interrelationships of their beliefs. By asking participants to explain how they know something or to identify sources supporting their views, the intervention naturally reveals patterns of correlation without directly challenging positions. The patient, structured nature of AI dialogue allows thorough exploration of information sources in ways that organic conversation might not achieve, while the non-judgmental framing reduces defensiveness. This positions participants as agents of their own epistemic recalibration rather than targets of external correction.

4 Context: Contentious topics in the U.S.

To test the conditions under which AI-guided Socratic dialogue can foster political moderation, I examine four contentious policy issues that vary systematically in their partisan polarization and, therefore, potential for cross-cutting agreement. The issues span gun policy, reproductive rights, and voting procedures, domains that have generated sustained political conflict in contemporary U.S. politics.

The issues also differ in their identity-based versus policy-based foundations. Gun rights and abortion involve deeply held moral and constitutional beliefs tied to partisan identity, while mental health screening and voter ID requirements may be more amenable to pragmatic consideration. This variation enables testing whether AI-guided reasoning interventions work differently for identity-protective versus instrumental political attitudes. Crucially, all four issues involve genuine trade-offs and reasonable disagreement rather than simple factual disputes. This ensures that moderation effects, if observed, reflect substantive attitude change rather than mere correction of misinformation.

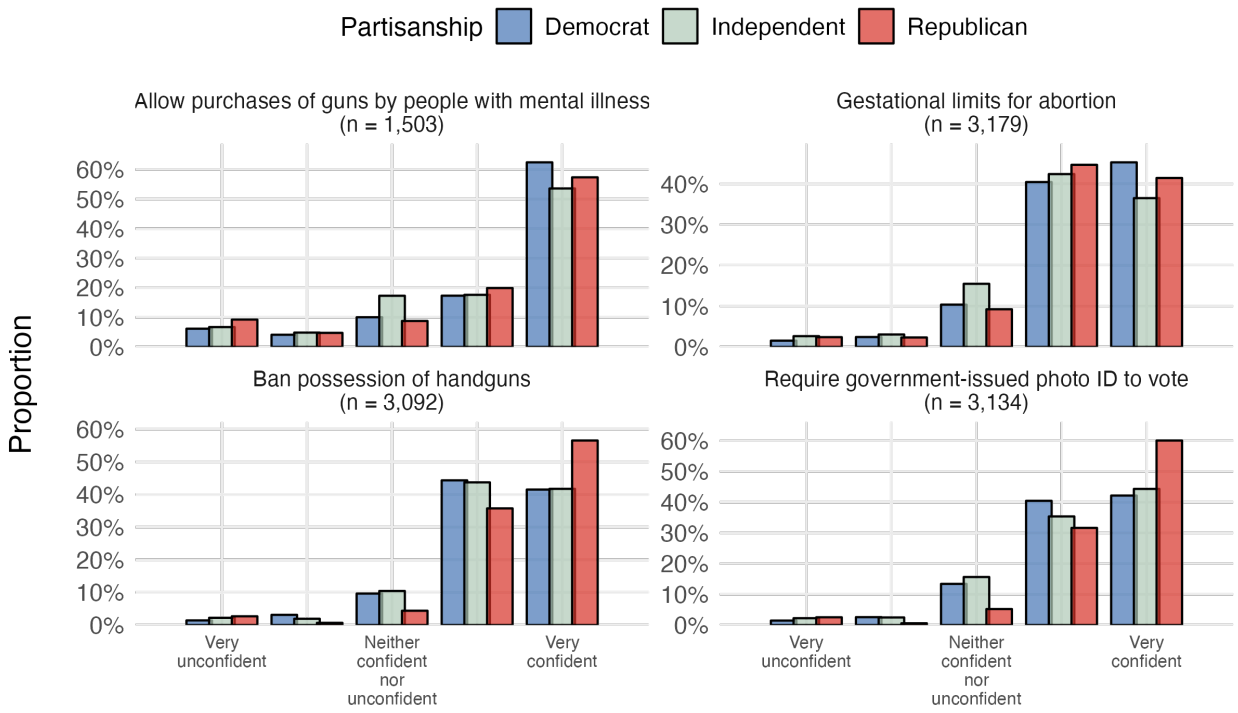
Figure 1: Pre-treatment opinion distributions by topic and partisanship



Notes: For each policy question, responses range from "Strongly oppose" (0) to "Strongly favor" (10). Densities are disaggregated by self-reported partisanship as reported by respondents to the crowdsourcing platform and verified through the survey instrument.

Gun policy and mental health. The first experiment centers on mental health-based gun purchase regulations, where participants face the question: "Would you favor or oppose allowing

Figure 2: Pre-treatment confidence distributions by topic



Notes: Each bar indicates the proportion of users with each given confidence level, ranging from very unconfident in red to very confident in blue.

people with mental illnesses to purchase guns?" This issue exemplifies the challenges of online political discourse by involving complex trade-offs between individual rights, public safety, and stigma considerations that resist simple partisan framing.

Current U.S. federal law restricts firearm access only for individuals who have been adjudicated as mentally ill or involuntarily committed to psychiatric treatment, affecting approximately 3,300 gun purchases annually from a potential pool of 59.3 million adults experiencing mental illness (Federal Bureau of Investigation, 2025; Substance Abuse and Mental Health Services Administration, 2023). This narrow restriction contrasts with broader public support for mental health screening, with polls showing 60-90% approval across party lines (Pew Research Center, 2021). However, the policy details reveal significant complexity around implementation, effectiveness, and civil liberties concerns that create space for reasoned disagreement.

As shown in Figure 1, there is ample cross-partisan overlap in participants' pre-treatment positions. Democrats, Republicans, and Independent-leaning voters alike concentrate around strong opposition to allowing gun purchases by people with mental health conditions. Figure 2 indicates that over 70% of participants, across partisan lines, reported feeling very confident about their positions, while only about 10% reported feeling somewhat unconfident.

Gestational limits for abortion. The second experiment includes reproductive rights policy, specifically around gestational limits: "Would you favor or oppose considering how long a woman has been pregnant in determining whether it is legal or illegal to have an abortion?" Unlike the binary pro-choice/pro-life framing that dominates partisan discourse, this question focuses on the nuanced middle ground where most Americans hold complex views. Both unconditional pro-choice and pro-life sympathizers may find common ground based on arguments about fetal development, maternal health, or practical implementation concerns.

Polling data reveals substantial cross-partisan overlap on gestational limits, with majorities supporting early abortion access while favoring restrictions in later stages of pregnancy (Gallup, 2025). The baseline distribution of respondents confirms this pattern, as there is ample overlap among the two partisan groups and independents on both sides of the spectrum, creating a bimodal distribution. While confidence remains high, with only 10-20% of respondents indicating they feel somewhat unconfident, "Confident" rather than "Very confident" is the most frequently chosen category.

Handgun bans. The second experiment also examines handgun possession bans: "Would you favor or oppose banning the possession of handguns, except by the police and other authorized persons?" This question represents a more traditionally polarized gun policy issue, with stark partisan divisions and limited middle ground in public opinion (Pew Research Center, 2021). As shown in Figure 1, we observe clear separation between Republicans and Democrats, producing two highly concentrated poles. Self-reported Independents lean toward opposing handgun bans. Confidence is very high, with the "confident" and "very confident" categories accounting for more than 80% of responses in roughly equal proportions, with independents being relatively less confident.

The inclusion of this highly polarized topic serves as a boundary condition for the intervention. If AI-guided Socratic dialogue can foster moderation even on issues with strong partisan identity attachments and limited cross-cutting appeal, it would suggest broad applicability for the intervention approach.

Photo ID Voting Requirements. Finally, the second experiment includes photo ID voting requirements: "Would you favor or oppose requiring all voters to show government-issued photo identification to vote?" This issue involves fundamental questions about democratic participation, election integrity, and voting access that generate significant partisan disagreement while maintaining some cross-cutting elements. The issue's framing also creates opportunities for nuanced discussion, as both sides can appeal to shared democratic values while disagreeing on implementation details.

Polling shows strong partisan divisions on voter ID requirements, with Republicans overwhelmingly supporting such measures (91% favor) while Democrats are more divided (48% favor) (Pew Research Center, 2021). This pattern holds for the experimental sample, showing high concentrations of Republican and self-reported Independent participants in the right-hand tail, while Democrats uniformly distribute across the spectrum, with a small jump at "Strongly favor." As

with the previous topics, the vast majority of participants feel confident or very confident about their policy positions, with a particular concentration in the latter category among Republicans.

5 Experimental design

This section describes the experimental design for two independent experiments. The first intervention evaluates the effects of AI-guided Socratic dialogue on political moderation against other active conditions that require user interaction and non-active conditions that require participants to passively wait (i.e., waiting control). The second experiment tests generalizability across multiple topics and examines persistence and spillover effects across topics. Figure 3 presents a diagram of the full experimental design.

Building on Experiment 1's findings, Experiment 2 employs a streamlined design with only two treatment arms: AI-Socratic dialogue and AI-Grammar correction. This simplification is motivated by several considerations from the initial experiment's results, as detailed in the preregistered analysis. First, as an active condition requiring participant engagement, AI-Socratic dialogue showed superior moderation effects, making it the preferred active treatment for scaling analysis. Second, the AI-emotional regulation condition showed minimal effects across all outcomes, while the waiting control provided limited insight into AI-specific mechanisms. The AI-Grammar correction condition emerged as the optimal control because it provides an active placebo that controls for AI interaction effects —including potential novelty, attention, and demand effects— while avoiding the cognitive engagement that characterizes Socratic questioning. This design choice enables cleaner identification of the specific mechanisms underlying AI-guided reflection by holding constant the experience of conversing with an AI system while varying only the type of guidance provided. Additionally, the simplified design reduces multiple testing concerns and increases statistical power for detecting treatment effects across the three diverse political topics examined in Experiment 2.

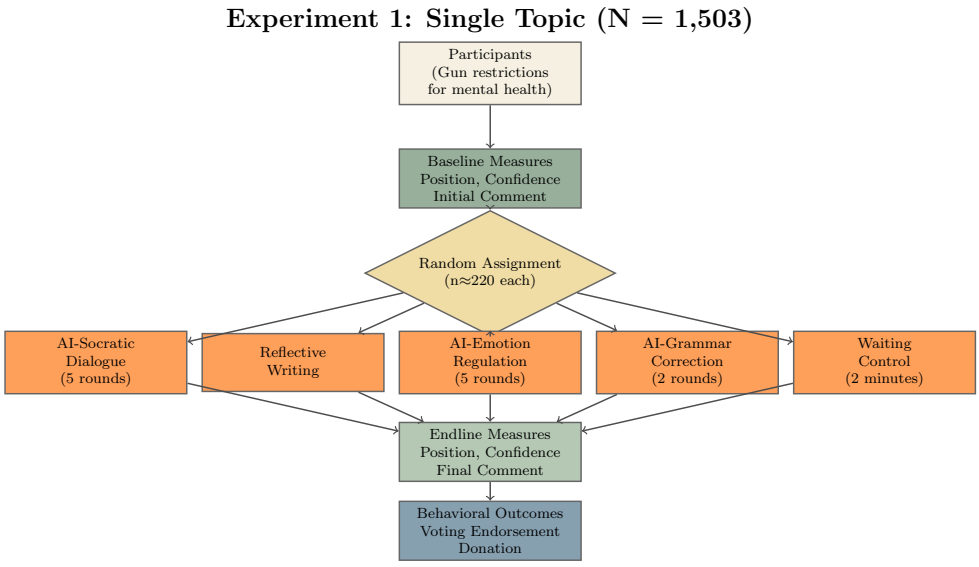
5.1 Experiment 1: AI-guided Socratic dialogue

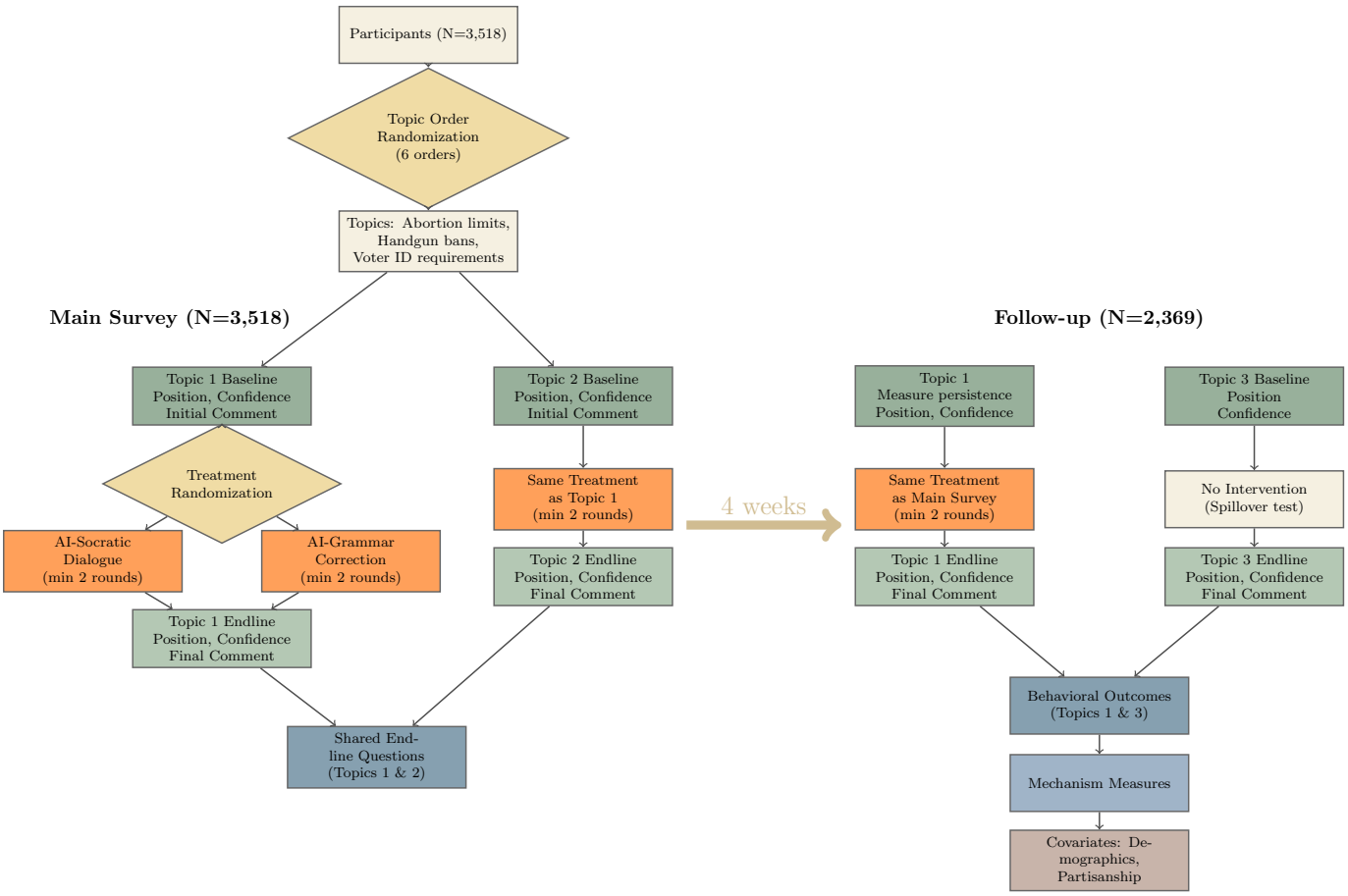
The first experiment employs a within- and between-subjects randomized design with 1,500 participants recruited through Prolific and directed to `deliberation.io`.[1] Participants shared their opinions on gun purchase regulations for individuals with mental health conditions through a structured, indirect conversation.

In the baseline period, participants share their initial position values and confidence level using a Likert scale, and initial argument as an open text question. They also respond to factual questions about gun regulations, as well as cognitive reflection test and a emotional state questions (see Section E.1). In the endline period, after completing the AI- or non-AI module, participants share their updated policy position, confidence level, and endline argument. Then, participants

[1] I ultimately analyze responses from 1,503 participants after excluding those whose personal characteristics were not reported by Prolific, did not complete the survey, or registered high latency and abnormal response times (i.e., over the 95th percentile of each treatment arm).

Figure 3: Experimental designs for AI-Socratic dialogue





respond to cognitive reflection test and emotional state questions, donation and endorsement behavior, attitudes toward the process, including agency, and finalize with a couple of quesitons on demographics (see Section E.2).

Participants are randomly assigned with equal probability to one of five treatment arms (See Section A for the LLM prompts and participant instructions for each treatment condition).

1. **AI-Socratic dialogue**: Participants engage in a five-round conversation with an LLM (Open AI's `gpt-4o-mini-2024-07-18`) that used the Socratic method, asking open-ended questions to probe their reasoning without providing direct arguments or opinions. The LLM was prompt-tuned and previously tested to avoid providing answers, opinions, or additional substantive input on the topic. The platform requires users to participate in five mandatory rounds.

2. **Reflective Writing**: Participants are asked to write a 4-5 sentence, structured paragraph elaborating on their position without AI assistance, serving as a self-guided reflection baseline to isolate the effect of AI interaction from general reflection.

3. **AI-Emotion Regulation**: Participants interact with an LLM focused on identifying and managing emotional responses to the topic. The LLM guided users to identify feelings about the issue and provided basic emotion-regulation strategies, such as reappraisal of the situation or taking a neutral perspective. The platform requires users to engage in five mandatory rounds.

4. **AI-Grammar Correction**: Participants receive only stylistic and grammatical feedback from an LLM, serving as a placebo to control for AI interaction effects. The LLM was prompt-tuned to avoid providing feedback on argument content, logic, or depth. The platform requires users to participate in two mandatory rounds.

5. **Waiting Control**: Participants wait for two minutes (the approximately time to complete AI conditions) without any task, controlling for time and potential incubation effects.

Following the treatment module, participants react to twelve seeded comments that vary in their degree of support for the initial policy premise. This comment endorsement exercise enables measurement of polarized information consumption patterns by capturing systematic preferences for ideologically aligned content. These comments were obtained from a demographically representative survey of 300 U.S. participants, balanced across ethnic, age, gender, and partisan characteristics. The comments span supportive, neutral, and opposing positions, as detailed in Section B.0.1.

5.2 Experiment 2: Multi-topic and Follow-up

The second experiment employs a two-stage design with approximately 3,500 participants to examine generalizability, persistence, and spillover effects.[2] Participants of Experiment 1 were excluded

[2]The final sample, after filtering out participants that did not complete the survey or registered high latency (i.e., completion times above the 95th percentile by treatment arm), is 3,518.

from participating in Experiment 2.

Participants were randomized to one of six topic orders based on the three policy topics: abortion gestational limits (bimodal but cross-partisan), handgun possession bans (bimodal and partisan), and voter ID requirements (strong Republican priors, weak Democratic priors).

In the first stage ("main survey"), participants covered two of the three topics, where they were asked to complete the assigned LLM module in each of them. After completing the baseline questionnaire, participants were randomized into AI-Socratic dialogue or AI-Grammar correction conditions, with a minimum of two mandatory rounds.

Participants then report their posterior beliefs and complete a comment endorsement exercise designed to identify extreme endorsement behavior. Similar to Experiment 1, in this task, they react to nine seeded comments that vary in their degree of support for the initial policy premise, allowing for the construction of voting scores that detect systematic endorsement of ideologically extreme positions. These comments were drawn from the same demographically representative survey of 300 U.S. participants, with balanced representation across ethnic, age, gender, and partisan characteristics. The comments encompass supportive, neutral, and opposing positions, as detailed in Section B.0.1.

In the second stage ("follow-up survey"), users revisit the first topic and complete the same AI-Grammar or AI-Socratic module they were assigned in the main survey. The objective is to assess persistence of the main survey's effects one month after completion by comparing baseline positions before the follow-up treatment. Then, participants move to the third topic, which they have not previously encountered. In this part, rather than being exposed to either treatment condition, I simply measure spillover effects by comparing baseline positions, confidence levels, and voting endorsement behavior between participants who were previously treated with AI-Socratic dialogue versus those who received AI-Grammar correction in the main survey.

Within each topic order, participants are randomly assigned to either the AI-Socratic dialogue or the AI-grammar correction condition. The design enables measurement of: (1) immediate effects on the first and second topics, (2) persistence effects at four-week follow-up, and (3) delayed spillover effects on the third topic.

At the end of the follow-up stage, in addition to being asked about partisanship, participants respond to 5 questions to measure individual characteristics regarding cross-partisan interactions, consumption of information, social conformity concerns, elite cues, and belief confidence.

5.3 Outcomes and measurement

Through the experimental design, the goal is to capture multiple dimensions of political attitude and behavioral change, as well as argumentative and conversational quality. All outcomes were preregistered, with measurement occurring at baseline, immediately post-treatment, and at four-week follow-up in Experiment 2. The design enables analysis of both individual-level changes and between-subject comparisons across randomized conditions.

5.3.1 Primary outcomes

Position moderation: I measure political positions for each topic on 10-point Likert scales before and after interventions.[3] I examine two primary outcomes related to position moderation that capture different aspects of polarization reduction.

The first outcome is *post-treatment extremity*, which directly measures whether participants maintain extreme positions after the intervention. This is operationalized as a binary indicator equal to one if the participant's post-treatment position falls in the extreme ranges (position value ≤ 2 or position value ≥ 8).

$$\text{Post-Extremity}_i = \mathbb{1}\{\text{Position}_{i,\text{post}} \in [1,2] \cup [8,10]\}$$

The second outcome is *movement toward the median*, which measures whether participants with initially extreme positions move closer to their topic's baseline median. I operationalize this as a binary indicator equal to one if the absolute distance between the post-treatment position and the baseline median is less than the absolute distance between the pre-treatment position and the baseline median. Since opinion distributions are not necessarily bimodal, this approach better captures moderation compared to simply moving toward the center of the scale (i.e., position value of 5).

$$\text{Move-to-Median}_i = \mathbb{1}\{|\text{Position}_{i,\text{post}} - \text{Median baseline}| < |\text{Position}_{i,\text{pre}} - \text{Median baseline}|\}$$

I focus the analysis on participants initially holding strong positions (position value ≤ 2 or position value ≥ 8), as these individuals have the greatest scope for moderation and represent the populations most relevant for understanding polarization reduction. More formally, the extreme participant indicator is defined as follows:

$$\text{Extreme}_{i,\text{pre}} = \mathbb{1}\{\text{Position}_{i,\text{pre}} \in [1,2] \cup [8,10]\}$$

Cross-partisan donation: I measure willingness to donate to organizations that bridge political divides as an incentive-compatible behavioral outcome. In experiment 2, participants could either keep or donate a \$1 bonus payment to Braver Angels, a cross-partisan organization dedicated to bridging partisan divides through structured dialogue between liberal and conservative representatives. This behavioral measure captures participants' revealed preferences for cross-partisan cooperation beyond self-reported attitude changes. The amount of the donation represents about 35% of the total payment for the survey completion.

Extreme endorsement: To measure issue polarization in information consumption, I construct an indicator for extreme endorsement patterns based on participants' voting behavior toward seeded comments representing different ideological positions. I calculate the voting score as a weighted

[3] While Experiment 1 included a scale ranging from -5 to 5, I re-scale it for comparability.

sum of votes, where approval corresponds to +1, disapproval to −1, and passing to 0, weighted by each seeded comment's position along the policy continuum (see Section B in the Appendix). The indicator equals one if the rescaled score is ≤ 0.20 or ≥ 0.80, identifying participants who systematically endorse comments from one ideological extreme. This measure captures behavioral polarization in comment endorsement patterns while maintaining comparability with standard position scales.

Argument quality: I assess comment thoughtfulness through structured rubrics evaluated by both human annotators and fine-tuned LLMs (i.e., Open AI's `gpt-4o-2024-08-06` model). Quality measures include depth of reasoning, perspective diversity, and logical coherence, rated on 3- to 6-point scales with high inter-rater reliability (ICC > 0.75). 100 college-educated annotators, blind to treatment assignment, evaluated a random subset of comments using detailed scoring guidelines (see Appendix G). The LLM evaluation system was trained on the human-coded subset (≈150 comments), evaluated on a testing set (≈150 comments), and then applied consistently across all comments. Each LLM was trained on a single rubric dimension to ensure focused and reliable assessment, and treatment assignment was consistently excluded from all models to prevent bias in quality evaluation. In addition to providing a score, the LLM is prompted to include a chain of thought justifying the specific value chosen in light of the rubric system.

5.3.2 Secondary outcomes

Belief confidence: I assess confidence in political positions on 5-point scales to test whether interventions operate through increased epistemic humility rather than substantive persuasion. Participants report how certain they are about their positions both before and after the intervention. Reduced confidence combined with position moderation would support the theoretical mechanism that AI-Socratic dialogue works by revealing gaps in reasoning rather than providing superior arguments.

Conversation perceptions and downstream effects: I measure participants' assessments of the AI intervention's impact on their perceived neutrality and further political engagement. These measures address concerns about AI manipulation and user agency while testing predictions from the epistemic openness mechanism. Specifically, I examine: (1) perceived bias of both the AI conversational agent and research team, assessed through 5-point Likert scales measuring whether participants detected liberal, conservative, or neutral positioning; (2) perceived agency and pressure, measuring whether participants felt coerced to modify their positions or maintained autonomy throughout the conversation; (3) openness to alternative perspectives, capturing participants' self-reported willingness to consider viewpoints different from their own; and (4) information-seeking behavior regarding both the specific deliberation topics and other contentious political issues. These measures are crucial for distinguishing genuine attitude change from demand effects or experimenter bias, for ensuring that the intervention preserves user autonomy, and for testing whether enhanced epistemic openness extends beyond the immediate conversational context to broader political en-

gagement.

5.4 Empirical strategy

Experiment 1: Single-topic analysis

To ensure robustness of the findings, I estimate Ordinary Least Squares together with three different specifications that address potential model misspecification. The use of multiple estimation strategies serves several purposes. First, it provides robustness checks against potential violations of identification assumptions. Second, it allows assessment of whether treatment effects are sensitive to different approaches for handling covariates and selection (i.e., model misspecification). Third, the comparison between estimated and known propensity score models provides a diagnostic for the quality of the propensity score estimation. Overall, results indicate consistency across specifications.

Model 1: OLS that estimates treatment effects conditional on initial position extremity:

$$Y_{i,\text{post}} = \beta_0 + \sum_{k=1}^{4} \beta_k T_{ik} + \beta_5 1[\text{Extreme at pre}_i] + \sum_{k=1}^{4} \beta_{k+5}(T_{ik} \times 1[\text{Extreme at pre}_i]) + \gamma Y_{i,\text{pre}} + \tilde{\mathbb{X}}_i' \boldsymbol{\delta} + \epsilon_i, \tag{10}$$

where $Y_{i,\text{post}}$ represents post-treatment outcomes including position changes, argument quality measures, and confidence levels. T_{ik} are indicators for the four active treatments (AI-Socratic dialogue, reflective writing, AI-emotion regulation, AI-grammar correction) with waiting control as the omitted category. Extreme$_i$ indicates initial positions with values ≤ 2 or ≥ 8 on the 10-point scale. $Y_{i,\text{pre}}$ controls for baseline outcomes. $\tilde{\mathbb{X}}_i$ includes covariates, de-meaned and interacted with treatment assignment. The coefficient β_{k+5} captures the differential treatment effect for extreme participants, with β_1, β_5, and β_5 representing the primary parameters of interest. Particularly interesting is the comparison against β_5, which captures the post-treatment position of extreme participants in the control condition. Standard errors are heteroskedasticity-robust.

Model 2: Post-Double-Selection LASSO (PDS-LASSO). Following Belloni et al. (2014), this approach first uses LASSO regularization to select relevant covariates for both the outcome equation and auxiliary selection equations, then estimates treatment effects using OLS with the selected controls. This method provides protection against model misspecification while maintaining valid inference.

Model 3: Augmented Inverse Propensity Weighting (AIPW). This doubly robust estimator combines propensity score weighting with outcome regression to provide consistent estimates when either the propensity score model or the outcome model is correctly specified (Robins et al., 1994; Bang and Robins, 2005; Athey et al., 2025b). For each treatment arm k, the AIPW estimator is:

$$\hat{\tau}k^{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_k(X_i) - \hat{\mu}_0(X_i) + \frac{T_{ik}}{\hat{e}_k(X_i)}(Y_i - \hat{\mu}_k(X_i)) - \frac{T_{i0}}{\hat{e}_0(X_i)}(Y_i - \hat{\mu}_0(X_i)) \right] \tag{11}$$

where $\hat{\mu}_k(X_i)$ represents the predicted outcome under treatment k, $\hat{e}_k(X_i)$ is the estimated propen-

sity score for treatment k, and T_{ik} and T_{i0} are treatment indicators. I estimate propensity scores using multinomial logistic regression and outcome models using linear regression, both including all available covariates.

Model 4: Known Propensity AIPW. This specification uses the experimental assignment probabilities as the true propensity scores rather than estimating them from observed data. Since treatment assignment was randomized, the known assignment probabilities provide the most accurate propensity weights. The estimator follows Equation 11 but replaces $\hat{e}_k(X_i)$ with the known randomization probabilities q_k:

$$\hat{\tau}k^{Known} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_k(X_i) - \hat{\mu}_0(X_i) + \frac{T_{ik}}{q_k}(Y_i - \hat{\mu}_k(X_i)) - \frac{T_{i0}}{q_0}(Y_i - \hat{\mu}_0(X_i))\right] \tag{12}$$

This approach provides the most efficient estimates under randomization while still allowing for covariate adjustment in the outcome models. The known assignment probabilities are: Waiting Control (18.90%), AI-Socratic dialogue (19.09%), Reflective paragraph (20.16%), AI-Emotional regulation (21.96%), and AI-Grammar correction (19.89%).

AIPW estimators focus on the subset of participants with extreme initial positions, as these individuals represent the primary population of interest for studying polarization reduction.

Experiment 2: Multi-topic analysis

For the multi-topic design examining generalizability across issues, I pool observations and estimate:

$$\begin{aligned} Y_{it,\text{post}} =& \beta_1 \text{Socratic}_i + \beta_2 \text{Extreme at pre}_i + \beta_3 \text{Topic}_{it} \\ &+ \beta_4(\text{Socratic}_i \times \text{Extreme at pre}_i) + \beta_5(\text{Socratic}_i \times \text{Topic}_{it}) \\ &+ \beta_6(\text{Extreme at pre}_i \times \text{Topic}_{it}) \\ &+ \gamma Y_{it,\text{pre}} + \tilde{\mathbb{X}}'_{it}\boldsymbol{\delta} + \boldsymbol{\phi}_i + \varepsilon_{it}, \end{aligned} \tag{13}$$

where Y_{it} represents outcomes for individual i on topic t. $\boldsymbol{\phi}_i$ are topic-order fixed effects. As indicated in the preregistered analysis, $\tilde{\mathbb{X}}_{it}$ includes selected covariates, de-meaned and interacted by treatment and topic indicators. Standard errors, ε_{it}, are clustered at the participant level to account for within-individual correlation across topics.

Similar to Experiment 1, I employ the same four estimation strategies to ensure robustness of findings across the multi-topic design. To do so, I adapt the AIPW approach to the multi-topic setting, where propensity scores are estimated using multinomial logistic regression that includes topic indicators and topic-varying covariates. The AIPW estimator accounts for the fact that treatment assignment occurs at the individual level while outcomes are measured at the individual-topic level. For Known Propensity AIPW, the specification uses the empirical experimental assignment probabilities (48.95% AI-Socratic dialogue, 51.05% AI-Grammar correction) as known propensity scores.

For individual-level outcomes in Experiment 2, I estimate:

$$D_i = \alpha_0 + \alpha_1 \text{Socratic}_i + \alpha_2 \text{Extreme}_i + \alpha_3 (\text{Socratic}_i \times \text{Extreme}_i) + \tilde{\mathbb{Z}}_i' \boldsymbol{\phi} + \nu_i, \tag{14}$$

where D_i represents donation and aggregate voting behavior. The intercept corresponds to the omitted category—non-extreme participants in the AI-Grammar condition. Standard errors, ν_i, are heteroskedasticity-robust.

As also pre-specified, given the treatment effects of the AI-Socratic dialogue intervention found in Experiment 2, I report one-sided p-values for the directional hypotheses.

5.5 Balance and attrition

Randomization balance. The randomization procedure successfully achieved covariate balance across treatment conditions in both experiments. For Experiment 1, joint tests across five treatment arms (emotional, grammar, paragraph, socratic, and control) indicate no systematic differences in observable characteristics ($F(66) = 0.82, p = 0.845$). The distribution of baseline covariates, including demographic characteristics, socioeconomic status, political attitudes, cognitive skills, and prior platform experience, does not differ significantly across treatment assignments.

Individual covariate analysis reveals that standardized differences remain well within conventional bounds, with most falling below 0.10 standardized mean differences. Notable exceptions include mixed ethnicity indicators (-0.22 to -0.11 across treatments), income in the $25,000-$49,999 bracket (-0.23 for emotional treatment), and certain occupational categories (transportation/logistics: 0.23 for socratic treatment; retail sales: -0.17 for socratic treatment). These imbalances, while statistically modest and below the standard 0.25 threshold, are addressed through covariate adjustment in all regression specifications (i.e., are unpenalized in PDS-LASSO specifications following Belloni et al. 2014).

The two-arm design in Experiment 2 similarly achieves satisfactory randomization balance ($F(44) = 1.12, p = 0.272$). Pre-treatment covariates exhibit minimal standardized differences between socratic and grammar conditions, with most differences below 0.08 standard deviations. Cross-party interaction frequency and reputation concerns show slightly larger but still acceptable imbalances (standardized differences of $0.05 - 0.08$).

Experiment 2 follow-up attrition. I do not find evidence of differential attrition. More specifically, the follow-up participation rate shows no systematic relationship with treatment assignment ($F(51) = 1.17, p = 0.196$). While participants with higher crowdsource platform experience and baseline confidence levels demonstrate marginally higher follow-up rates (standardized differences of 0.20 and 0.42-0.87 respectively across confidence categories), these patterns are consistent across both treatment arms, preserving internal validity. Furthermore, to address potential concerns about selective attrition, I include the predicted participation probability (demeaned) as a covariate in all follow-up estimates. This indicator is obtained by regressing follow-up participation on the full

set of pre-treatment covariates.

Figures C1-C7 present detailed balance figures for standardized mean differences across treatment arms, divided by groups of individual attributes and priors. Figures C8-C10 show estimates for attrition in the follow-up round, examining both overall and treatment-specific differential attrition.

5.6 First stage: Engagement with the AI conversational agent

Both interventions were successfully implemented as designed with high compliance rates across experimental conditions. In Experiment 1, as shown in Figure 4, all participants in the AI-Grammar condition completed the mandatory two rounds. Participants in the other two active conditions, the AI-Socratic and AI-Emotional Regulation, completed all five rounds.

In Experiment 2, where only two rounds were mandatory across both conditions, participants exhibited differential engagement patterns consistent with treatment design. Those assigned to AI-Socratic dialogue engaged in an average of 4.27 rounds (SD = 1.12) with a median of five rounds, indicating substantial voluntary participation beyond the minimum requirement. As shown in Figure 4, 65% of participants in this condition completed all five available rounds. Conversely, participants in the AI-Grammar correction condition completed an average of 2.63 rounds (SD = 0.96) with a median of two rounds, with 62% participating only in the mandatory rounds. This differential engagement provides evidence that the Socratic questioning format successfully sustained participant interest and cognitive involvement.

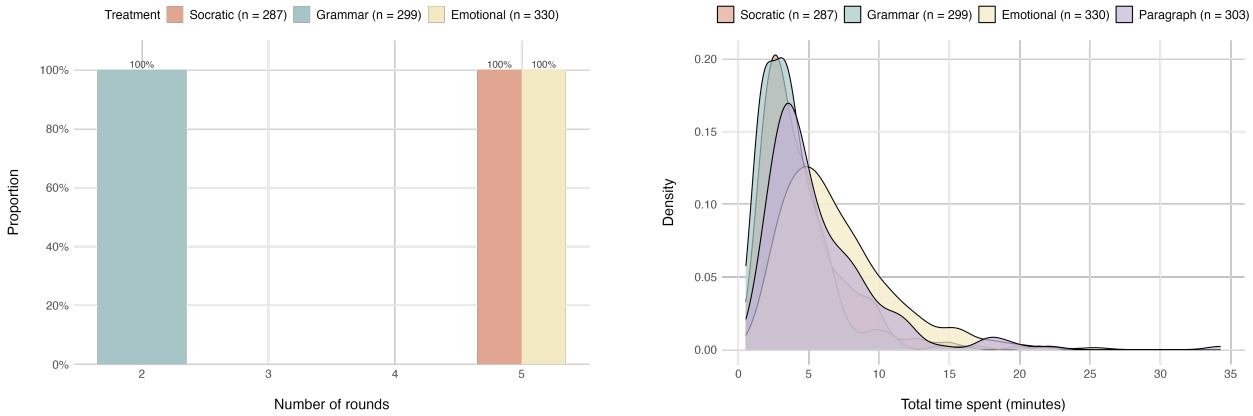
Temporal engagement patterns, displayed in the right panels of Figure 4, reveal particularly high participant investment in cognitively demanding conditions. In Experiment 1, participants in the AI-Socratic condition registered a median completion time of 4.43 minutes, while the median of those in the AI-Emotional regulation was 3.61 minutes. Participants in the reflective writing condition exhibited longer times, with a median of 6.02 minutes, confirming the distinct nature of the treatment arms. By design, control group participants waited exactly two minutes without further engagement.

Experiment 2 demonstrated lower overall engagement levels due to the reduced mandatory participation requirements, yet maintained clear differentiation between treatment conditions. Participants in the AI-Socratic condition registered a mean time of 3.81 minutes (SD = 2.4) compared to 1.66 minutes (SD = 0.71) for those in the AI-Grammar condition. Notably, the temporal distribution for AI-Socratic participants exhibited greater variance, which, combined with balanced observable characteristics across conditions, indicates higher voluntary engagement levels with the Socratic intervention.

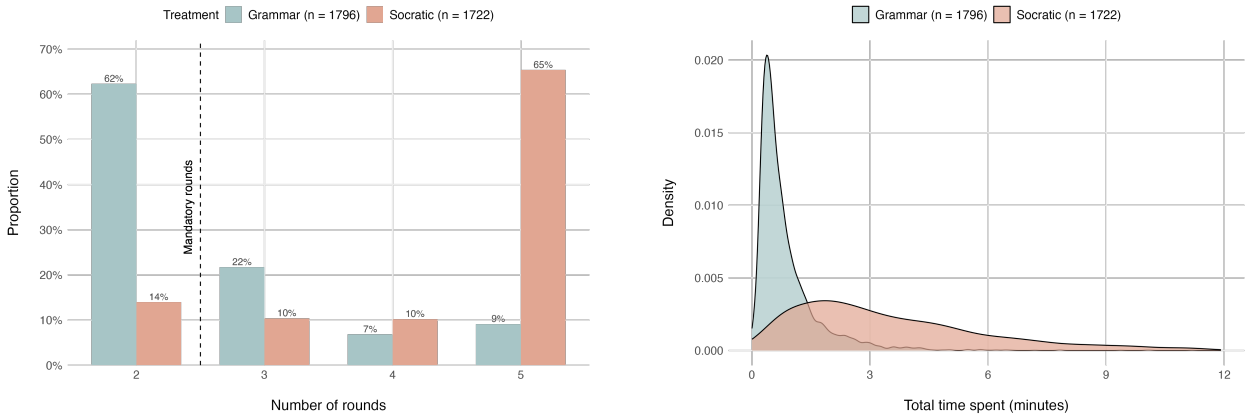
Response length patterns, presented in Appendix Table D1, vary systematically by treatment assignment in ways consistent with intervention design. Across both experiments, responses to the AI-Grammar conversational agent were systematically shorter, reflecting the nature of grammar-focused interactions. The reflective writing task generated the longest responses with mean and standard deviation of 53.4 and 31.3 words respectively, along with the highest variance, indicating

Figure 4: Experiment 1 and 2 first stage

Experiment 1



Experiment 2



substantial heterogeneity in participant elaboration. Importantly, median response lengths in Experiment 2 approximated those observed in the reflective paragraph condition, nearly doubling the median response length for the corresponding condition in Experiment 1. Overall, these patterns suggest that participants maintained substantive engagement through the intervention.

6 Results

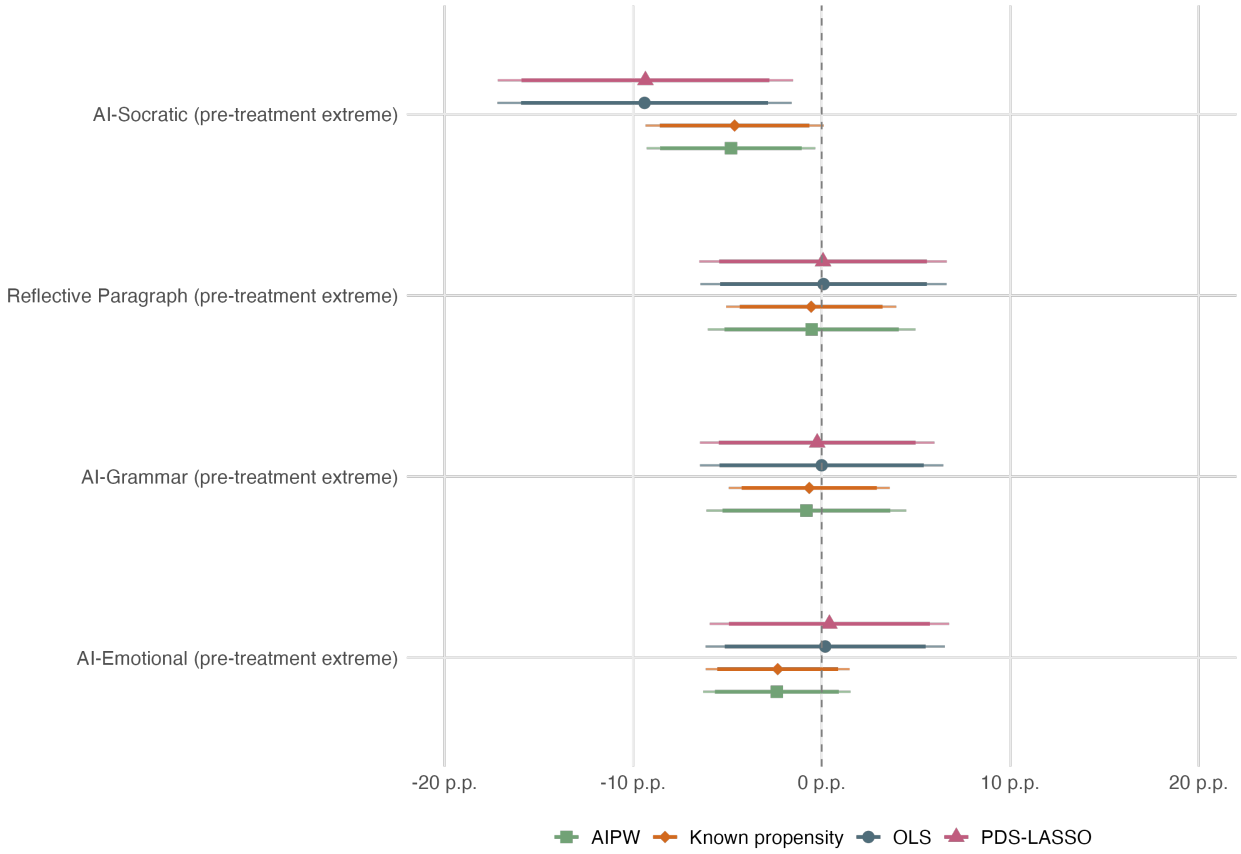
This section presents results from both experiments on position moderation and argument quality. I first present Experiment 1's findings comparing AI-Socratic dialogue to other active and non-active treatments. Then, I analyze Experiment 2's results across multiple topics in the light of the theoretical prediction that topics with larger baseline overlap will exhibit greater effects of the AI-Socratic condition. Next, I assess behavioral outcomes, mid-term persistence effects, and spillovers. Finally, I examine the theoretical mechanisms underlying the observed moderation effects.

6.1 Experiment 1: AI-Socratic dialogue vs. other treatments

6.1.1 Policy position moderation and confidence among extreme participants

Figure 5 provides self-reported positioning results of Experiment 1, focusing on treatment effects among participants with initially extreme positions. More specifically, the figure shows how each treatment arm performs relative to the waiting control among extreme participants. Appendix Table D.2.1 reports detailed regression results for both positioning and confidence outcomes.

Figure 5: Experiment 1: Conditional average treatment effects on position extremity among initially extreme participants



Notes: The figure indicates average treatment effects among extreme participants on a binary indicator of extreme positioning at post-treatment. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the waiting control condition. Full results including Waiting control are reported in Appendix Table D.2.1. Bars indicate 90 and 95% confidence intervals from two-sided tests, with heteroskedasticity-robust standard errors.

Position moderation effects. Figure 5 shows that AI-Socratic dialogue produces significant moderation effects compared to the waiting control baseline. Based on PDS-LASSO estimates, extreme participants in the AI-Socratic condition are 9.4 p.p. less likely to maintain extreme positions compared to extreme participants in the waiting control (SE = 3.98; $p < 0.05$). This effect represents approximately 20% of the control standard deviation (0.476), indicating a meaningful reduction in issue polarization. Estimates for AIPW and Known propensity models correspond to

about 50% of that magnitude and are also statistically significant different from zero. In contrast, other treatment arms show much smaller and statistically insignificant effects. With the exception of the AI-Emotional, for which AIPW and Known propensity estimates yield a negative but noisily estimates at around 2 p.p., other treatment arms show effects close to zero.

As shown in columns (3) and (4) of Appendix Table D.2.1, the moderation effects operate through systematic movement toward the baseline median rather than random position changes. Extreme participants in the AI-Socratic condition are 4.6 p.p. more likely to move toward the topic median compared to extreme participants in the waiting control (SE = 1.89; $p < 0.05$). Given the low baseline rate of movement toward the median among extreme participants (control mean = 4.9%), this represents nearly a doubling of the probability of centrist movement and corresponds to approximately 21% of the control standard deviation (0.217). This systematic pattern contrasts sharply with other treatment arms, which show negligible movement toward the median among extreme participants.

The pattern of results across treatment arms provides crucial insight into mechanisms underlying the moderation effects. While AI-Socratic dialogue consistently produces significant moderation, the other active treatments —AI-Emotional regulation, AI-Grammar correction, and Reflective Paragraph writing— show substantially smaller and statistically insignificant effects on both outcome measures. This specificity suggests that the particular questioning approach of Socratic dialogue, rather than general engagement with AI systems or reflective writing exercises, drives the observed attitude moderation.

The contrast is particularly notable with the Reflective Paragraph condition, which might be expected to produce similar effects through self-examination and articulation of reasoning. However, the reflective writing intervention shows virtually no movement towards the median (coefficient = −0.002) and minimal reduction in extremity (−1.9 p.p.), suggesting that structured questioning by an external agent may be more effective than self-directed reflection.

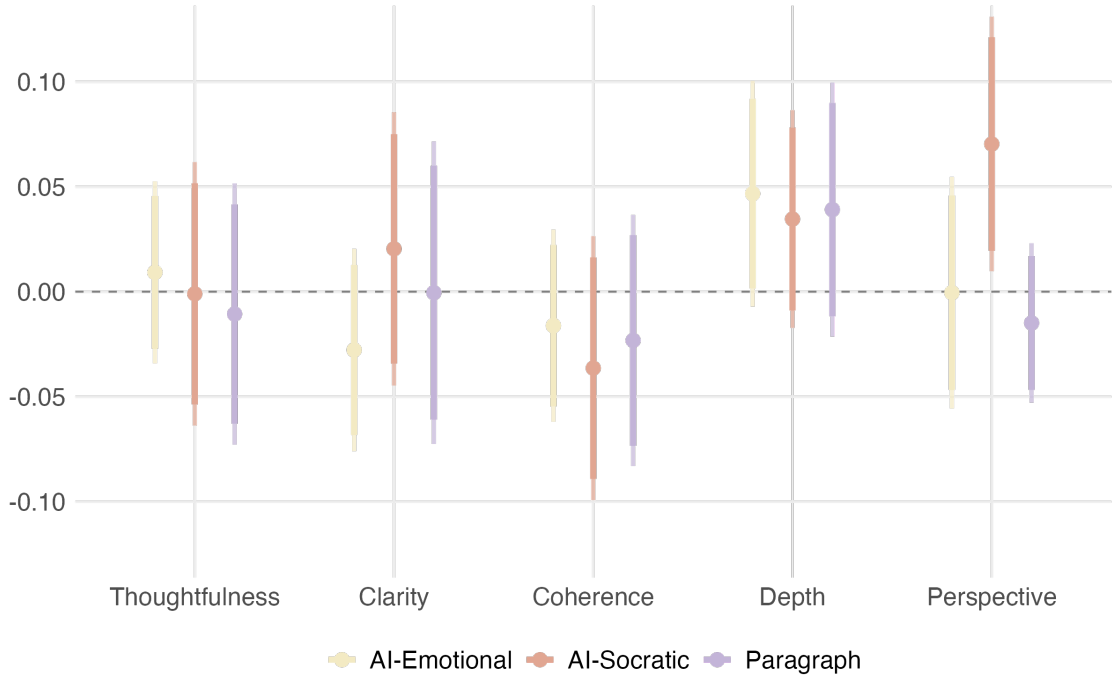
Confidence. Columns (5) and (6) of Appendix Table D.2.1 examine confidence at post-treatment. Overall, AI-Socratic participants show a significant decrease in confidence (−17.8 p.p., SE = 0.093; p = 0.056), representing approximately 0.14 SDs of the control group (1.295). Among extreme participants specifically, the AI-Socratic intervention produces a confidence reduction of 15.7 p.p. (SE = 0.092; p = 0.089), or about 12% of the control standard deviation. While these effects are modest in magnitude and only marginally significant, they suggest that the intervention works in part by revealing gaps in participants' knowledge through systematic questioning. Notably, the confidence reduction is specific to AI-Socratic dialogue, with other treatment arms showing smaller and statistically insignificant effects.

6.1.2 Effects on argument quality

Figure 6 and Appendix Table D.3.1 present treatment effects on a comprehensive thoughtfulness score and four dimensions of argument quality: clarity, coherence, depth, and consideration of

diverse perspectives.

Figure 6: Experiment 1: Conditional average treatment effects on argumentative quality scores among extreme participants



Notes: The figure indicates average effects among extreme participants on standardized argumentative quality scores. Estimates are from PDS-LASSO. The horizontal dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the waiting control condition. AI-Grammar is omitted from this panel due to substantially larger coefficient magnitudes relative to other arms. Full results including AI-Grammar are reported in Appendix Table D.3.1 and Figure D3. Bars indicate 90 and 95% confidence intervals from two-sided tests, with heteroskedasticity-robust standard errors.

I find no significant effects of the AI-Socratic module on overall thoughtfulness scores—an overarching measure that captures whether arguments contain clear reasoning, evidence, and logical development and that is measured independently of the component scores. This is the case for both non-Extreme (0.013, SE = 0.063) and Extreme participants in the AI-Socratic dialogue (−0.001, SE = 0.074).

In contrast, the AI-Grammar correction condition reduced thoughtfulness by 0.52 SDs (SE = 0.078; $p < 0.01$) among non-Extreme participants and by 0.93 SDs among Extreme participants. This is likely because participants focused on surface-level editing rather than substantive improvement of their arguments. Participants appear to have allocated cognitive resources toward linguistic improvements at the expense of substantive reasoning.

Similarly, the Reflective Paragraph condition displays a significant decrease in thoughtfulness compared to non-extreme participants in the waiting control ($p < 0.10$). This is particularly true for non-extreme participants. The contrast with AI-Socratic dialogue is particularly revealing since the former condition lacks the structured guidance that characterizes the latter. These findings also suggest that participants may lack clear frameworks for productive introspection when left to their own devices. The moderating effect of the Socratic approach appears to be driven by its ability to

direct participants' attention toward examining their own reasoning processes in a systematic way, rather than leaving them to determine how best to engage with their positions on their own.

Examining sub-components of argument quality reveals more nuanced patterns that illuminate the mechanisms underlying attitude moderation. While the AI-Socratic, AI-Emotional, and Reflective paragraph treatments likely increased argument depth, the experiment is underpowered to detect statistically significant effects. The depth measure captures whether participants engaged thoroughly with core issues rather than making superficial claims, suggesting that structured questioning may prompt more substantive reflection even when overall argument quality remains unchanged.

More importantly, AI-Socratic dialogue increases the diversity of perspectives considered by extreme participants by 0.07 SDs (SE = 0.031; $p < 0.05$), compared to extreme participants in the waiting control. This is equivalent to a 0.03-point increase in the scoring scale. Perspective diversity assesses recognition and integration of alternative viewpoints—specifically measuring whether participants acknowledge competing considerations, recognize trade-offs, or demonstrate awareness of how others might reasonably disagree with their position.

This finding provides crucial insight into why AI-facilitated interventions reduce polarization without improving argument quality. The Socratic questioning process appears to prompt participants to consider multiple dimensions of complex policy issues rather than simply strengthening their existing arguments. When asked to articulate their reasoning step-by-step, participants may spontaneously recognize limitations in their initial positions or acknowledge valid concerns raised by opposing viewpoints. This increased cognitive complexity —captured by higher perspective diversity scores— may create the psychological space necessary for attitude moderation by reducing the certainty and rigidity that typically characterize extreme political positions.

The disconnect between perspective diversity improvements and overall thoughtfulness scores suggests that traditional measures of argument quality may miss important dimensions of political reasoning. Although participants do not construct more logically rigorous arguments, they develop more nuanced understanding of others' perspectives. This pattern aligns with recent empirical research showing that the ability to consider multiple perspectives is associated with more moderate political attitudes (Tetlock, 1984; Adida et al., 2018; Houck et al., 2020; Kalla and Broockman, 2023). The AI intervention appears to foster this cognitive flexibility without requiring participants to master formal argumentation skills or engage directly with opposing arguments that might trigger defensive responses.

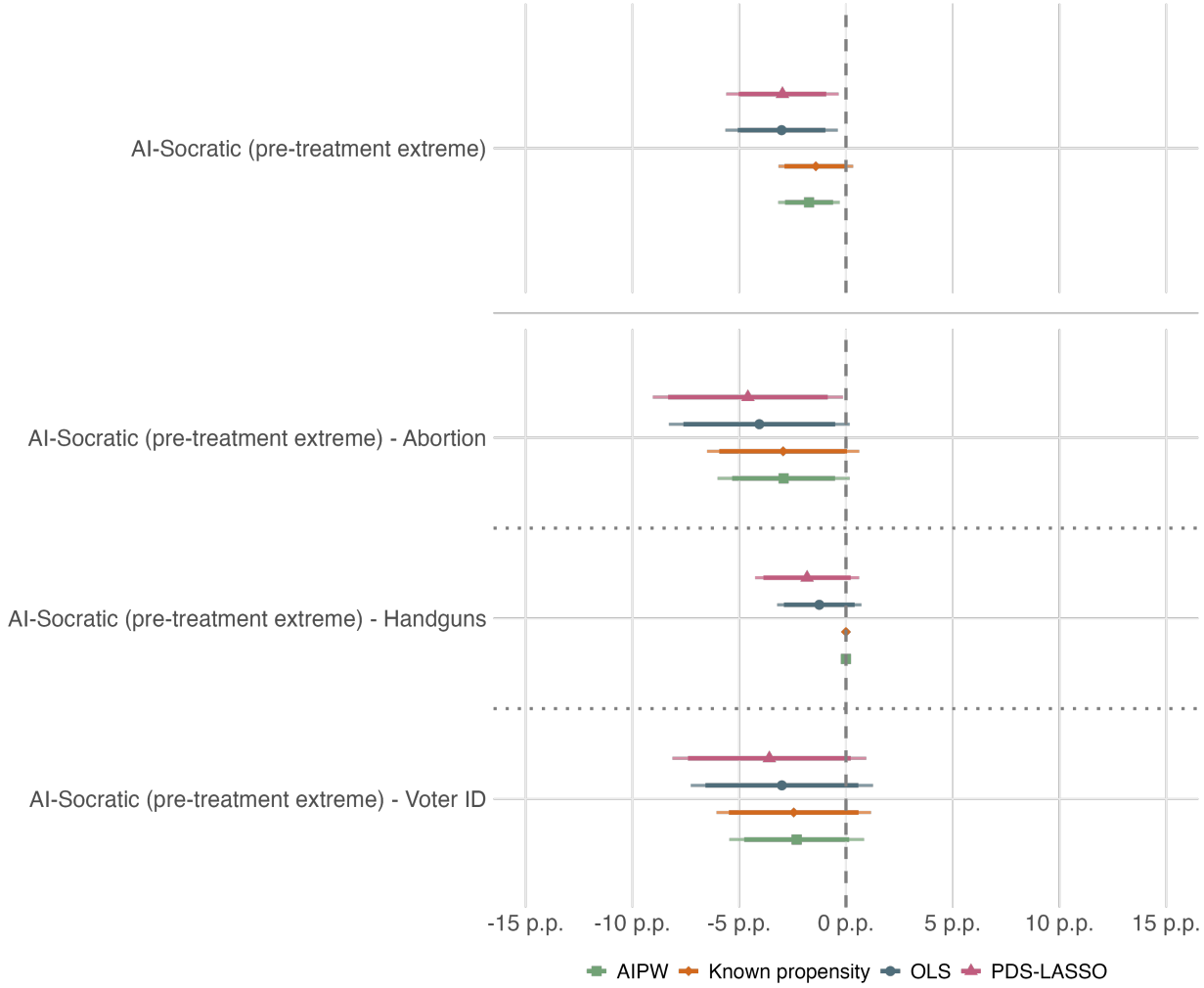
6.2 Experiment 2: Effects of Socratic dialogue across topics

6.2.1 Policy position moderation and confidence among extreme participants

Position moderation effects. Figure 7 presents the central findings from Experiment 2, examining treatment effects among participants with initially extreme positions across three political topics. The top panel shows overall pooled effects, while the bottom panel reveals heterogeneous

effects by topic. Appendix Table D3 provides comprehensive results for movement toward the median, position change, and confidence outcomes. The results corroborate and extend the position moderation findings from Experiment 1.

Figure 7: Experiment 2: Conditional average treatment effects on position extremity among initially extreme participants, overall and by topic



Notes: The figure indicates average treatment effects among extreme participants on a binary indicator of extreme positioning at post-treatment. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Full results including AI-Grammar are reported in Appendix Table D3. Bars indicate 90 and 95% confidence intervals from one-sided tests, with standard errors clustered at the participant level.

The top panel of Figure 7 shows that AI-Socratic dialogue produces significant moderation effects when pooling across the three policy topics. Extreme participants in the AI-Socratic condition are 3 p.p. less likely to maintain extreme positions compared to extreme participants in the AI-Grammar control condition (SE = 1.6; $p < 0.05$ for one-sided test). While smaller in magnitude than Experiment 1, this effect represents approximately 6% of the control standard deviation (0.495), indicating meaningful moderation across diverse political issues.

The bottom panel reveals substantial heterogeneity across topics, with the strongest effects

concentrated on abortion policy. For gestational limits on abortion —the topic with the highest cross-partisan overlap— extreme participants in the AI-Socratic condition are 4.6 p.p. less likely to maintain extreme positions (SE = 2.3; $p < 0.05$ for one-sided test). This represents nearly 10% of the control standard deviation and aligns with theoretical expectations that moderation is most achievable where some common ground exists. Effects for handgun bans and voter ID requirements are smaller and statistically insignificant, consistent with these topics' higher baseline polarization.

Columns (3) and (4) of Appendix Table D3 confirm that the moderation effects operate through systematic movement toward the baseline median rather than random position changes. When pooling across topics, extreme participants in the AI-Socratic condition are 2.3 p.p. more likely to move toward the topic median (SE = 1.7; $p < 0.10$ for one-sided test). Topic-specific analysis reveals meaningful movement toward the median for abortion (4.6 p.p, SE = 2.6; $p < 0.10$), while handgun policies (1.6 p.p., SE = 1.4) and voter ID show minimal movement (1.5 p.p., SE = 2.4).

Given the baseline rate of movement toward the median among extreme participants (control mean = 10.6%), the abortion effect represents a substantial 43% increase in centrist movement and corresponds to approximately 15% of the control standard deviation (0.308). This systematic pattern mirrors Experiment 1 findings and reinforces that AI-guided deliberation promotes genuine moderation rather than arbitrary position shifts.

Heterogeneity across topics provides crucial insights into the boundary conditions for AI-facilitated moderation. Abortion policy shows the strongest and most consistent effects across both extremity reduction and median movement, reflecting this issue's greater potential for nuanced discussion despite partisan divisions. The topic's complex trade-offs around gestational timing, maternal health, and fetal development appear to provide more entry points for reflective reconsideration compared to the more binary framing of handgun bans or voter ID requirements.

Handgun policy shows moderate effects on extremity reduction but smaller movement toward the median, suggesting that while some participants abandon the most extreme positions, they may not necessarily converge on the center. Voter ID requirements show the weakest effects across both measures, consistent with this issue's strong partisan sorting, where Democrats have weak priors but Republicans concentrate on one side of the spectrum, and limited cross-cutting appeal.

Furthermore, the comparison between AI-Socratic dialogue and AI-Grammar correction provides important evidence about mechanisms. The AI-Grammar condition controls for general AI interaction effects while avoiding the structured questioning that characterizes Socratic dialogue. The consistent pattern of larger effects for AI-Socratic across topics suggests that the specific questioning approach, rather than mere AI engagement, drives the observed moderation.

Confidence. Columns (5) and (6) of Appendix Table D3 examine confidence post-treatment, revealing consistent patterns with Experiment 1. The interaction between AI-Socratic treatment and extreme positioning shows a negative but statistically insignificant effect on confidence in the pooled model (−7.2 p.p., SE = 8.7). When compared directly to the AI-Grammar control condition among extreme participants, the AI-Socratic intervention produces significantly lower confidence

(−16.2 p.p., SE = 8.3; p = 0.025 for one-sided test), representing approximately 0.11 SDs of the control group (1.442). Topic-specific analysis reveals the strongest confidence reduction for abortion (−20.9 p.p., SE = 10.0; $p = 0.037$), equivalent to 0.14 standard deviations of the control group, followed by voter ID requirements (−17.0 p.p., SE = 0.104; $p = 0.101$), while handgun show smaller and statistically insignificant effects. These results indicate that the intervention operates through a reduction in participants' confidence particularly on complex issues with multiple competing considerations, as in the case of abortion, which is consistent with the proposed mechanism that Socratic questioning reveals gaps in knowledge and understanding.

The overall pattern of results closely mirrors Experiment 1 findings, with somewhat smaller effect magnitudes. This consistency across different experimental designs (single-topic vs. multi-topic), control conditions (waiting vs. AI-Grammar), and political issues strengthens confidence in the robustness of AI-guided Socratic dialogue's moderation effects. The smaller magnitudes in Experiment 2 may reflect the more constrained comparison group (active AI control vs. passive waiting) or dilution effects from including highly polarized topics alongside more moderate ones.

6.2.2 Behavioral outcomes: Cross-partisan donations and comment endorsement

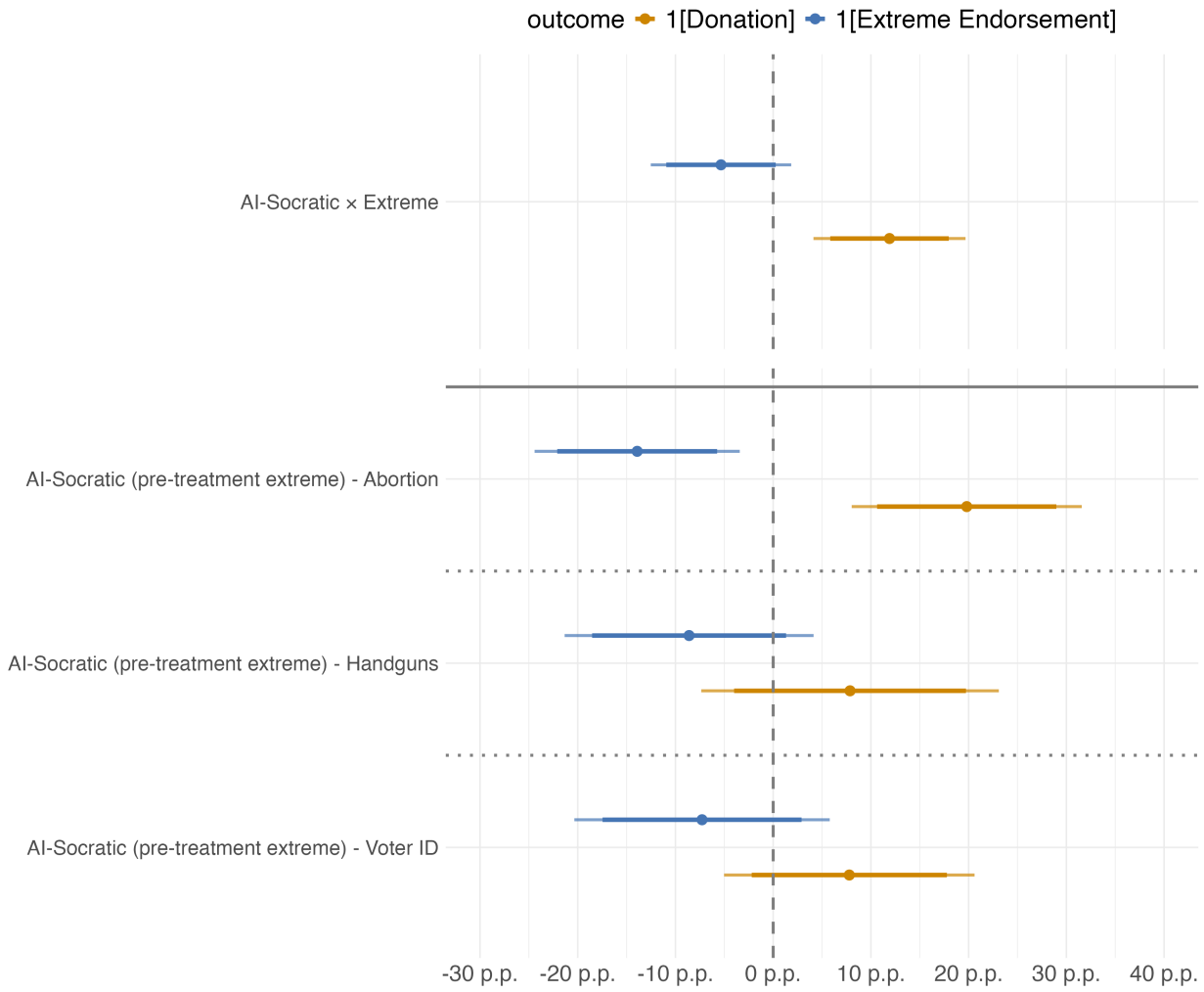
Figure 8 presents evidence that the attitude moderation effects from AI-guided Socratic dialogue translate into meaningful behavioral changes. The figure shows treatment effects on two key behavioral measures: willingness to donate a $1 bonus payment to cross-partisan advocacy groups and patterns of endorsing extreme political comments. These outcomes provide validation that the observed position changes reflect genuine preference shifts rather than superficial survey responses.

Cross-partisan donation behavior. In yellow, the top panel shows that extreme participants in the AI-Socratic condition are significantly more willing to donate to cross-partisan advocacy organizations compared to extreme participants in the AI-Grammar control. Overall, AI-Socratic participants are 12 p.p. more likely to donate their bonus payment (SE = 4.74; $p < 0.01$). This represents a real monetary sacrifice, since participants forfeit actual compensation equivalent to approximately 30% of the total compensation for completing the study, to support organizations that bridge partisan divides, suggesting genuine attitude moderation.

The bottom panel reveals that this effect is primarily driven by participants who discussed abortion policy. For gestational limits on abortion, extreme AI-Socratic participants are 20 p.p. more likely to donate compared to extreme participants in the AI-Grammar condition (SE = 7.12; $p < 0.01$). This represents nearly a 100% increase relative to the control condition baseline.

Comment endorsement patterns. In blue, the figure also shows that AI-guided deliberation affects how participants evaluate political arguments, with extreme AI-Socratic participants showing 6 p.p. lower likelihood of endorsing extreme comments overall (SE = 4.34; $p < 0.10$). This pattern suggests that the intervention not only moderates participants' own positions but also influences their receptivity to polarizing political rhetoric.

Figure 8: Experiment 2: Conditional average treatment effects on Donation to cross-partisan advocacy group and comment endorsing behavior among extreme participants, overall and by topic



Notes: The figure indicates the average effects on cross-partisan donations and extreme comment endorsement among extreme participants. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Estimates from PDS-LASSO. Bars indicate 90 and 95% confidence intervals, calculated using one-sided tests with heteroskedasticity-robust standard errors.

Topic-specific analysis reveals particularly strong effects for abortion policy, where extreme AI-Socratic participants are 14 p.p. less likely to endorse extreme comments compared to extreme participants in the AI-Grammar condition (SE = 6.29; $p < 0.01$). This behavioral change complements the donation findings and provides converging evidence that AI-facilitated deliberation reduces both active and passive forms of political polarization.

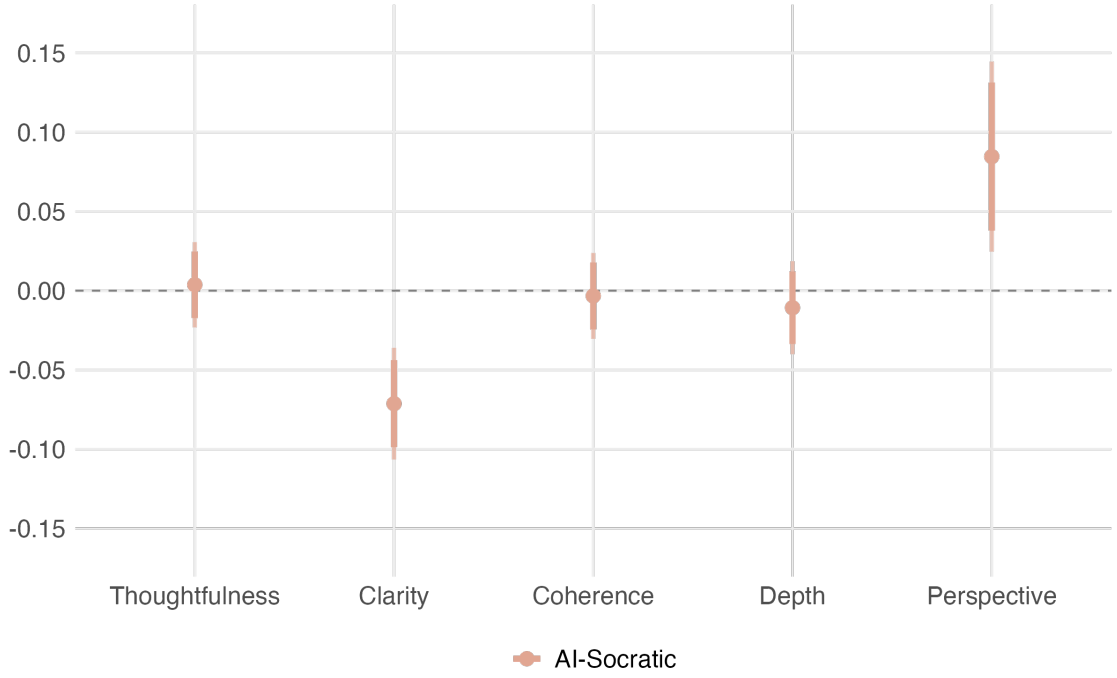
Consistent with the positioning results, the behavioral effects are concentrated on abortion policy—the topic with the greatest cross-partisan overlap and potential for nuanced discussion. While similar patterns emerge for handgun bans and voter ID requirements, with extreme AI-Socratic participants showing greater donation likelihood and lower extreme endorsement, these effects are smaller in magnitude and statistically insignificant. This heterogeneity reinforces that

AI-guided deliberation may be most effective for issues that allow for complex reasoning and consideration of multiple perspectives.

6.2.3 Effects on argument quality

Results in Figure 9 and Appendix Table D8 confirm the findings from Experiment 1 on argument quality. More specifically, I do not find evidence indicating that AI-Socratic dialogue encourages greater argumentative rigor, as measured by the comprehensive thoughtfulness score (0.004 SDs, SE = 0.017). In contrast, I find that it increases perspective diversity by 0.076 SDs (SE = 0.036; $p < 0.05$) among extreme participants, compared to extreme participants in the AI-Grammar condition. This consistent pattern across both experiments strengthens confidence in the robustness of the core finding that attitude moderation occurs independently of improvements in logical reasoning quality.

Figure 9: Experiment 2: Conditional average treatment effects on argumentative quality scores among extreme participants



Notes: The figure indicates average effects among extreme participants on standardized argumentative quality scores. Estimates are from PDS-LASSO. The horizontal dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar condition. Full results are reported in Appendix Table D8. Bars indicate 90 and 95% confidence intervals, calculated using one-sided tests with standard errors clustered at the participant level.

The results also indicate that AI-Grammar and AI-Socratic conditions, on average, do not differ significantly across the other argument quality sub-components —coherence and depth— except for clarity. As expected given its design, AI-Grammar differentially increases clarity by helping participants improve the linguistic precision and readability of their arguments relative to pre-treatment baselines. However, this surface-level improvement does not translate into enhanced substantive

reasoning, consistent with the pattern observed in Experiment 1 where grammar correction actually reduced overall thoughtfulness scores.

The selective effect on perspective diversity, while other argument quality measures remain largely unchanged, provides additional evidence for the proposed mechanism. The AI-Socratic intervention appears to target a specific cognitive process, the consideration of alternative viewpoints, without affecting participants' ability to construct logical arguments, present evidence coherently, or engage deeply with policy details. This precision suggests that Socratic questioning operates through enhanced cognitive flexibility rather than general improvements in reasoning capacity, supporting the epistemic openness hypothesis over traditional models that emphasize argument quality as the pathway to political moderation.

6.2.4 One-month follow-up and spillover effects

I assess the durability and scope of the AI-guided conversation effects. The main results, with detailed regression outputs are in Appendix Figures D1 and D2 and Appendix Tables D5 and D6.

One-month follow-up effects. The results provide evidence for persistence of treatment effects relative to immediate main survey impacts. The top panel of Figure D1 shows that when pooling across all three topics, extreme participants who received AI-Socratic dialogue are 3.7 p.p. less likely to maintain extreme positions one month later compared to extreme participants in the AI-Grammar control (SE = 4.1; p = 0.185 for one-sided test). While not statistically significant due to reduced sample size and statistical power, this effect represents a similar magnitude to that of the main survey effect observed for this same subsample, suggesting that immediate treatment effects do not decay substantially over one month.

The bottom panel reveals meaningful heterogeneity across topics that exceeds main survey benchmarks. For abortion, extreme AI-Socratic participants show a 9.0 p.p. lower likelihood of maintaining extreme positions one month later (SE = 5.6; p = 0.110). For voter ID requirements, the follow-up effect is 7.8 p.p. (SE = 5.4; p = 0.145). For handgun policies, the follow-up effect of 3.1 p.p. (SE = 5.0) represents complete persistence relative to the negligible main survey effect.

Columns (3) and (4) of Appendix Table D5 examine movement toward the topic median. Overall, extreme AI-Socratic participants show a 3.6 p.p. lower likelihood of moving toward the median at follow-up (SE = 2.8; p = 0.100), compared to a 1.9 p.p. higher likelihood in the main survey for this subsample. Topic-specific analysis shows substantial heterogeneity. For abortion, follow-up movement toward the median (5.6 p.p.). For handgun regulation, there is a significant reversal with extreme AI-Socratic participants 9.9 p.p. less likely to move toward the median (SE = 4.4; p = 0.024)—a potential backlash effect relative to the negligible main survey effect. For voter ID, the modest negative follow-up effect (−2.1 p.p.) compares to a positive main survey effect (3.5 p.p.).

Spillover effects. Figure D2 examines whether the AI-guided deliberation experience influences participants' positions on an entirely new topic encountered one month later. These spillover effects

provide evidence that the intervention may teach generalizable skills or attitudes that transfer across political domains.

The results show limited evidence for systematic spillover effects. When pooling across all topics, extreme participants who received AI-Socratic dialogue are 1.6 p.p. less likely to hold extreme positions on new topics compared to extreme participants in the AI-Grammar control, but this difference is small and statistically insignificant (SE = 4.7; p = 0.365 for one-sided test).

As expected, topic-specific analysis reveals considerable heterogeneity in spillover effects, but the pattern is inconsistent and largely insignificant. For abortion as the new topic, extreme AI-Socratic participants show a modest 1.3 p.p. reduction in extreme positioning (SE = 6.4), while for handgun policy, the effect is 1.8 p.p. (SE = 5.9). Only for voter ID requirements is there a larger estimated spillover effect of 6.9 p.p. (SE = 5.6), but even this effect fails to reach statistical significance.

The limited transfer suggests that the moderation effects may be more issue-specific, operating through participants' engagement with the particular content and considerations of specific political topics rather than through the development of generalizable deliberative skills. This pattern aligns with research showing that changes in political attitudes often fails to generalize across domains, particularly when interventions target specific reasoning processes tied to particular issues. The Socratic questioning approach may be most effective when it helps participants work through the specific complexities, trade-offs, and considerations inherent in individual political topics rather than teaching broad meta-cognitive skills that transfer across domains.

6.3 Mechanisms

To examine the theoretical mechanisms, I estimate heterogeneous treatment effects through two complementary approaches. First, I test pre-specified interactions with individual characteristics theoretically linked to the proposed pathways. Second, I employ a data-driven method using causal forests to identify subgroups with differential treatment responsiveness based on predicted conditional average treatment effects (CATEs) (i.e., $\tau(X_i) = E[Y_i(1) - Y_i(0)|X_i = X]$).

6.3.1 Pre-specified heterogeneity analysis

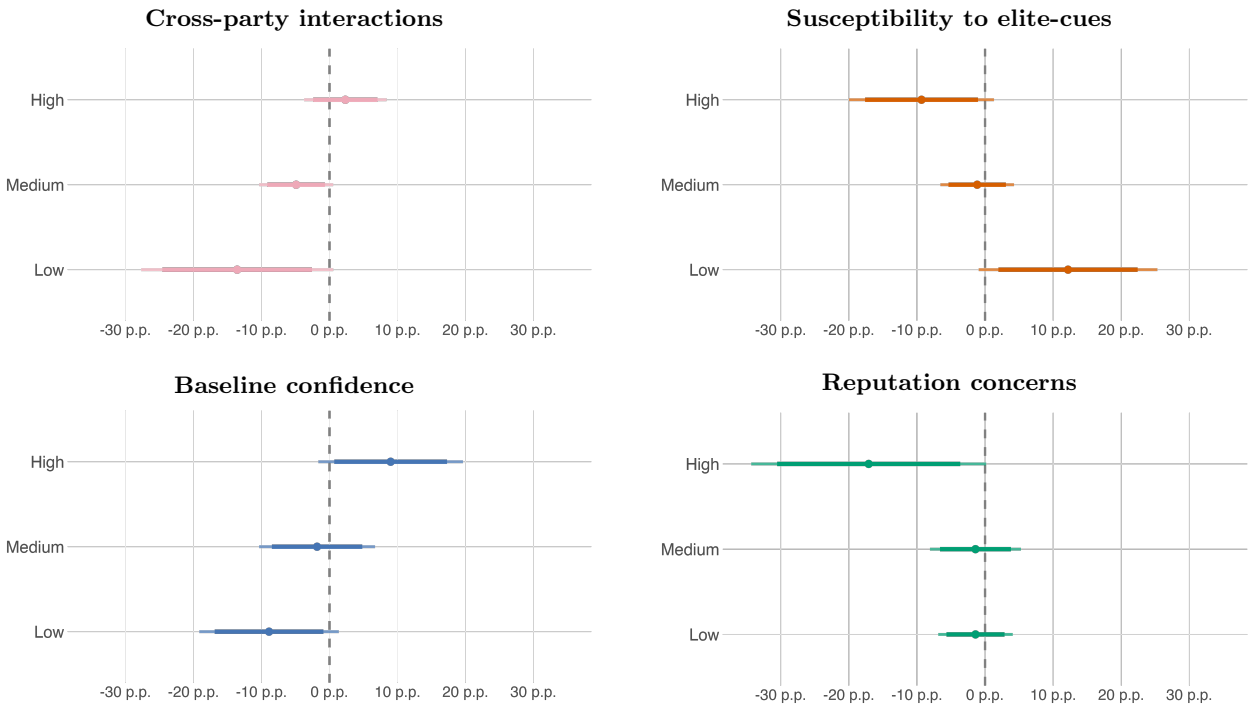
I begin by estimating heterogeneous treatment effects via triple interactions with theoretically motivated individual characteristics:

$$\begin{aligned} Y_i =& \beta_1 \text{Socratic}_i + \beta_2 1[\text{Extreme at pre}_i] + \beta_3 \text{Char}_i \\ &+ \beta_4(\text{Socratic}_i \times 1[\text{Extreme at pre}_i]) + \beta_5(\text{Socratic}_i \times \text{Char}_i) \\ &+ \beta_6(\text{Extreme}_i \times \text{Char}_i) + \beta_7(\text{Socratic}_i \times 1[\text{Extreme at pre}_i] \times \text{Char}_i) \\ &+ \gamma Y_{i,\text{pre}} + \tilde{\mathbb{X}}_i' \boldsymbol{\delta} + \boldsymbol{\delta}_i + \boldsymbol{\phi}_i + \upsilon_i, \end{aligned} \tag{15}$$

where Y_i represents a binary indicator for extreme positioning at post-treatment. Char_i captures individual characteristics including elite cue susceptibility, cross-partisan interaction frequency,

baseline belief confidence, and social conformity concerns.[4] The sum coefficients $\beta_5 + \beta_7$ identifies how individual characteristics moderate the treatment effect among extreme participants. Covariates $\tilde{\mathbb{X}}_i$ are selected via PDS-LASSO and de-meaned, with interactions included for both treatment and characteristic variables. $\boldsymbol{\delta}_i$ and $\boldsymbol{\phi}_i$ are topic and topic-order fixed effects, respectively. Errors υ_i are clustered at the participant level.

Figure 10: Experiment 2: Heterogeneous treatment effects on position extremity



Notes: Each panel presents CATEs by baseline participant characteristics, presented as the difference between extreme AI-Socratic and extreme AI-Grammar on position extremity at post-treatment. Top-left: cross-party interaction frequency (Low: Never; Medium: Rarely/Sometimes; High: Often/Always). Top-right: susceptibility to elite cues measured by agreement with randomly labeled policy proposals (Low: 0/3; Medium: 1/3 and 2/3; High: 3/3). Bottom-left: belief confidence assessed through willingness to wager on factual claims from an initial $100 endowment (Low: $0-$40; Medium: $41-$80; High: $81-$100). Bottom-right: social conformity concerns regarding opinion changes (Low: Not at all; Medium: Slightly/Moderately; High: Very/Extremely). The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Bars indicate 90 and 95% confidence intervals, calculated using one-sided tests with standard errors clustered at the individual participant level.

Belief uncertainty and social influence. Figure 10 presents CATEs on endline position extremity across four participant characteristics that moderate influence responsiveness to AI-guided reflection. Cross-party interaction frequency is measured by participants' self-reported level of interaction with members of other political parties, categorized as Low (Never), Medium (Rarely or Sometimes), and High (Often or Always). Susceptibility to elite cues is operationalized through

[4]Since each moderator represents a theoretically distinct causal pathway rather than multiple operationalizations of a single hypothesis, I do not apply multiple hypothesis testing corrections that would control for family-wise error rates across the four mechanisms examined.

an experimental task where participants evaluated policy proposals with randomized endorsements by prominent national figures and organizations. To ensure strict comparability, participants were exposed to four policies at a time: two endorsed by the same Republican figure and two by the same Democratic figure. Participants are classified based on how many proposals (0, 1-2, or 3 out of 3) they preferred when partisan labels were present versus absent. Belief confidence is measured by participants' willingness to wager money on the accuracy of their factual claims about policy issues, with amounts ranging from 0 to 100 and grouped into Low ($0 to $100) and grouped into Low ($0–$40), Medium ($41–$80), and High ($81–$100) categories. Social conformity concerns capture participants' reported worry about others' reactions to opinion changes, classified as Low (not at all concerned), Medium (slightly or moderately concerned), and High (very or extremely concerned).

Results reveal systematic patterns consistent with the proposed mechanisms, though with caveats about statistical precision. Most of the moderating effects are driven by participants with low cross-party interaction, while medium and high levels report limited effects. Moreover, the effect of AI-Socratic dialogue decreases monotonically with susceptibility to elite cues, with the strongest moderation occurring among participants most responsive to partisan endorsements. Belief confidence shows a positive relationship with treatment effectiveness: participants with low baseline confidence in their factual claims drive the observed moderation effects, while those with high confidence show minimal, or even a negative response to the intervention. Finally, the moderation effect increases with social conformity concerns, with participants most worried about others' reactions to opinion changes showing the strongest response to AI-Socratic dialogue, while the rest show no effects.

Consistent with the previous results, Appendix Figure D8 examines CATEs on position moderation, defined as movement toward the baseline topic-specific median. The outcome is a binary indicator equal to one when a participant's post-treatment position is closer to the baseline median than their pre-treatment position. Consequently, positive coefficients indicate moderation toward the center, representing the mirror image of the extremity reduction results presented above.

These patterns suggest that the intervention is most effective among individuals who are epistemically uncertain and socially constrained. Results are consistent with a mechanism operating through reduced certainty and social pressure rather than improved reasoning quality.

6.3.2 Data-driven heterogeneity analysis

To complement the pre-specified heterogeneity analysis, I employ honest causal forests (Athey et al., 2019; Wager and Athey, 2018) to identify treatment effect heterogeneity in a data-driven manner. This approach estimates CATEs and ranks observations into quartiles based on predicted treatment effects using five-fold honest sample splitting (Athey and Imbens, 2016). Within each fold, I fit a causal forest on the remaining folds and use out-of-bag predictions to rank observations, ensuring that no observation's ranking depends on a model fitted using that observation's data.

Appendix Figure D5 and Table D13 display CATEs by quartile of predicted CATE for both

the overall sample and the subset of initially extreme participants. I estimate these effects using three approaches: OLS, AIPW, and AIPW with known propensity scores. To account for multiple hypothesis testing across quartiles while also preserving power, I report one-sided Romano–Wolf adjusted p-values (Romano and Wolf, 2005)

For the overall sample, quartile 1 (highest predicted negative treatment effect) shows an estimated reduction in extremity of 3.8 percentage points (SE = 2.2 p.p.) using AIPW with known propensity, while quartile 4 shows a near-zero effect of -0.9 p.p. (SE = 2.1 p.p.). Among initially extreme participants, the pattern is more pronounced: quartile 1 shows a 7.0 p.p. reduction (SE = 3.3 p.p.), while quartile 4 shows a positive but insignificant effect of 0.6 p.p. (SE = 2.5 p.p.). The Romano–Wolf adjusted p-value for the quartile 1 effect among extreme participants is 0.021, providing strong evidence against the null of no heterogeneity.

Appendix Figures D6 and D7 examine how baseline covariates vary across predicted CATE quartiles for the overall sample and extreme participants respectively. Each heatmap cell displays the average covariate value within a quartile along with its heteroskedasticity-robust standard error. The color intensity represents the scaled deviation from the sample mean, computed as $\Phi\left(\frac{X_q - X}{SD(Xq)}\right)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function, X_q is the mean of covariate X in quartile q, and $SD(X_q)$ is the standard deviation of the quartile-specific means.

Among all participants, those in quartile 1 (strongest estimated treatment effects) exhibit lower baseline belief confidence (0.64, SE = 0.01) compared to quartile 4 (0.70, SE = 0.01), higher social conformity concerns (0.19, SE = 0.01 versus 0.14, SE = 0.01), and greater susceptibility to elite cues (0.07, SE = 0.04 versus -0.06, SE = 0.04). Cross-partisan interaction shows a similar gradient, with quartile 1 averaging 0.59 (SE = 0.01) compared to 0.64 (SE = 0.01) in quartile 4. Gender composition also varies substantially: females comprise 22.4% (SE = 1.8) of quartile 1 but only 8.1% (SE = 2.0) of quartile 2, declining to -20.6% (SE = 1.9) in quartile 4, indicating males dominate groups with weaker treatment response. Democratic partisanship decreases monotonically from 44.9% (SE = 2.0) in quartile 1 to 25.1% (SE = 1.8) in quartile 4, while Republican identification increases from 23.9% (SE = 1.7) to 33.9% (SE = 2.0). Baseline argument quality score —encompassing overall quality, coherence, and depth— are systematically lower in quartile 1 and increase through quartile 4, suggesting that participants with weaker initial reasoning skills exhibit the largest effects from the intervention.

The pattern intensifies among initially extreme participants. Quartile 1 shows markedly lower belief confidence (0.44, SE = 0.02) relative to quartile 4 (0.56, SE = 0.02), higher reputation concerns (0.27, SE = 0.02 versus 0.17, SE = 0.01), and higher elite cue susceptibility (0.55, SE = 0.01 versus 0.51, SE = 0.02). However, quartile 1 of cross-partisan interaction exhibits similar levels than those of quartile 4 (0.73, SE = 0.02 versus 0.75, SE = 0.02). Gender disparities remain pronounced, with females representing 16.6% (SE = 2.5) of quartile 1 versus -12.7% (SE = 2.7) in quartile 4. Among extreme participants, Democratic identification is more evenly distributed across quartiles (ranging from 31.0% to 37.9%) compared to the full sample, while Republican identification shows a positive gradient from 27.8% (SE = 2.4) to 39.4% (SE = 2.7). Baseline

quality scores again demonstrate that participants with lower initial reasoning quality, particularly on coherence (quartile 1: -0.41, SE = 0.04; quartile 4: 0.17, SE = 0.03) and overall quality (quartile 1: -0.36, SE = 0.06; quartile 4: 0.15, SE = 0.06), exhibit the strongest treatment responsiveness.

6.3.3 Conversational text analysis

To complement the survey-based and annotation-based analyses, I examine the conversational dynamics during the AI-guided dialogue using text analysis. This approach allows me to test whether the theoretical mechanisms operate through observable changes in participants' language use during their interactions with the AI conversational agent. I construct seven theory-grounded linguistic indices that correspond to the proposed mechanisms and estimate treatment effects on these indicators across the participant (user) and the (AI) assistant messages.

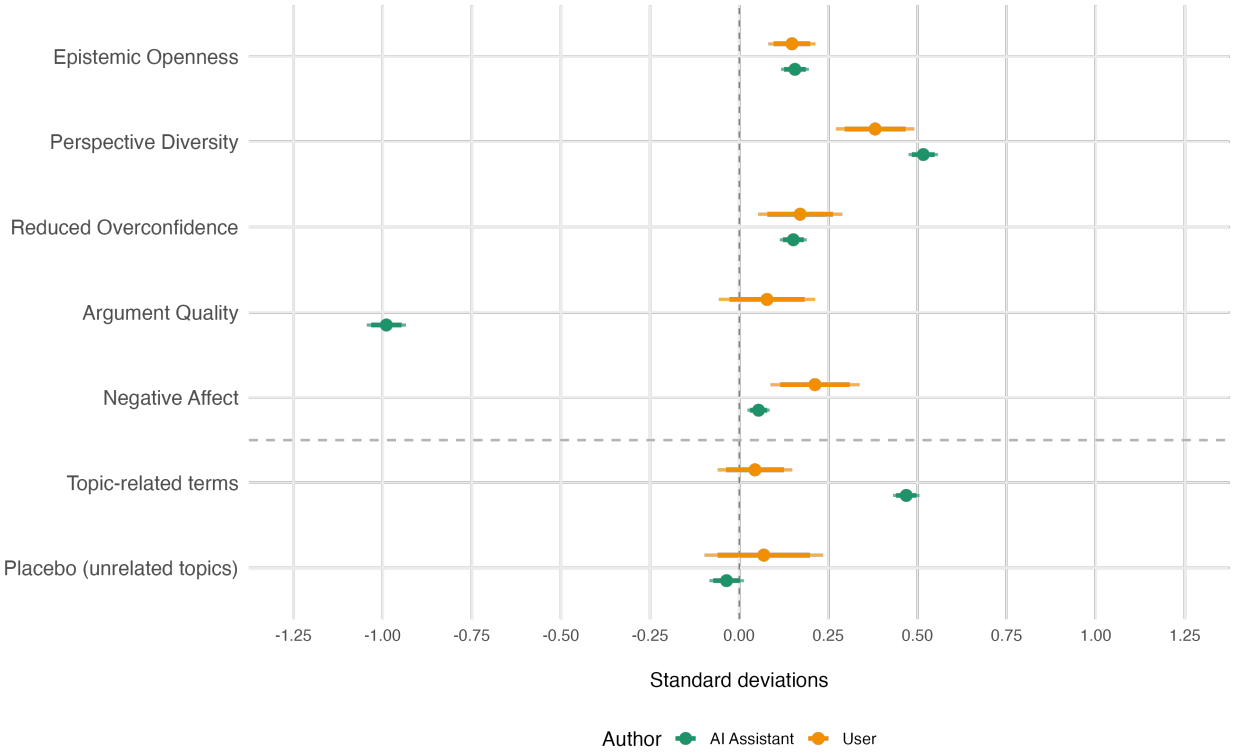
All text is normalized, tokenized, and stemmed using standard NLP pre-processing, with stopwords removed. I construct the linguistic indicators by counting curated stemmed terms and normalizing by the total token count within each document. The indices are: (1) Epistemic openness, capturing acknowledgment of uncertainty (e.g., "unsure," "depends," "might"); (2) Perspective diversity, measuring consideration of multiple viewpoints (e.g., "viewpoint," "tradeoff," "complicated"); (3) Reduced overconfidence, quantifying hedging language and qualifiers (e.g., "somewhat," "in my view," "generally"); (4) Argument quality, tracking evidence-based reasoning and logical connectors (e.g., "evidence," "data," "demonstrates"); (5) Negative affect, using NRC emotion lexicon terms for anger and fear. I also include two falsification checks: (6) topic-related terms (e.g., "abortion," "firearm," "voting"); and (7) Placebo, counting terms from unrelated policy domains (immigration, climate, healthcare) that should not be included in the conversations (e.g., "asylum," "carbon," "medicare"). Appendix Table D9 provides the complete list of terms used to construct each linguistic index.

For each index and message type, I estimate treatment effects using token-weighted least squares regression with topic fixed effects and participant-clustered standard errors, controlling for (demeaned) text length. I standardize all indices to z-scores to facilitate comparability across measures. Estimates are obtained separately for assistant messages and user messages.

Figure 11 presents treatment effects during the AI-guided dialogue for participants with initially extreme positions. The results indicate marked differences across indices supporting the theoretical predictions. For the first three indices, corresponding to the epistemic openness mechanism, the patterns strongly support the theoretical framework. The AI assistant in the Socratic condition uses substantially more epistemic openness language (0.156 SD, SE = 0.024; RW-adjusted $p < 0.01$) compared to the grammar correction condition, and participants exposed to this questioning respond with similarly epistemic openness language (0.148 SD, SE = 0.040; RW-adjusted $p < 0.01$). Perspective diversity shows even larger effects, with the Socratic AI using 0.516 SD more perspective-taking language (SE = 0.025; RW-adjusted $p < 0.01$) and participants responding with 0.381 SD more such language (SE = 0.067; RW-adjusted $p < 0.01$). Reduced overconfidence also increases significantly in both AI messages (0.152 SD, SE = 0.023; RW-adjusted $p < 0.01$) and

participant responses (0.171 SD, SE = 0.072; RW-adjusted p = 0.018). Notably, AI assistant coefficients show standard errors 2-3 times smaller than user coefficients despite similar effect sizes.

Figure 11: Linguistic differences between AI-Socratic and AI-Grammar treatments among extreme participants



Notes: Figure indicates average effects among extreme participants on standardized linguistic indices. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar condition. Bars indicate 90 and 95% confidence intervals for one-sided tests, with standard errors clustered at the participant level.

Argument quality shows no evidence of improvement and actually decreases markedly in AI assistant messages (-0.990 SD, SE = 0.033; $p < 0.01$) while remaining unchanged in participant responses (0.078 SD, SE = 0.082; p = 0.171). This divergence is theoretically and practically important: the Socratic AI deliberately avoids providing evidence-based arguments or logical reasoning, instead focusing exclusively on questioning. Participants do not compensate by increasing their own argumentative rigor, confirming that the intervention operates through epistemic openness rather than enhanced reasoning quality.

The control and falsification indices provide supporting evidence for the validity of the approach. Negative affect shows small positive effects in both AI (0.054 SD, SE = 0.019) and participant messages (0.212 SD, SE = 0.076), suggesting that Socratic questioning may slightly increase emotional engagement without being the primary mechanism. Topic-related terms show a large positive effect for the AI (0.468 SD, SE = 0.022) but no effect for participants (0.044 SD, SE = 0.063), indicating that the Socratic AI uses more topic-specific vocabulary in its questions while participants maintain similar usage regardless of condition. Most importantly, the placebo index shows no significant effects for either AI or participants, confirming that the intervention does not increase discussion

of unrelated policy topics.

Appendix Figure D4 presents the results for non-extreme participants. The results largely mirror those for extreme participants, with 30-40% smaller in magnitude for the epistemic openness and perspective diversity. The reduced overconfidence effect is similar in magnitude but noisier, while argument quality again shows no improvement. This pattern suggests that the intervention works through similar mechanisms for both extreme and non-extreme participants, but with stronger effects among those with more polarized initial positions who have greater room for epistemic opening.

6.3.4 Additional results

Perceived bias and agency. Appendix Tables D10 and D11 examine whether the AI intervention affects participants' perceptions of neutrality and autonomy. The results provide reassuring evidence that AI-Socratic dialogue does not systematically undermine perceptions of neutrality or user agency.

Extreme participants in the AI-Socratic condition show modest decreases in perceiving the AI as neutral, but the estimates are not statistically significant (6.0 p.p., SE = 6.2; $p = 0.335$ for a two-sided test). Similarly, there are no significant differences in perceived research team neutrality or feelings of pressure from either the AI system or research team across treatment conditions. These patterns suggest that while participants may detect the AI's more active questioning approach compared to grammar correction, they do not interpret this enhanced engagement as inappropriate bias or coercion. The preservation of perceived agency is particularly important given concerns about AI manipulation in political contexts (e.g., Hackenburg et al., 2025).

Openness and information seeking behavior. Appendix Table D12 examines whether the intervention produces broader changes in political openness that align with the proposed epistemic openness mechanism, and explores potential downstream effects on information seeking behavior across other contentious political topics.

The results indicate that extreme participants in the AI-Socratic condition show increased openness to others' perspectives in the main survey (8.6 p.p., SE = 6.6; $p < 0.10$ for one-sided test), with a similar though underpowered pattern in the follow-up survey (10.0 p.p., SE = 11.2 p.p; $p = 0.186$). This result suggests that AI-guided Socratic dialogue increases participants' willingness to consider alternative viewpoints and acknowledge the legitimacy of opposing perspectives.

Regarding downstream effects on information seeking behavior, extreme participants show no increased interest in the specific deliberation topics from the main survey but demonstrate suggestive increases in seeking information about other polarized political topics such as police funding, climate policy, and immigration (8.4 p.p., SE = 8.1; $p = 0.147$). This selective increase suggests that the intervention may foster general curiosity about political complexity as a downstream consequence of enhanced epistemic openness that is not tied to the original conversational experience.

7 Discussion

The study results demonstrate that AI-facilitated Socratic dialogue can significantly reduce political polarization across multiple contentious issues, offering a scalable technological solution to one of democracy's most pressing challenges. Rather than working through improved argumentation as traditionally assumed, the intervention operates by fostering epistemic openness among politically disengaged citizens with low belief confidence, high social conformity concerns, and limited cross-partisan exposure.

Implications for understanding polarization. These findings advance our understanding of the psychological mechanisms underlying political polarization. The concentration of effects among participants with low belief confidence, high social conformity concerns, and greater susceptibility to elite cues reveals that a substantial portion of the electorate may hold strong political positions publicly while remaining privately uncertain about complex policy issues. This challenges assumptions that polarization is primarily driven by deeply committed partisans and suggests that much political extremism may reflect social pressure and cognitive shortcuts rather than genuine ideological commitment.

The independence of moderation effects from argument quality improvements represents a fundamental challenge to existing models of political attitude change. Standard approaches often focus on providing better information or enhancing reasoning capabilities (Nyhan and Reifler, 2010; Dunning et al., 2019; Roozenbeek and van der Linden, 2020), but my results suggest these may miss the primary barriers to compromise: psychological rigidity and social pressure rather than cognitive limitations.

Technological solutions to democratic dysfunction. For policymakers and platform designers, my findings suggest several promising directions. AI facilitation addresses how to provide high-quality deliberative experiences without prohibitive resource costs that limit traditional forums (Fishkin, 2018). Platforms could identify users exhibiting characteristics associated with responsiveness—low cross-partisan exposure, high elite cue susceptibility, or expressions of uncertainty—and offer them personalized reflection opportunities.

The behavioral effects extend beyond stated preferences to actual resource allocation decisions, with increased willingness to donate to cross-partisan organizations and endorse moderate content. This demonstrates that attitude moderation translates into meaningful changes in political behavior, addressing concerns about whether laboratory-measured changes reflect genuine belief revision.

Applications and limitations. The findings have direct implications for digital civic participation where polarization poses challenges. Participatory budgeting processes, online citizen assemblies, and policy consultations could benefit from structured questioning that helps participants examine their reasoning. Corporate contexts dealing with sensitive topics like diversity initiatives,

environmental policies, or workplace changes could also utilize AI-facilitated dialogue as a neutral mechanism for exploring perspectives.

However, important limitations warrant consideration. The study's controlled environment differs from natural political conversations with real consequences, requiring field experiments for external validity tests. More fundamentally, the concentration of effects among politically disengaged participants raises normative concerns about systematically influencing less engaged citizens' political views. The techniques could potentially be misused to shift opinions problematically, highlighting the urgent need for governance frameworks and ethical guidelines for AI in democratic contexts as these technologies mature.

8 Conclusion

This study provides novel experimental evidence that AI-facilitated conversations can reduce political polarization across multiple contentious issues through scalable technological interventions. By prompting participants to articulate their reasoning through structured questioning, AI systems achieve genuine movement away from extreme positions toward greater acceptance of complexity and compromise—outcomes with significant implications for democratic governance and economic policy-making.

The findings offer important insights into the behavioral foundations of political polarization. The concentration of effects among epistemically uncertain and socially constrained participants suggests that much political extremism reflects information frictions and social constraints rather than deep preference heterogeneity. This challenges standard political economy models that treat polarization as reflecting fundamental ideological differences and points toward interventions that address cognitive and social barriers to preference updating.

From a policy design perspective, the results demonstrate that technological solutions to democratic dysfunction can generate measurable welfare improvements at low marginal cost. Traditional deliberative interventions face classic public goods problems—high fixed costs, free-riding, and limited scalability that prevent efficient provision of democratic discourse quality. AI-facilitated dialogue offers a technological solution that reduces these transaction costs while preserving the benefits of structured political engagement. As political divisions impose growing costs on economic policy-making and institutional effectiveness, such tools represent potentially high-return investments in democratic infrastructure.

The welfare implications extend beyond immediate attitude changes to downstream economic effects. Political polarization reduces the efficiency of policy-making by preventing compromise on potential welfare-enhancing reforms. By demonstrating that targeted interventions can moderate extreme positions and increase willingness to support cross-partisan initiatives, my findings suggest pathways for improving the quality of democratic decision-making with direct implications for economic outcomes.

However, the promising results also highlight important questions about optimal intervention

design and targeting. The concentration of effects among politically disengaged participants raises concerns about distributional impacts: systematic attitude changes among less informed citizens could alter electoral outcomes and policy adoption in ways that may not enhance aggregate welfare. Future research should examine how AI interventions affect not only individual attitude formation but also the aggregation of preferences through democratic institutions.

As democratic societies face growing political divisions that impede efficient policy-making, understanding how to design technology that improves rather than degrades democratic discourse represents a critical research frontier. My findings suggest that AI, appropriately designed with attention to the underlying behavioral mechanisms, could serve as a valuable tool to address the lack of compromise and cross-partisan cooperation. In an era where political polarization threatens both governance and economic efficiency, such technological solutions merit serious consideration as complements to traditional institutional reforms.

References

Adida, Claire L, Adeline Lo, and Melina R Platas, "Perspective taking can promote short-term inclusionary behavior toward Syrian refugees," *Proceedings of the National Academy of Sciences*, 2018, *115* (38), 9521–9526.

Akbiyik, Ahmet, Jeremy Bowles, Horacio Larreguy, and Shelley Lui, "How Does Exposure to Discordant Media Sources Affect Political Attitudes and Behavior? Experimental Evidence from Turkey," *Working paper*, 2025.

Alesina, Alberto, Marlon Seror, David Y Yang, Yang You, Weihong Zeng et al., *Persistence through revolutions*, National Bureau of Economic Research, 2020.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The welfare effects of social media," *American Economic Review*, 2020, *110* (3), 629–676.

_ , Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon et al., "The effects of Facebook and Instagram on the 2020 election: A deactivation experiment," *Proceedings of the National Academy of Sciences*, 2024, *121* (21), e2321584121.

Argyle, Lisa P, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate, "Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale," *Proceedings of the National Academy of Sciences*, 2023, *120* (41), e2311627120.

Athey, Susan and Guido Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

_ , Julie Tibshirani, and Stefan Wager, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

_ , Matias Cersosimo, Kristine Koutout, and Zelin Li, "Emotion-versus Reasoning-based Drivers of Misinformation Sharing: A field experiment using text message courses in Kenya," 2025.

_ , Niall Keleher, and Jann Spiess, "Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal," *Journal of Econometrics*, 2025, *249*, 105945.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky, "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, 2018, *115* (37), 9216–9221.

Bang, Heejung and James M Robins, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 2005, *61* (4), 962–973.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski, "Toxic content and user engagement on social media: Evidence from a field experiment," *CESifo Working Paper*, 2025.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, 2014, *81* (2), 608–650.

Bénabou, Roland and Jean Tirole, "Identity, morals, and taboos: Beliefs as assets," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Benson, Jonathan, "The epistemic value of deliberative democracy: how far can diversity take us?," *Synthese*, 2021, *199* (3), 8257–8279.

_ , "Is fake news a threat to deliberative democracy?: Partisanship, inattentiveness, and deliberative capacities," *Social Theory and Practice*, 2023.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro, "Greater Internet use is not associated with faster growth in political polarization among US demographic groups," *Proceedings of the National Academy of Sciences*, 2017, *114* (40), 10612–10617.

_ , _ , and _ , "Cross-country trends in affective polarization," *Review of Economics and Statistics*, 2024, *106* (2), 557–565.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, 2017, *114* (28), 7313–7318.

Braghieri, Luca, Ro'ee Levy, and Hannah Trachtman, "Frictions in News Consumption: Evidence from Social Media," *Available at SSRN 5494670*, 2025.

_ , Ro'ee Levy, and Alexey Makarin, "Social media and mental health," *American Economic Review*, 2022, *112* (11), 3660–3693.

_ , Sarah Eichmeyer, R Levy, MM Mobius, J Steinhardt, and R Zhong, "Article level slant and polarization of news consumption on social media," *Available at SSRN*, 2024, *4932600*.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin, "From extreme to mainstream: The erosion of social norms," *American Economic Review*, 2020, *110* (11), 3522–3548.

Carney, Kevin, "The effect of social media on voters: experimental evidence from an Indian election," *Working paper*, 2025.

Chambers, Simone, "Making Referendums Safe for Democracy: A Call for More and Better Deliberation," *Swiss Political Science Review*, 2018, *24* (3), 305–311.

Chinoy, Sahil and Martin Koenen, "Political Sorting in the US Labor Market: Evidence and Explanations," *Working paper*, 2024.

Cohen, Joshua, *Deliberative Democracy*, London: Palgrave Macmillan UK,

_ , Archon Fung, Lucy Bernholz, Hélène Landemore, and Rob Reich, "Democracy and the digital public sphere," *Digital Technology and Democratic Theory*, 2021, pp. 23–61.

Cruz, Ceci and Horacio Larreguy, "Political Polarization," *Working paper*, 2025.

DellaVigna, Stefano and Ethan Kaplan, "The Fox News effect: Media bias and voting," *The Quarterly Journal of Economics*, 2007, *122* (3), 1187–1234.

Ding, Yuyang, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He, "Boosting large language models with socratic method for conversational mathematics teaching," in "Proceedings of the 33rd ACM International Conference on Information and Knowledge Management" 2024, pp. 3730–3735.

Djourelova, Milena, Ruben Durante, and Gregory J Martin, "The impact of online competition on local newspapers: Evidence from the introduction of Craigslist," *Review of Economic Studies*, 2025, *92* (3), 1738–1772.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, and Gareth Nellis, *Information, accountability, and cumulative learning: Lessons from Metaketa I*, Cambridge University Press, 2019.

Durán, Rafael Jiménez, Karsten Müller, and Carlo Schwarz, "The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG," *Available at SSRN 4230296*, 2024.

Eady, Gregory, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler, "News sharing on social media: mapping the ideology of news media, politicians, and the mass public," *Political Analysis*, 2025, *33* (2), 73–90.

Enke, Benjamin and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

_ and Thomas Graeber, "Cognitive uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser, "Accountability under polarization," *Working paper*, 2024.

Erceg, Nikola, Zvonimir Galić, and Mitja Ružojčić, "A reflection on cognitive reflection–testing convergent/divergent validity of two measures of cognitive reflection," *Judgment and Decision making*, 2020, *15* (5), 741–755.

Eyster, Erik, Matthew Rabin, and Dimitri Vayanos, "Financial markets where traders neglect the informational content of prices," *The Journal of Finance*, 2019, *74* (1), 371–399.

Federal Bureau of Investigation, "Federal Denials," https://www.fbi.gov/file-repository/federal_denials.pdf/view 2025.

Fishkin, James S, *Democracy when the people are thinking: Revitalizing our politics through public deliberation*, Oxford University Press, 2018.

Frederick, Shane, "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 2005, *19* (4), 25–42.

Gallup, "Abortion," https://news.gallup.com/poll/1576/abortion.aspx 2025. Historical trends and polling data on abortion.

Gastil, John, *By popular demand: Revitalizing representative democracy through deliberative elections*, Univ of California Press, 2000.

Gentzkow, Matthew and Jesse M Shapiro, "Ideological segregation online and offline," *The Quarterly Journal of Economics*, 2011, *126* (4), 1799–1839.

Goldberg, Beth, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer et al., "AI and the Future of Digital Public Squares," *arXiv preprint arXiv:2412.09988*, 2024.

Greene, Kenneth F, Erin L Rossiter, Enrique Seira, and Alberto Simpser, "Interacting as equals reduces partisan polarization in Mexico," *Nature Human Behaviour*, 2025, *9* (1), 147–155.

Gutmann, Amy and Dennis F. Thompson, *Why Deliberative Democracy?*, Princeton, NJ: Princeton University Press, 2004.

Habermas, Jürgen, *The theory of communicative action: Volume 1: Reason and the rationalization of society*, Vol. 1, Beacon press, 1985.

_ , *Between facts and norms: Contributions to a discourse theory of law and democracy*, John Wiley & Sons, 2015.

Hackenburg, Kobi, Ben M Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G Rand, and Christopher Summerfield, "The Levers of Political Persuasion with Conversational AI," *arXiv preprint arXiv:2507.13919*, 2025.

Hobolt, Sara B, Katharina Lawall, and James Tilley, "The polarizing effect of partisan echo chambers," *American Political Science Review*, 2024, *118* (3), 1464–1479.

Houck, Shannon C, Kathleen J Huber, Mackenzie Ess, and Morgan L Proulx, "Cognitive complexity in political contexts," in "Oxford Research Encyclopedia of Politics" 2020.

Jungherr, Andreas and Adrian Rauchfleisch, "Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide," *arXiv preprint arXiv:2503.07690*, 2025.

Kalla, Joshua L and David E Broockman, "Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting," *American Journal of Political Science*, 2023, *67* (1), 185–204.

Kim, Woojin, "Political polarization in medicine," *Working paper*, 2024.

Kirk, Hannah Rose, Bertie Vidgen, Paul Röttger, and Scott A Hale, "The benefits, risks and bounds of personalizing the alignment of large language models to individuals," *Nature Machine Intelligence*, 2024, *6* (4), 383–392.

Klayman, Joshua and Young-Won Ha, "Confirmation, disconfirmation, and information in hypothesis testing.," *Psychological Review*, 1987, *94* (2), 211.

Konya, Andrew, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel Procaccia, Lisa Schirch, and Michiel Bakker, "Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders," in "Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency" 2025, pp. 312–333.

Kriplean, Travis, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett, "Supporting reflective public thought with ConsiderIt," in "Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work" 2012, pp. 265–274.

Landemore, Hélène, "Open Democracy and Digital Technologies," *Digital Technology and Democratic Theory*, 2021, pp. 62–89.

Larreguy, Horacio and Pia J Raffler, "Accountability in Developing Democracies: The Impact of the Internet, Social Media, and Polarization," *Annual Review of Political Science*, 2025, *28.*

Larsen, Randy J and Edward Diener, "Promises and problems with the circumplex model of emotion," 1992.

Lelkes, Yphtach, Gaurav Sood, and Shanto Iyengar, "The hostile audience: The effect of access to broadband internet on partisan affect," *American Journal of Political Science*, 2017, *61* (1), 5–20.

Levy, Ro'ee, "Social media, news consumption, and polarization: Evidence from a field experiment," *American Economic Review*, 2021, *111* (3), 831–870.

Lodge, Milton and Charles S Taber, *The rationalizing voter*, Cambridge University Press, 2013.

Mansbridge, Jane, "Reconstructing democracy," in "Revisioning the political," Routledge, 1996, pp. 117–138.

_ , Janette Hartz-Karp, Matthew Amengual, and John Gastil, "Norms of deliberation: An inductive study," in "Multi-Party Dispute Resolution, Democracy and Decision-Making," Routledge, 2017, pp. 139–185.

Martin, Gregory J and Ali Yurukoglu, "Bias in cable news: Persuasion and polarization," *American Economic Review*, 2017, *107* (9), 2565–2599.

McCarty, Nolan, Keith T Poole, and Howard Rosenthal, *Polarized America: The dance of ideology and unequal riches*, MIT Press, 2016.

McConnell, Christopher, Yotam Margalit, Neil Malhotra, and Matthew Levendusky, "The economic consequences of partisanship in a polarized era," *American Journal of Political Science*, 2018, *62* (1), 5–18.

Moehring, Alex, Avinash Collis, Kiran Garimella, M Amin Rahimian, Sinan Aral, and Dean Eckles, "Providing normative information increases intentions to accept a COVID-19 vaccine," *Nature Communications*, 2023, *14* (1), 126.

Mohammad, Saif M and Peter D Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, 2013, *29* (3), 436–465.

Mousa, Salma, Lennard Naumann, and Alexandra Scacco, "Intergroup Contact, Empathy Education, and Refugee-Native Integration: Evidence from a Field Experiment in Lebanon," 2025.

Müller, Karsten and Carlo Schwarz, "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, 2021, *19* (4), 2131–2167.

Munger, Kevin, "Tweetment effects on the tweeted: Experimentally reducing racist harassment," *Political Behavior*, 2017, *39*, 629–649.

Nyhan, Brendan and Jason Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 2010, *32* (2), 303–330.

Ortoleva, Pietro and Erik Snowberg, "Overconfidence in political behavior," *American Economic Review*, 2015, *105* (2), 504–535.

Paul, Richard, "Critical thinking: What every student needs to survive in a rapidly changing world," *Critical Thinking, Moral Integrity and Citizenship*, 1993.

Pei, Jiaxin, José Ramón Enríquez, Umar Patel, Alia Braley, Nuole Chen, Lily Tsai, and Alex Pentland, "DELIBERATION.IO: Facilitating Democratic and Civil Engagement at Scale with Open-Source and Open-Science," *Working paper*, 2025.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, 2021, *592* (7855), 590–595.

Pentland, Alex and Lily Tsai, "Toward building deliberative digital media: from subversion to consensus," *PNAS Nexus*, 2024, *3* (10), pgae407.

Peterson, Erik and Shanto Iyengar, "Partisan gaps in political information and information-seeking behavior: Motivated reasoning or cheerleading?," *American Journal of Political Science*, 2021, *65* (1), 133–147.

Pew Research Center, "Amid a Series of Mass Shootings in the U.S., Gun Policy Remains Deeply Divisive," https://www.pewresearch.org/politics/2021/04/20/amid-a-series-of-mass-shootings-in-the-u-s-gun-policy-remains-deeply-divisive/ April 2021.

Pinzón Hernández, Daniel Alejandro, Jungkyu Rhys Lim, Michelle Dugas, Ellen Elizabeth Moscoe, Mohamad Chatila, Corey Cameron, Renos Vakis, Zeina Afif, and Victor Hugo Orozco Olvera, "Behaviorally Informed Messages Increase COVID-19 Vaccination Intentions : Insights from a Global Meta-Analysis," *World Bank Policy Research Working Papers*, 2024.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 1994, *89* (427), 846–866.

Romano, Joseph P and Michael Wolf, "Stepwise multiple testing as formalized data snooping," *Econometrica*, 2005, *73* (4), 1237–1282.

Roozenbeek, Jon and Sander van der Linden, "Breaking Harmony Square: A game that "inoculates" against political misinformation," 2020.

Rossiter, Erin L and Taylor N Carlson, "Cross-partisan conversation reduced affective polarization for republicans and democrats even after the contentious 2020 election," *The Journal of Politics*, 2024, *86* (4), 1608–1612.

Sanders, Lynn M, "Against deliberation," *Political Theory*, 1997, *25* (3), 347–376.

Schiff, Kaylyn Jackson, Daniel S Schiff, and Bryce Dietrich, "The Second Order Effects of LLMs and Generative AI on Political Communication," *Working paper*, 2024.

Schwardmann, Peter, Egon Tripodi, and Joël J Van der Weele, "Self-persuasion: Evidence from field experiments at international debating competitions," *American Economic Review*, 2022, *112* (4), 1118–1146.

Sorensen, Taylor, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri et al., "A roadmap to pluralistic alignment," *arXiv preprint arXiv:2402.05070*, 2024.

Substance Abuse and Mental Health Services Administration, "Results from the 2022 National Survey on Drug Use and Health: Detailed Tables," https://www.samhsa.gov/data/sites/default/files/reports/rpt42731/2022-nsduh-annual-national-web-110923/2022-nsduh-nnr.htm#:~:text=In%202022%2C%2021.8%20percent%20of,treatment%20in%20the%20past%20year 2023.

Taber, Charles S and Milton Lodge, "Motivated skepticism in the evaluation of political beliefs," *American Journal of Political Science*, 2006, *50* (3), 755–769.

Tessler, Michael Henry, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes et al., "AI can help humans find common ground in democratic deliberation," *Science*, 2024, *386* (6719), eadq2852.

Tetlock, Philip E, "Cognitive style and political belief systems in the British House of Commons.," *Journal of Personality and Social Psychology*, 1984, *46* (2), 365.

Thaler, Michael, "The fake news effect: Experimentally identifying motivated reasoning using trust in news," *American Economic Journal: Microeconomics*, 2024, *16* (2), 1–38.

Tsai, Lily L, Alex Pentland, Alia Braley, Nuole Chen, José Ramón Enríquez, and Anka Reuel, "Generative AI for pro-democracy platforms," *An MIT Exploration of Generative AI*, 2024.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan, "Social media, political polarization, and political disinformation: A review of the scientific literature," *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

Velez, Yamil Ricardo, "Crowdsourced adaptive surveys," *Political Analysis*, 2024, pp. 1–14.

Voelkel, Jan G, James Chu, Michael N Stagnaro, Joseph S Mernyk, Chrystal Redekopp, Sophia L Pink, James N Druckman, David G Rand, and Robb Willer, "Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes," *Nature Human Behaviour*, 2023, *7* (1), 55–64.

_ , **Michael N Stagnaro, James Y Chu, Sophia L Pink, Joseph S Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G Allen et al.**, "Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity," *Science*, 2024, *386* (6719), eadh4764.

Wager, Stefan and Susan Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Wright, Scott and John Street, "Democracy, deliberation and design: the case of online discussion forums," *New Media & Society*, 2007, *9* (5), 849–869.

Young, Iris Marion, *Inclusion and democracy*, OUP Oxford, 2002.

Appendix contents

A Treatments **A-3**

A.1 Experiment 1 treatments . . . A-3

A.1.1 Reasoning/active treatments . . . A-3

A.1.2 Passive treatment . . . A-4

A.2 Experiment 2 treatments . . . A-4

B Voting endorsement **A-7**

B.0.1 Seeded comments . . . A-7

B.0.2 Extreme voting . . . A-10

C Randomization balance tests **A-12**

C.1 Experiment 1 Balance Analysis . . . A-12

C.2 Experiment 2 Balance Analysis . . . A-16

C.3 Experiment 2 Follow-up Attrition Analysis . . . A-19

D Results **A-21**

D.1 First stage . . . A-21

D.2 Policy issue position . . . A-22

D.2.1 Experiment 1 . . . A-22

D.2.2 Experiment 2 . . . A-23

D.3 Argumentative quality . . . A-29

D.3.1 Experiment 1 . . . A-29

D.3.2 Experiment 2 . . . A-31

D.4 Additional results . . . A-34

D.4.1 Experiment 2 . . . A-34

D.5 Heterogeneous effects . . . A-36

D.5.1 Experiment 2 . . . A-36

E Experiment 1 survey instruments **A-41**

E.1 Baseline survey . . . A-41

E.2 Endline survey, Part 1 . . . A-42

E.3 Endline survey, Part 2 . . . A-44

F Experiment 2 survey instrument **A-48**

F.1 Main survey . . . A-48

F.2 Follow-up survey . . . A-52

G Rubrics for argument scoring **A-61**

H Ethical considerations **A-62**

H.1 Informed consent A-62

H.2 Confidentiality A-62

H.3 Risk sssessment A-62

H.4 Institutional Review Board (IRB) approval A-62

I Additional notes **A-62**

A Treatments

A.1 Experiment 1 treatments

A.1.1 Reasoning/active treatments

Between baseline and endline surveys, study participants are randomized to one treatment condition. Each of these groups is explained in detail in Section 5.1. Below I include the exact prompt or instructions given to the LLM (OpenAI's `gpt-4o-mini-2024-07-18`) or the participant.

AI-Socratic dialogue prompt

> You are an expert engaging in a Socratic dialogue with a user. The topic is whether people with mental illnesses should be allowed or banned from purchasing guns in the US. Your role is to raise concerns in the form of clear, polished questions in response to the user's comments or answers. Be succinct and respond with one question at a time, encouraging the user to reason through their position on the issue. Avoid being overly pushy. Rather than acting as a teacher asking for an explanation, raise constructive questions.

Reflective paragraph instructions

> Please share your thoughts and perspectives in one paragraph. Aim for 4-5 sentences.

AI-Emotions regulation LLM prompt

> You are an expert engaging in a Socratic dialogue with a user. The topic is whether people with mental illnesses should be allowed or banned from purchasing guns in the US. Your role is to raise concerns in the form of clear, polished questions in response to the user's comments or answers. Be succinct and respond with one question at a time, encouraging the user to reason through their position on the issue. Avoid being overly pushy. Rather than acting as a teacher asking for an explanation, raise constructive questions.

AI-Grammar correction LLM prompt

> You are a meticulous language expert specializing in correcting typos, grammar, and syntax. Your role is to refine the user's text while preserving their original tone and intent. In each round, make precise edits and provide a brief explanation of your changes. If the text is already clear, confirm its correctness and suggest minor improvements if applicable. Be respectful and concise in your responses.
>
> Start with "I will propose an edited version of your comment with respect to typos, grammar, and syntax, while ensuring your tone and intent are preserved. Here's my refined version:"
>
> Finish every iteration, except for the last one, asking the user to respond with the edited version of the comment.

A.1.2 Passive treatment

Waiting control instructions

> Loading page. This may take up to 2 minutes....

A.2 Experiment 2 treatments

In the main survey, participants are randomized to one of two treatments: AI-Socratic dialogue or AI-Grammar correction. Each participant interacts with the AI conversational agent on three occasions: twice during the main survey (once per topic) and once more during the follow-up survey refresher. Below I include the exact prompts and instructions provided to the LLM (OpenAI's `gpt-4-turbo`).

AI-Socratic dialogue prompt

You are an expert engaging in a Socratic dialogue with a user. The topic is [`topic`]. Your role is to raise concerns in the form of clear, polished questions in response to the user's comments or answers. Be succinct and gentle, and respond with one question at a time, encouraging the user to reason through their position on the issue. Avoid being overly pushy. Rather than acting as a teacher asking for an explanation, raise constructive questions.

You may point out logical flaws when appropriate. Do not accept any responses that stray from comments or thoughts about the issue. If the user shares unrelated comments, gently redirect the conversation by asking a relevant question.

To start the conversation, please first say then following: "Hello, I am here to help you reflect by asking you questions. You may talk to me for up to five rounds. You may also move to the next page at any time. Click "Next" when you are ready to move on. Nothing you say to me will be seen publicly."

Then, please repeat the user's initial comment, which is [`participant comment`], affirm that the user is making a good point with their initial comment, and begin asking the user questions.

Limit all your responses to 25 words, regardless of what the user says. If the comment is empty, let the user know.

AI-Grammar correction LLM prompt

You are a meticulous language expert specializing in correcting typos, spellcheck, grammar, and syntax. Your role is to refine the user's text while preserving their original tone and intent. In each round, make precise edits and provide a brief explanation of your changes. If the text is already clear, confirm its correctness and suggest minor improvements if applicable. Be respectful and concise in your responses.

The initial comment you will edit is [`participant comment`]. Start with "I will propose an edited version of your comment to help with typos, spellcheck, grammar, and syntax, while ensuring your tone and intent are preserved. Here's my version:"

Finish every iteration, except for the last one, asking the user if they like this new version or if the user has any suggestions.

B Voting endorsement

B.0.1 Seeded comments

As indicated in Section 5, participants are required to react at least to 5 seeded comments for each of the topics. Comments are shown in random order. Participants have the options to indicate "Approve," "Disapprove," or "Pass."

The comments are obtained from a 300-online participant sample collected in late 2024, with representation by age, gender, ethnicity, and partisanship, as reported by the Prolific platform. The distribution of self-reported opinions follow distributions clos to that of national-representative studies, as reported by Gallup (2025) and Pew Research Center (2021). Position and scores are self-reported by survey participants and verified for consistency by the research team.

I narrow down the list to 12 seeded comments in the first experiment and 9 comments, by topic, in the second experiment, ensuring that they are balanced across positions.

Response	Position	Score
Gun purchase restrictions for people with mental health conditions		
Between a quarter to a third of Americans will have a mental health problem at some point that doesn't mean you can disarm a third of adults.	Strongly oppose	0/100
I am strongly opposed to limiting gun purchases based on mental issues because diagnosis are too easily handed out so Doctors can write more Rxs	Strongly oppose	10/100
Mental illness covers a wide range of things, most of which don't make you a danger to others.	Somewhat oppose	18/100
We shouldn't suspend constitutional rights because someone has a "mental illness." Extremists think transgender people are mentally ill.	Somewhat oppose	40/100
Most crimes involving mental illness are not committed by but against those people. I don't think such laws would help them.	Somewhat oppose	55/100
It feels like a well-intended idea but I'm skeptical how it would be operationalized and enforced.	Somewhat favor	69/100
I favor people with mental illnesses being able to purchase guns as it is a constitutional right to bear arms.	Somewhat oppose	70/100
It depends on the mental illness and the severity. If a person is not dangerous, they shouldn't be discriminated against.	Somewhat favor	80/100

I somewhat agree that people with mental illness should not be able to own guns. Anxiety is different than schizophrenic narcissistic	Somewhat favor	82/100
If someone has mental issues, they don't have the capacity to think straight and giving them access to guns is dangerous	Strongly favor	94/100
Mentally ill individuals do not possess the capability to judge when a firearm can be lawfully used.	Strongly favor	100/100
People with mental illness may not have the rationality or clear judgement not to harm themselves or others with a gun.	Strongly favor	100/100

Gestational limits for abortion		
A woman has the right to have an abortion at any stage of pregnancy. Her body, her choice.	Strongly oppose	0/100
Abortion at any stage is still killing the baby and the person did not take responsibility for their actions.	Strongly oppose	0/100
I oppose considering gestational age as a determinant for the legality of abortion because it restricts reproductive rights	Somewhat oppose	22/100
I am against abortion but if a life is in danger time shouldn't matter and an abortion should be allowed to save the mother's life.	Somewhat oppose	41/100
I believe a woman should be able to term a pregnancy up to a viable fetus age. At that point, the fetus should be given up but not aborted.	Somewhat favor	60/100
I am pro choice, but at full term or almost full term it should be wrong to do abortions unless theres a medical reason to end pregnancy.	Somewhat favor	75/100
Once a fetus is viable it is unethical to abort unless the woman's life is in danger.	Strongly favor	100/100
At some point we must consider a fetus to be developed enough to have it's own right to live.	Strongly favor	100/100
I think that a baby should not be aborted after a certain time period unless there is danger to the woman.	Strongly favor	100/100

Handgun possession		
People will always get guns whether banned or not. It's not a good idea to disarm the innocent.	Strongly oppose	0/100
I strongly oppose restricting handgun possession to only law enforcement etc. because this leads to authoritarianism and tyranny.	Strongly oppose	0/100
The Constitution gives us the right to bear arms. But there should be stiffer penalties for the people that illegally have them	Somewhat oppose	12/100
If "authorized persons" includes citizens who can earn the privilege to do so then yes I think it's a good idea.	Somewhat favor	50/100
If it would get less guns off the streets would be a positive though I'm not sure if banning out right is necessarily the way.	Somewhat favor	50/100
I agree but also disagree. People will always find gun and own guns no matter what the ban is. So banning them could cause more harm.	Somewhat favor	55/100
Handgun usage needs to be better controlled. US is one of the only countries with our level of gun murders & crime	Somewhat favor	80/100
I am a firm believer that the only guns available should be those for hunting. Handguns do not keep people safer as some think.	Strongly favor	100/100
Gun control would lead to safer communities, as we can see in other countries.	Strongly favor	100/100

Photo-ID Vote requirements		
Until everyone has equal actual access to their documents then this is an unfair burden to impose.	Strongly oppose	0/100
This is overkill and would disenfranchise the elderly, handicapped and marginalized members of our nation	Strongly oppose	0/100
In theory its reasonable, but it presents roadblocks to vulnerable people	Somewhat oppose	30/100

In theory this sounds great but doesn't take into account vote by mail	Somewhat oppose	40/100
Although fraud is extremely low I think some sort of id is acceptable to ask for, perhaps issued when a voter is automatically registered.	Somewhat favor	59/100
I would have nothing against showing ID as long as it free and easy to obtain for every qualified citizen.	Somewhat favor	72/100
Yes, I think this will give total legitimacy to the process.	Strongly favor	100/100
Voting is important so people should have to prove that they are qualified in order to participate.	Strongly favor	100/100
Yes and this will promote free and fair elections because there will be no manipulation and a fraud election.	Strongly favor	100/100

B.0.2 Extreme voting

I construct an indicator for extreme voting based on endorsement behavior toward seeded comments. First, I re-scale the positions across a [-100,100] range, where negative values indicate opposition to the policy premise. Second, as detailed in the preregistered analysis plan, I obtain a participant-level weighted average by multiplying each of the votes by each of the scores. Approval of the comment corresponds to 1, while disapproval and passing correspond to -1 and zero, respectively. This yields a value between -100 and 100.

$$s_j^* \in [-100, 100] \tag{B.1}$$

$$v_{ij} = \begin{cases} 1 & \text{if participant } i \text{ approves comment } j \\ -1 & \text{if participant } i \text{ disapproves comment } j \\ 0 & \text{if participant } i \text{ passes on comment } j \end{cases} \tag{B.2}$$

$$W_i = \frac{\sum_{j=1}^{J} v_{ij} \cdot s_j^*}{\sum_{j=1}^{J} |v_{ij}|} \in [-100, 100] \tag{B.3}$$

$$\tilde{W}_i = \frac{W_i - (-100)}{100 - (-100)} = \frac{W_i + 100}{200} \in [0, 1] \tag{B.4}$$

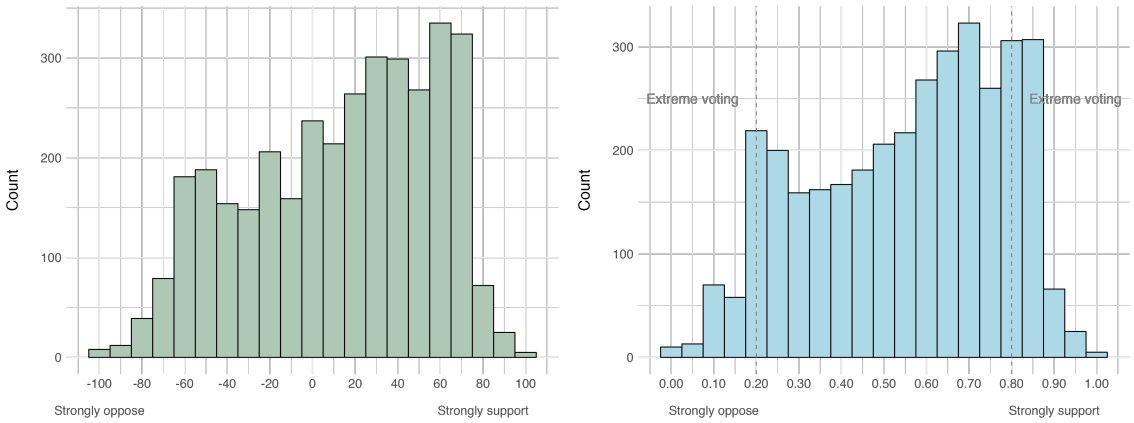
To generate an indicator of extreme voting, I re-scale the weighted average to the [0,1] interval. I then create a binary indicator equal to one if the re-scaled weighted average is equal to or below 0.20 or equal to or above 0.80. While an alternative is to obtain the 20th and 80th percentiles, I select this approach to maintain comparability with the original scale among the general population

and extremity characterization in the rest of the paper. However, the behavioral results are robust to this alternative approach.

$$\text{Extreme}_i = \mathbb{1}\{\tilde{W}_i \leq 0.20 \text{ or } \tilde{W}_i \geq 0.80\} \tag{B.5}$$

As shown in Figure B1, pooling across topics, the re-scaled weighted score is concentrated among central values with fatter right tails.

Figure B1: Voting weighted sum and re-scaled weighted score

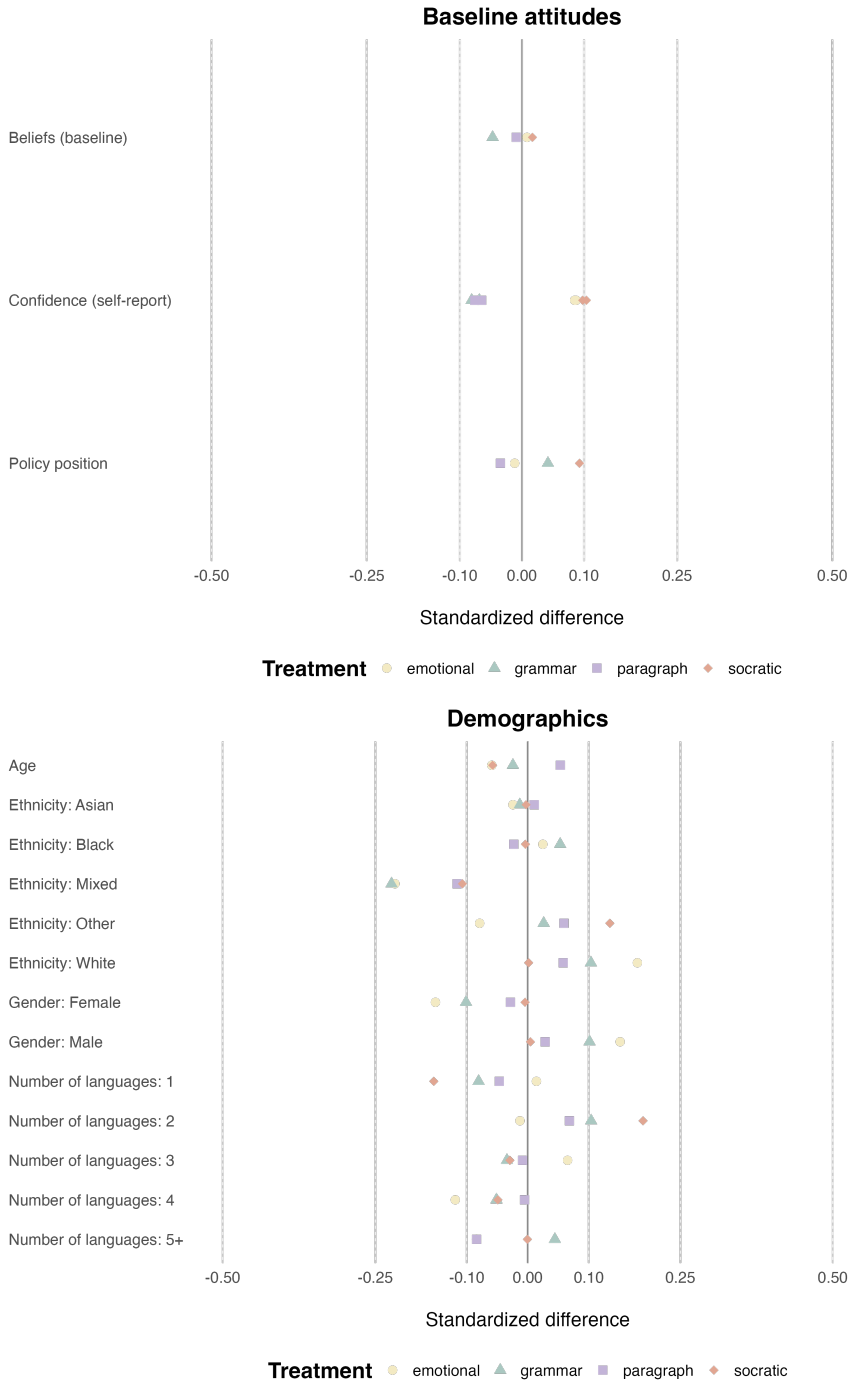


Notes: Vertical dashed lines in the right figure indicate the extreme-vote region, corresponding to the 0.20 and 0.80 thresholds.

C Randomization balance tests

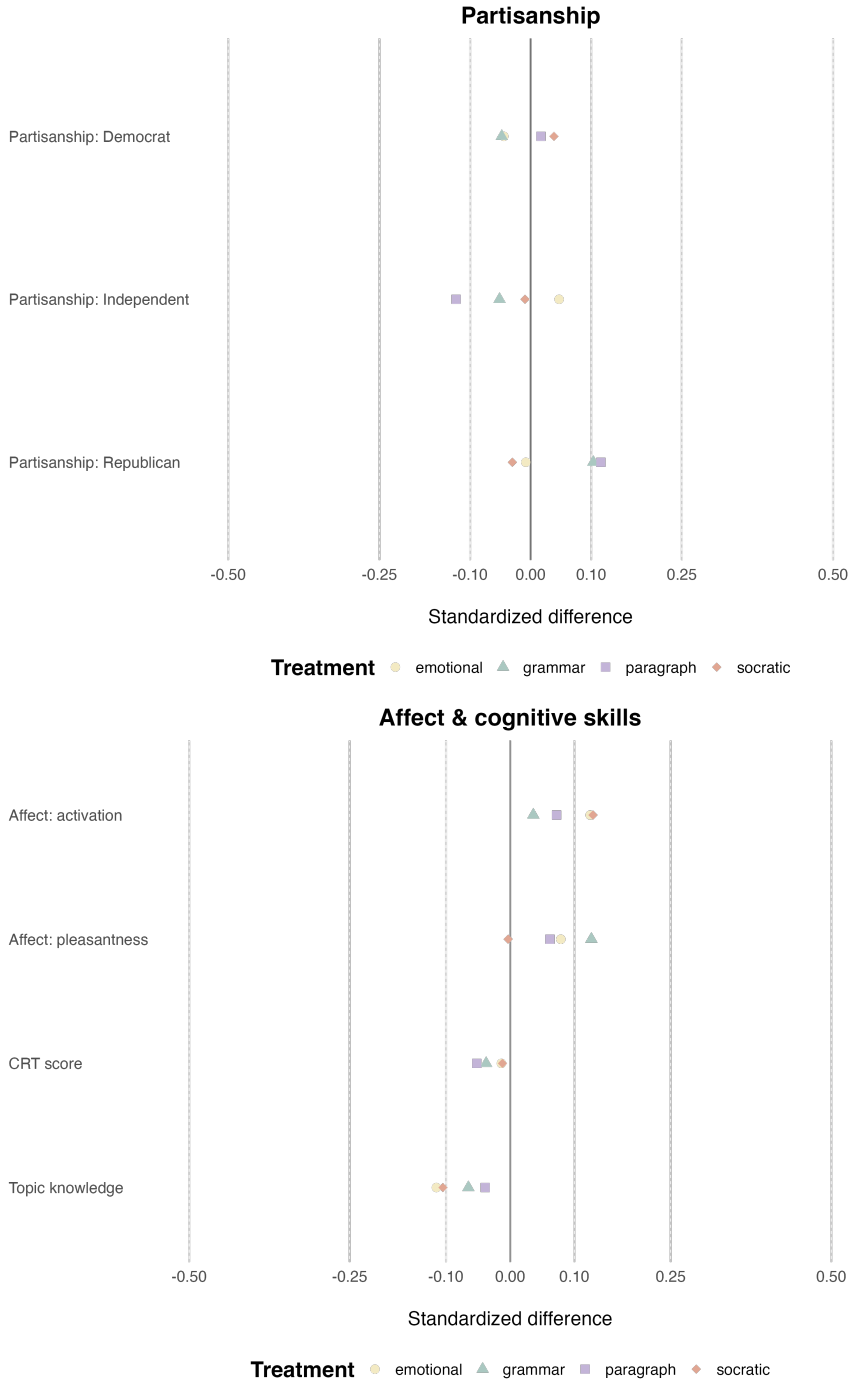
C.1 Experiment 1 Balance Analysis

Figure C1: Standardized differences in baseline position, confidence, and demographics by treatment arm



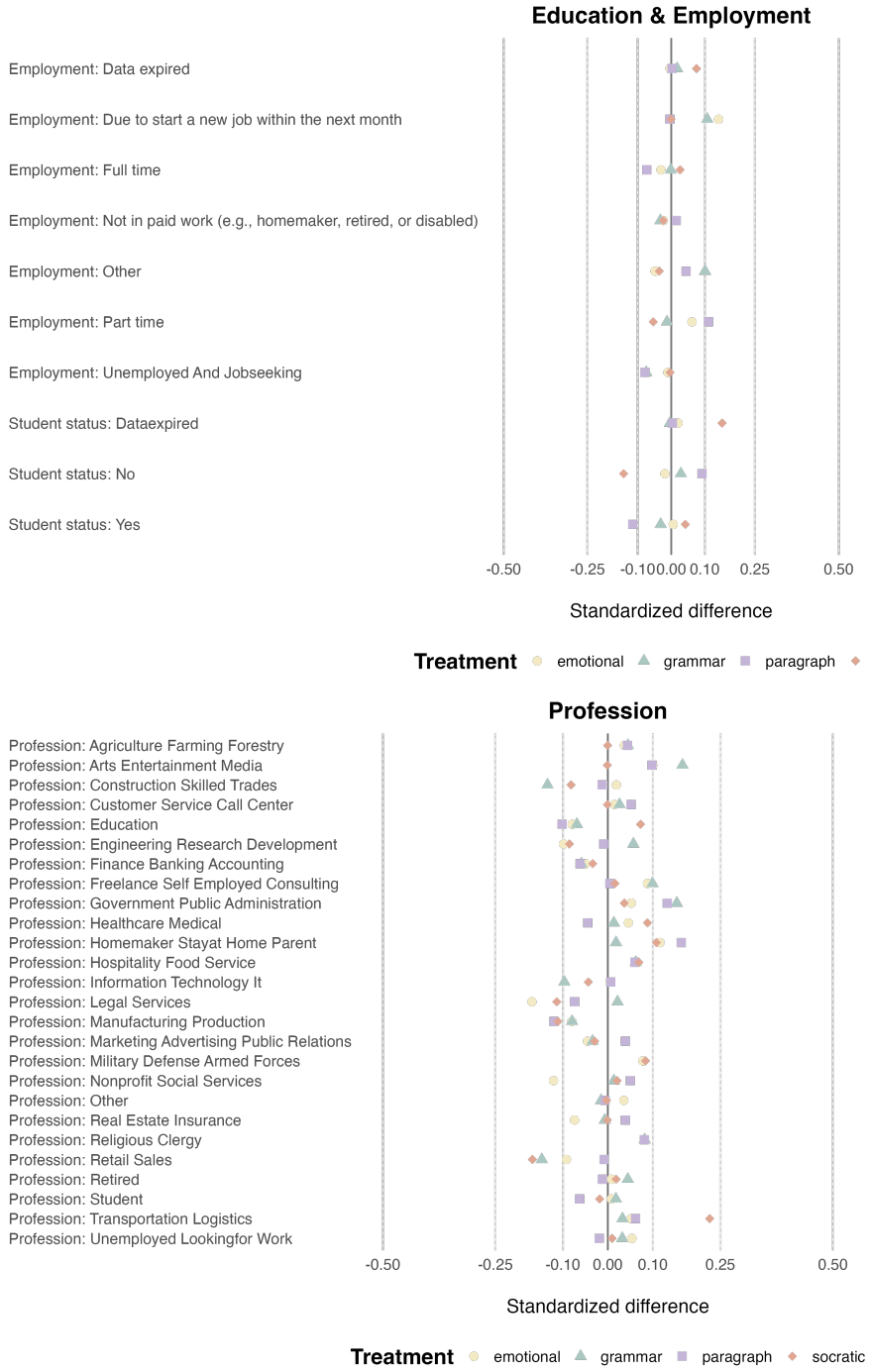
Notes: The sample includes 1,503 participants across the five treatment arms. Differences with respect to the reference category (Waiting control group) are standardized. All characteristics are reported at baseline.

Figure C2: Standardized differences in partisanship, affect, and cognitive skills by treatment arm



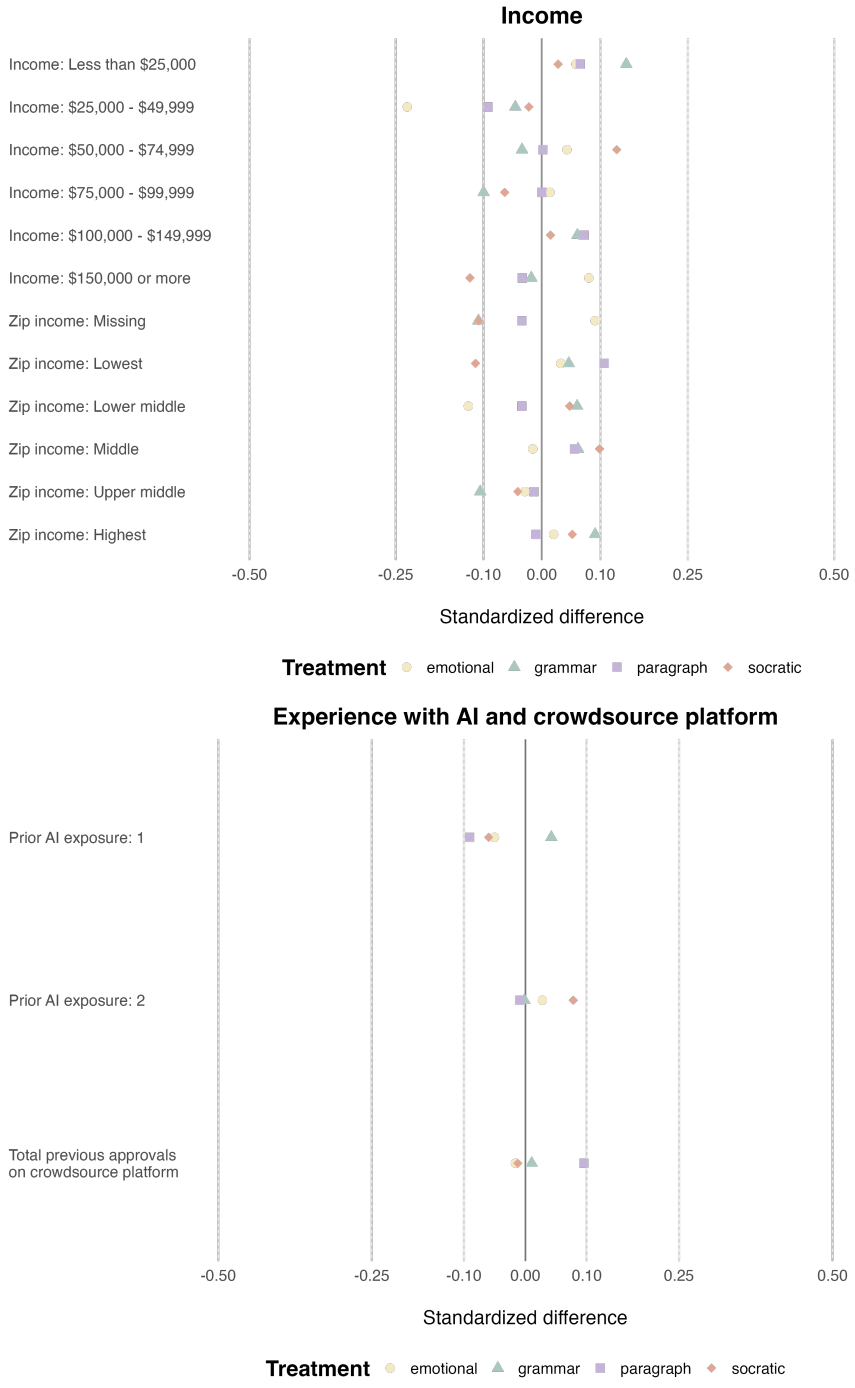
Notes: The sample includes 1,503 participants across the five treatment arms. Differences with respect to the reference category (Waiting control group) are standardized. All characteristics are reported at baseline.

Figure C3: Standardized differences in education, employment, and profession by treatment arm



Notes: The sample includes 1,503 participants across the five treatment arms. Differences with respect to the reference category (Waiting control group) are standardized. All characteristics are reported at baseline. With the exception of professions, characteristics are reported by Prolific.

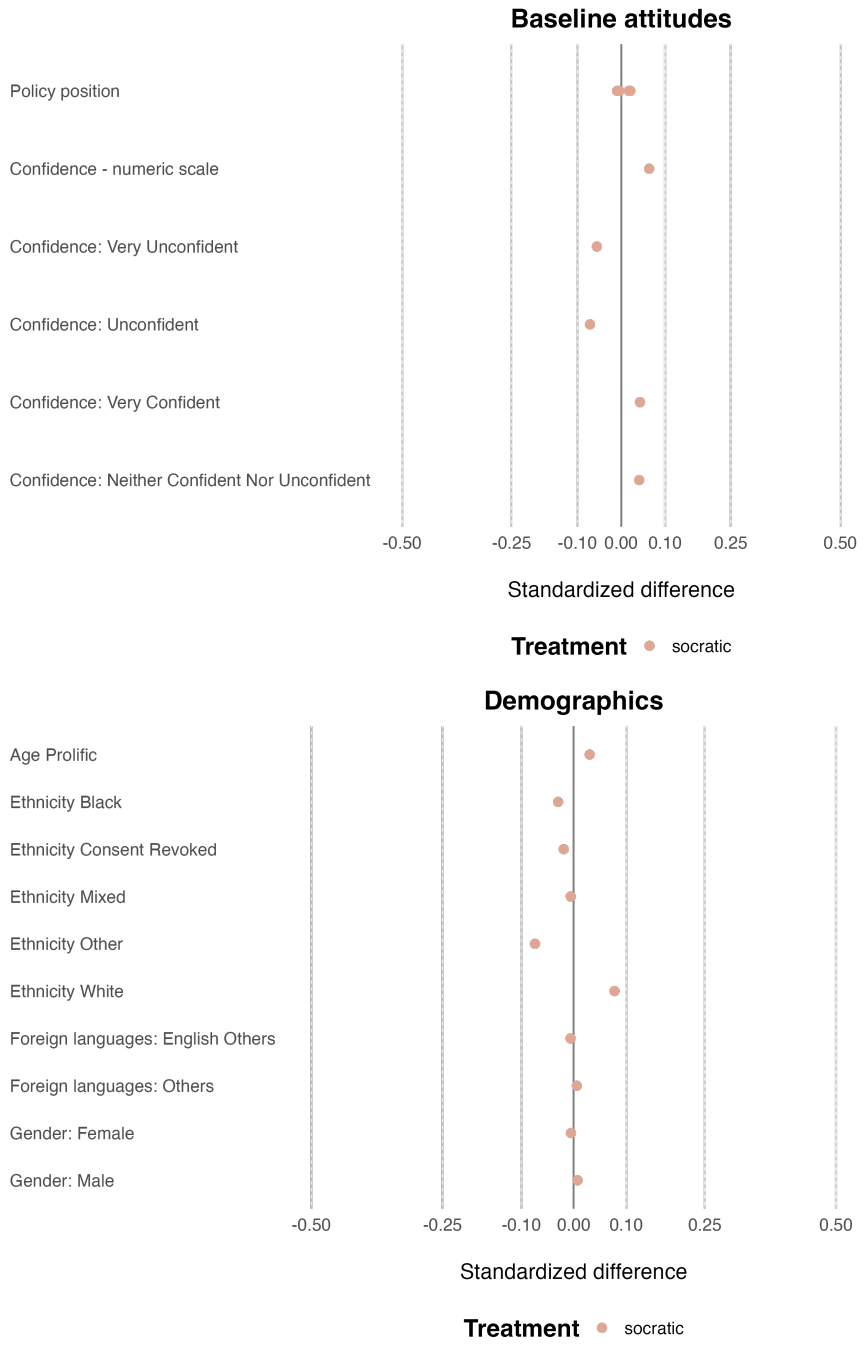
Figure C4: Standardized differences in income, previous exposure to AI, and experience in the crowdsource platform



Notes: The sample includes 1,503 participants across the five treatment arms. Differences with respect to the reference category (Waiting control group) are standardized. All characteristics are reported at baseline.

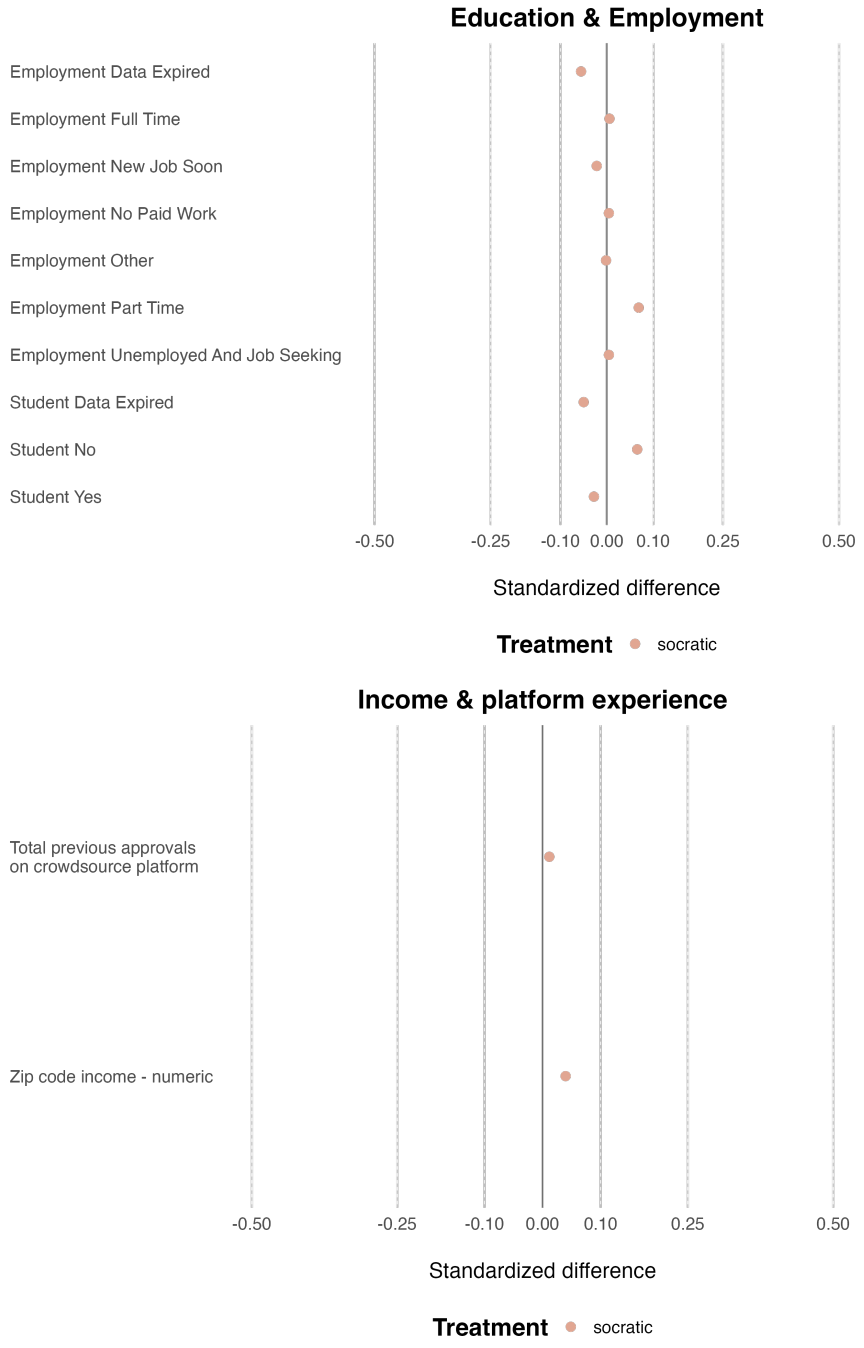
C.2 Experiment 2 Balance Analysis

Figure C5: Standardized differences in baseline position, confidence, and demographics by treatment arm



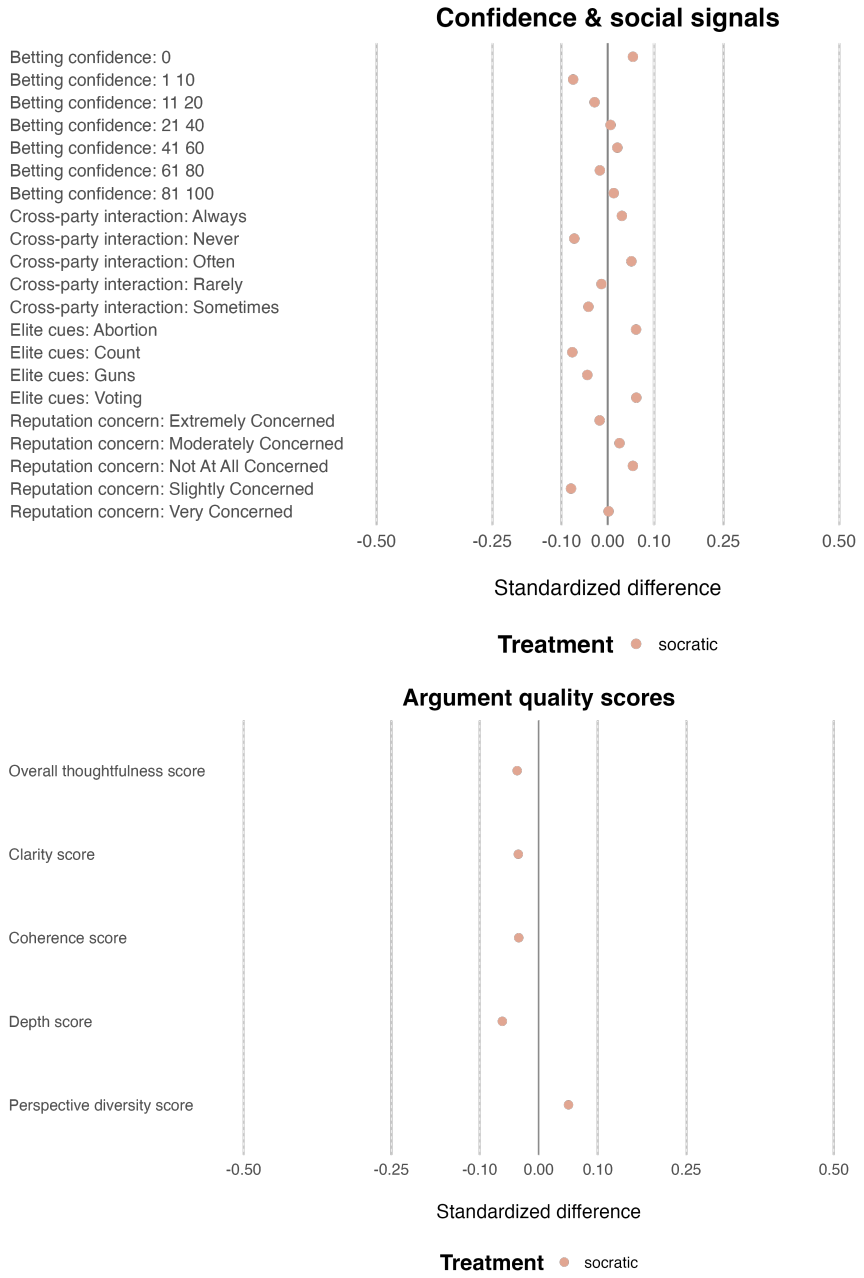
Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category (AI-Grammar group) are standardized. All characteristics are reported at baseline. Age, ethnicity, foreign language fluency, and gender are reported by Prolific.

Figure C6: Standardized differences in education, employment, income, and experience in the crowdsource platform by treatment arm



Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category (AI-Grammar group) are standardized. All characteristics are reported at baseline. Employment and student status is reported by Prolific.

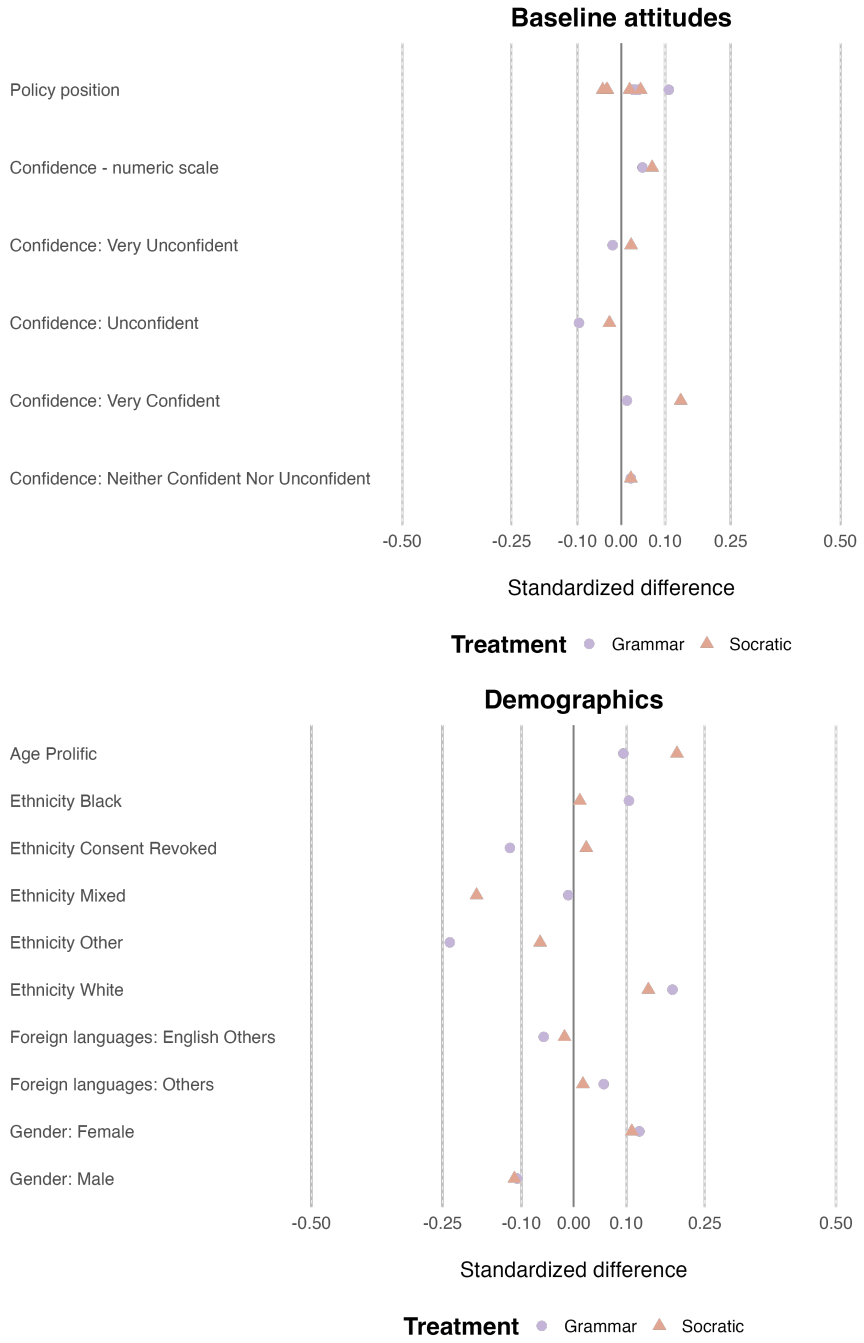
Figure C7: Standardized differences in confidence, cross-partisan and social interactions, and argument quality scores by treatment arm



Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category (AI-Grammar group) are standardized. All characteristics are reported at baseline.

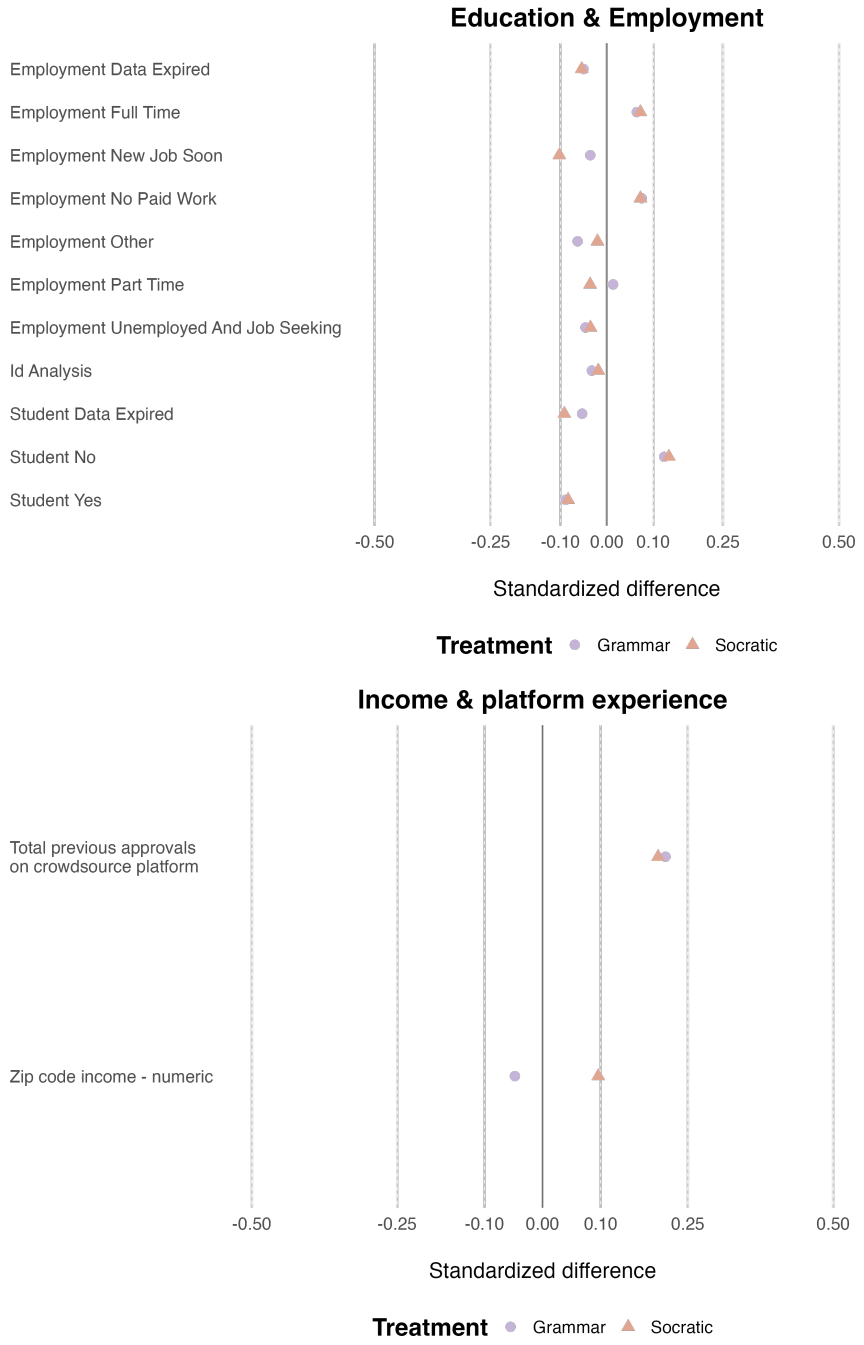
C.3 Experiment 2 Follow-up Attrition Analysis

Figure C8: Standardized differences in confidence, policy position, and demographics by follow-up participation status



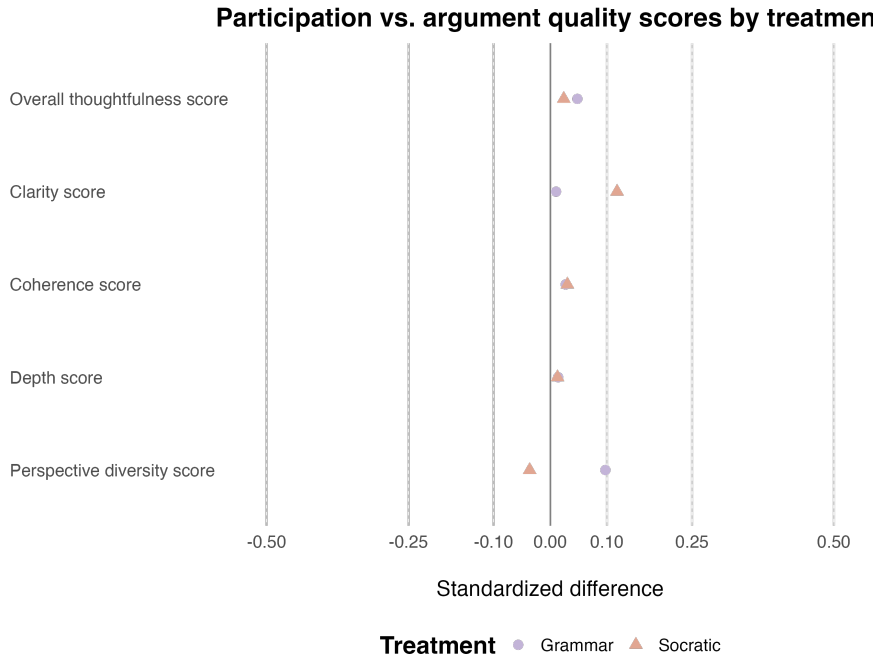
Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category (not participating in the follow-up survey) are standardized. All characteristics are reported at baseline Age, ethnicity, language fluency, and gender are reported by Prolific.

Figure C9: Standardized differences in education employment, and income by follow-up participation status



Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category (not participating in the follow-up survey) are standardized. All characteristics are reported at baseline. Employment and student status are reported by Prolific.

Figure C10: Standardized differences in argument quality scores by follow-up participation status



Notes: The sample includes 3,518 participants across the two treatment arms. Differences with respect to the reference category a(not participating in the follow-up survey) are standardized. All characteristics are reported at baseline.

D Results

D.1 First stage

Table D1: Response length by treatment arm and experiment

Treatment	N	Mean	SD	Median	Min	Max
Experiment 1						
Waiting Control	284	–	–	–	–	–
AI-Emotional	330	37.2	27.2	31	5	266
AI-Grammar	299	37.7	24.5	33	1	210
Paragraph	303	53.4	31.3	49	1	162
AI-Socratic	287	34.3	22.5	27	8	130
Experiment 2						
AI-Grammar	1796	9.6	16.9	5	1	356
AI-Socratic	1722	70.2	63.4	55	1	598

Notes: Length is measured as the number of words, operationalized by whitespace-delimited tokens counts.

D.2 Policy issue position

D.2.1 Experiment 1

Table D2: Experiment 1: Treatment effects on self-reported positioning and confidence

	Self-reported position				Self-reported confidence	
	1[Extreme at Post]		1[Moved to topic median]		Confidence at post	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Socratic	−0.109	−0.101**	0.046**	0.222	−0.178*	−0.204
	(0.071)	(0.042)	(0.019)	(0.183)	(0.093)	(0.144)
AI-Emotional	−0.040	−0.031	0.031*	0.161	−0.002	−0.095
	(0.068)	(0.046)	(0.018)	(0.135)	(0.140)	(0.151)
Reflective Paragraph	0.082	0.020	0.012	0.015	−0.041	0.023
	(0.066)	(0.054)	(0.019)	(0.016)	(0.112)	(0.162)
AI-Grammar	−0.019	0.028	0.010	0.015	−0.059	−0.079
	(0.069)	(0.050)	(0.022)	(0.016)	(0.141)	(0.170)
1[Pre extreme]		0.851***		−0.321		0.040
		(0.034)		(0.242)		(0.073)
AI-Socratic × 1[Pre extreme]		0.007		−0.176		0.047
		(0.046)		(0.182)		(0.141)
AI-Emotional × 1[Pre extreme]		0.033		−0.130		0.132
		(0.045)		(0.138)		(0.138)
Reflective Paragraph × 1[Pre extreme]		−0.019		−0.002		−0.083
		(0.048)		(0.021)		(0.130)
AI-Grammar × 1[Pre extreme]		−0.029		−0.005		0.034
		(0.049)		(0.023)		(0.125)
Num. obs	1,503	1,503	1,503	1,503	1,503	1,503
Control mean	0.655	0.655	0.049	0.049	4.077	4.077
Control SD	0.476	0.476	0.217	0.217	1.295	1.295
Test: Waiting control = AI-Socratic	-0.109		0.046		-0.178	
	(0.071)		(0.019)		(0.093)	
p-value (two-sided)	0.123		0.015		0.056	
Test: Pre Extreme in Waiting control = Pre Extreme in AI-Socratic		-0.094		0.046		-0.157
		(0.039)		(0.019)		(0.092)
p-value (two-sided)		0.018		0.015		0.089
Socratic effect (pp)	−0.109	−0.094	0.046	0.046	−0.178	−0.157
SE (Socratic, pp)	(0.071)	(0.040)	(0.019)	(0.019)	(0.093)	(0.092)
p-value (Socratic)	0.123	0.018	0.015	0.015	0.056	0.089
R^2	0.031	0.717	0.085	0.145	0.706	0.707
Adjusted R^2	0.022	0.707	0.074	0.104	0.701	0.701

Notes: Table reports treatment effects on self-reported positioning and confidence. Odd-numbered columns report average treatment effects. Even-numbered columns report conditional average treatment effects by extreme position. All estimates include pre-treatment self-reported position, confidence in a 5-item scale, partisanship, and PDS-LASSO selected covariates. Heteroskedasticity-robust standard errors are reported in parentheses. Control group statistics are calculated for the Waiting condition. *** p<0.01, ** p<0.05, * p<0.1 for two-sided tests.

D.2.2 Experiment 2

Main survey

Table D3: Experiment 2: Treatment effects on self-reported positioning and confidence

	Self-reported position				Self-reported confidence	
	1[Extreme at post]		1[Moved to topic median]		Confidence at post	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Socratic	0.001	−0.006	0.014	0.018	−0.090	−0.119
	(0.017)	(0.025)	(0.017)	(0.024)	(0.080)	(0.096)
1[Pre extreme]	0.796***	0.683***	0.049***	0.074***	1.050***	0.944***
	(0.012)	(0.021)	(0.012)	(0.022)	(0.059)	(0.085)
AI-Socratic × 1[Pre extreme]	−0.031**	−0.040	0.009	0.027	−0.072	−0.090
	(0.018)	(0.031)	(0.017)	(0.031)	(0.087)	(0.121)
Handguns		−0.175***		−0.127***		−0.076
		(0.016)		(0.016)		(0.072)
1[Pre extreme] × Handguns		0.293***		−0.059***		0.195**
		(0.020)		(0.021)		(0.102)
AI-Socratic × Handguns		−0.004		−0.003		0.051
		(0.022)		(0.021)		(0.145)
AI-Socratic × 1[Pre extreme] × Handguns		0.031		−0.027		0.051
		(0.029)		(0.030)		(0.145)
Voter ID		−0.053***		0.004		−0.068
		(0.022)		(0.022)		(0.072)
1[Pre extreme] × Voter ID		0.048**		−0.018		0.125
		(0.028)		(0.029)		(0.101)
AI-Socratic × Voter ID		0.020		−0.008		0.045
		(0.032)		(0.030)		(0.103)
AI-Socratic × 1[Pre extreme] × Voter ID		−0.010		−0.023		−0.006
		(0.041)		(0.041)		(0.144)
Num. obs	7,036	7,036	7,036	7,036	7,036	7,036
Control mean	0.570	0.570	0.106	0.106	2.517	2.517
Control SD	0.495	0.495	0.308	0.308	1.442	1.442
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order
Fixed effects	Topic		Topic		Topic	
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	-0.030	-0.046/-0.018/-0.036	0.023	0.046/0.016/0.015	-0.162	-0.209/-0.114/-0.170
	(0.016)	(0.023)/(0.012)/(0.023)	(0.017)	(0.026)/(0.014)/(0.024)	(0.083)	(0.100)/(0.100)/(0.104)
p-value (one-sided)	0.032	0.043/0.143/0.121	0.079	0.073/0.252/0.545	0.025	0.037/0.251/0.101
R^2	0.650	0.668	0.076	0.078	0.222	0.223
Adjusted R^2	0.649	0.667	0.071	0.073	0.219	0.219

Notes: The table reports treatment effects on self-reported positioning and confidence measures. All outcomes are binary indicators. Observations include 3,518 participants across two topics. Odd-numbered columns report treatment effects pooling the three topics (abortion gestational limits, banning possession of handguns, voter ID requirements). Even-numbered columns report treatment effects by topic. All models include pre-treatment self-reported position in a 1-10 scale and confidence in a 5-item scale, self-reported partisanship and additional LASSO-selected covariates, if applicable. Standard errors are clustered at the individual level and reported in parentheses. Control group statistics are calculated for the AI-Grammar condition. *** p<0.01, ** p<0.05, * p<0.1 for one-sided tests.

Follow-up survey

Table D4: Experiment 2: Treatment effects on self-reported positioning and confidence at one-month follow-up

	Self-reported position				Self-reported confidence	
	1[Extreme position at post]		1[Moved to topic median]		Confidence at post	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Socratic	−0.007	−0.014	−0.044*	0.011	0.119	0.270**
	(0.043)	(0.056)	(0.032)	(0.044)	(0.115)	(0.156)
1[Pre extreme]	0.447***	0.410***	0.054**	0.183***	0.579***	0.565***
	(0.029)	(0.045)	(0.028)	(0.044)	(0.086)	(0.132)
AI-Socratic × 1[Pre extreme]	−0.030	−0.076	0.008	0.045	−0.108	−0.149
	(0.043)	(0.069)	(0.041)	(0.068)	(0.127)	(0.205)
Handguns		−0.041		0.033		0.025
		(0.044)		(0.039)		(0.032)
1[Pre extreme] × Handguns		0.030		−0.037		0.102
		(0.058)		(0.058)		(0.171)
AI-Socratic × Handguns		0.039		−0.051		−0.139
		(0.066)		(0.059)		(0.186)
AI-Socratic × 1[Pre extreme] × Handguns		0.082		−0.103		0.072
		(0.086)		(0.085)		(0.255)
Voter ID		−0.032		0.208***		0.019
		(0.049)		(0.048)		(0.037)
1[Pre extreme] × Voter ID		0.081		−0.372***		−0.093
		(0.064)		(0.066)		(0.193)
AI-Socratic × Voter ID		−0.030		−0.108*		−0.342*
		(0.073)		(0.074)		(0.212)
AI-Socratic × 1[Pre extreme] × Voter ID		0.042		0.032		0.085
		(0.096)		(0.099)		(0.290)
Num. obs	2,369	2,369	2,369	2,369	2,369	2,369
Control mean	0.564	0.564	0.259	0.259	2.498	2.498
Control SD	0.496	0.496	0.438	0.438	1.451	1.451
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	-0.037	-0.090/0.031/-0.078	-0.036	0.056/-0.099/-0.021	0.011	0.121/0.054/-0.136
	(0.041)	(0.056)/(0.050)/(0.054)	(0.028)	(0.052)/(0.044)/(0.042)	(0.105)	(0.157)/(0.143)/(0.154)
p-value (one-sided)	0.185	0.110/0.536/0.145	0.100	0.282/0.024/0.624	0.458	0.442/0.704/0.378
R^2	0.292	0.295	0.018	0.048	0.218	0.221
Adjusted R^2	0.283	0.285	0.010	0.038	0.210	0.211

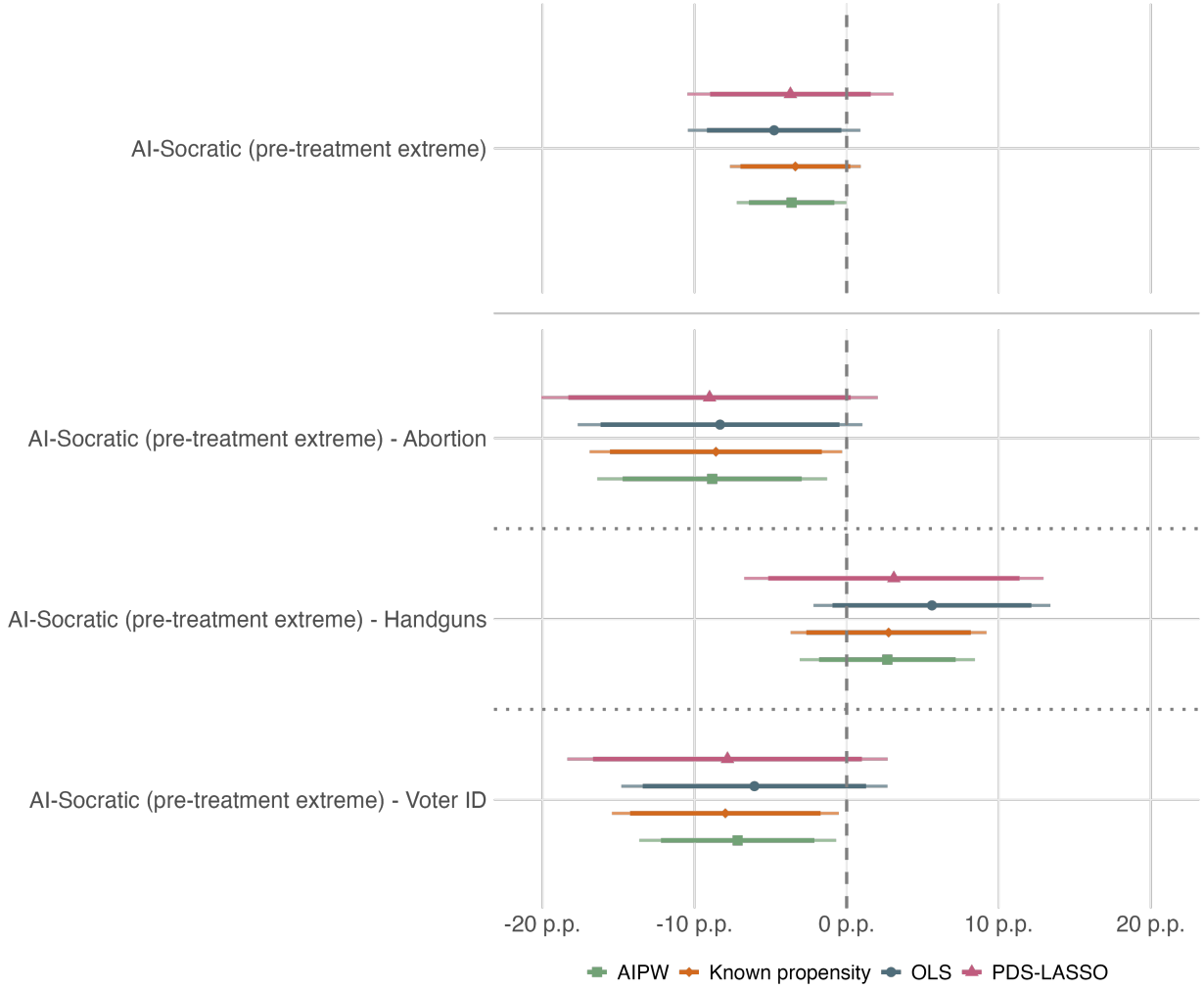
Notes: The table reports treatment effects on self-reported positioning and confidence measures. All outcomes are binary indicators. Observations include 3,518 participants across two topics. Odd-numbered columns report treatment effects pooling the three topics (abortion gestational limits, banning possession of handguns, voter ID requirements). Even-numbered columns report treatment effects by topic. All models include pre-treatment self-reported position in a 1-10 scale and confidence in a 5-item scale, self-reported partisanship, predicted probability of participating in the follow-up survey based on all the set of covariates, and additional LASSO-selected covariates, if applicable. Standard errors are clustered at the individual level and reported in parentheses. Control group statistics are calculated for the AI-Grammar condition. *** p<0.01, ** p<0.05, * p<0.1 for one-sided tests.

Table D5: Experiment 2: Treatment effects on self-reported positioning and confidence in main survey on one-month follow-up sample

	Self-reported position				Self-reported confidence	
	1[Extreme position at post]		1[Moved to topic median]		Confidence at post	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Socratic	0.003	0.020	0.014	−0.0003	−0.003	0.021
	(0.019)	(0.029)	(0.016)	(0.028)	(0.020)	(0.026)
1[Pre extreme]	0.813***	0.703***	−0.001	−0.062***	0.0001	0.003
	(0.013)	(0.025)	(0.011)	(0.023)	(0.017)	(0.021)
AI-Socratic × 1[Pre extreme]	−0.035**	−0.048	0.006	−0.010	0.008	−0.002
	(0.021)	(0.038)	(0.017)	(0.034)	(0.025)	(0.033)
Handguns		−0.165***		−0.149***		0.042**
		(0.019)		(0.019)		(0.024)
1[Pre extreme] × Handguns		0.276***		0.075***		0.010
		(0.023)		(0.022)		(0.030)
AI-Socratic × Handguns		−0.016		−0.005		−0.026
		(0.027)		(0.027)		(0.036)
AI-Socratic × 1[Pre extreme] × Handguns		0.040		0.009		0.013
		(0.035)		(0.033)		(0.045)
Voter ID		−0.044**		−0.094***		0.052**
		(0.027)		(0.024)		(0.028)
1[Pre extreme] × Voter ID		0.038		0.110***		−0.021
		(0.034)		(0.030)		(0.033)
AI-Socratic × Voter ID		0.013		0.011		−0.043
		(0.041)		(0.036)		(0.041)
AI-Socratic × 1[Pre extreme] × Voter ID		−0.012		0.034		0.021
		(0.051)		(0.045)		(0.048)
Num. obs	4,738	4,738	4,738	4,738	4,738	4,738
Control mean	0.569	0.569	0.072	0.072	0.118	0.118
Control SD	0.495	0.495	0.258	0.258	0.323	0.323
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order
Fixed effects	Topic		Topic		Topic	
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	-0.033	-0.027/-0.003/-0.026	0.019	-0.010/-0.006/0.035	-0.010	0.003/-0.010/-0.019
	(0.016)	(0.024)/(0.008)/(0.025)	(0.014)	(0.019)/(0.007)/(0.022)	(0.020)	(0.024)/(0.026)/(0.027)
p-value (one-sided)	0.022	0.122/0.353/0.147	0.081	0.307/0.210/0.055	0.300	0.449/0.352/0.239
R^2	0.663	0.680	0.054	0.064	0.047	0.048
Adjusted R^2	0.661	0.678	0.049	0.057	0.042	0.042

Notes: The table reports treatment effects on self-reported positioning and confidence measures in the main survey. All outcomes are binary indicators. Observations include 2,369 participants that completed the follow-up survey. Odd-numbered columns report treatment effects pooling the three topics (abortion gestational limits, banning possession of handguns, voter ID requirements). Even-numbered columns report treatment effects by topic. All models include pre-treatment self-reported position in a 1-10 scale and confidence in a 5-item scale, self-reported partisanship, predicted probability of participating in the follow-up survey based on all the set of covariates, and additional LASSO-selected covariates, if applicable. Standard errors are clustered at the individual level and reported in parentheses. Control group statistics are calculated for the AI-Grammar condition. *** p<0.01, ** p<0.05, * p<0.1 for one-sided tests.

Figure D1: Experiment 2: Conditional treatment effects on position extremity among initially extreme participants at one-month follow-up on initial topic, overall and by topic



Notes: The figure indicates average treatment effects among extreme participants on a binary indicator of extreme positioning on the first topic under discussion one month after treatment. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Bars indicate 90 and 95% confidence intervals from one-sided test, with standard errors clustered at the participant level.

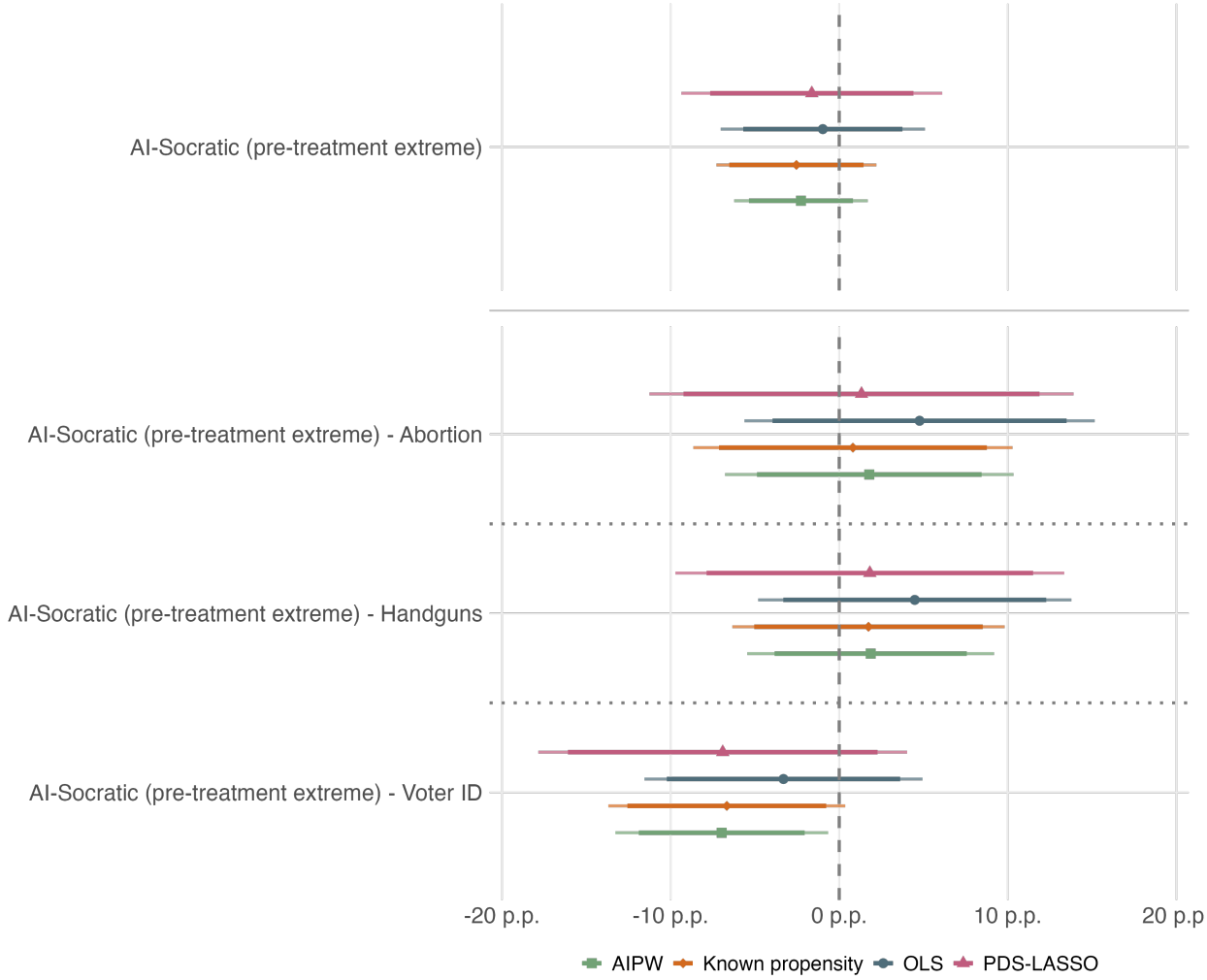
Spillovers

Table D6: Experiment 2: Spillover effects on self-reported positioning and confidence at one-month follow-up (on new, untreated topic)

	Self-reported position				Self-reported confidence	
	1[Extreme position at post]		1[Moved to topic median]		Confidence at post	
	(1)	(2)	(3)	(4)	(5)	(6)
AI-Socratic	0.065*	−0.038	−0.016	−0.023	0.212*	0.012
	(0.050)	(0.073)	(0.035)	(0.055)	(0.137)	(0.198)
1[Pre extreme]	0.140***	0.028	0.183***	0.238***	0.278***	0.302**
	(0.030)	(0.053)	(0.026)	(0.048)	(0.088)	(0.157)
AI-Socratic × 1[Pre extreme]	−0.082**	0.051	0.052*	0.035	−0.109	0.068
	(0.045)	(0.080)	(0.039)	(0.070)	(0.133)	(0.230)
Handguns		−0.102**		0.0003		−0.002
		(0.054)		(0.045)		(0.039)
1[Pre extreme] × Handguns		0.139**		−0.106**		−0.032
		(0.071)		(0.064)		(0.208)
AI-Socratic × Handguns		0.162**		0.026		0.246
		(0.085)		(0.068)		(0.239)
AI-Socratic × 1[Pre extreme] × Handguns		−0.157*		0.036		−0.134
		(0.108)		(0.092)		(0.311)
Voter ID		−0.047		−0.077**		0.031
		(0.050)		(0.040)		(0.035)
1[Pre extreme] × Voter ID		0.171***		−0.053		−0.029
		(0.067)		(0.061)		(0.197)
AI-Socratic × Voter ID		0.127*		0.0003		0.279*
		(0.078)		(0.061)		(0.217)
AI-Socratic × 1[Pre extreme] × Voter ID		−0.209**		0.007		−0.327
		(0.100)		(0.089)		(0.290)
Num. obs	2,369	2,369	2,369	2,369	2,369	2,369
Control mean	0.606	0.606	0.281	0.281	2.622	2.622
Control SD	0.489	0.489	0.450	0.450	1.425	1.425
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	-0.016	0.013/0.018/-0.069	0.036	0.012/0.074/0.019	0.031	0.046/0.047/0.005
	(0.047)	(0.064)/(0.059)/(0.056)	(0.036)	(0.055)/(0.047)/(0.051)	(0.028)	(0.048)/(0.042)/(0.039)
p-value (one-sided)	0.365	0.837/0.758/0.216	0.165	0.829/0.119/0.716	0.134	0.332/0.259/0.897
p-value (one-sided)	0.365	0.837/0.758/0.216	0.165	0.829/0.119/0.716	0.209	0.659/0.252/0.841
R^2	0.099	0.103	0.130	0.132	0.091	0.093
Adjusted R^2	0.087	0.088	0.123	0.122	0.080	0.080

Notes: The table reports treatment effects on self-reported positioning and confidence measures. All outcomes are binary indicators. Observations include 3,518 participants across two topics. Odd-numbered columns report treatment effects pooling the three topics (abortion gestational limits, banning possession of handguns, voter ID requirements). Even-numbered columns report treatment effects by topic. All models include pre-treatment self-reported position in a 1-10 scale and confidence in a 5-item scale, self-reported partisanship, predicted probability of participating in the follow-up survey based on all the set of covariates, and additional LASSO-selected covariates, if applicable. Standard errors are clustered at the individual level and reported in parentheses. Control group statistics are calculated for the AI-Grammar condition. *** p<0.01, ** p<0.05, * p<0.1 for one-sided tests.

Figure D2: Experiment 2: Spillover effects on position extremity among initially extreme participants at one-month follow-up on a new topic, overall and by topic



Notes: The figure indicates average treatment effects among extreme participants on a binary indicator of extreme positioning on a new topic one month after treatment. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Bars indicate 90 and 95% confidence intervals, calculated using one-sided tests with standard errors clustered at the participant level.

D.3 Argumentative quality

D.3.1 Experiment 1

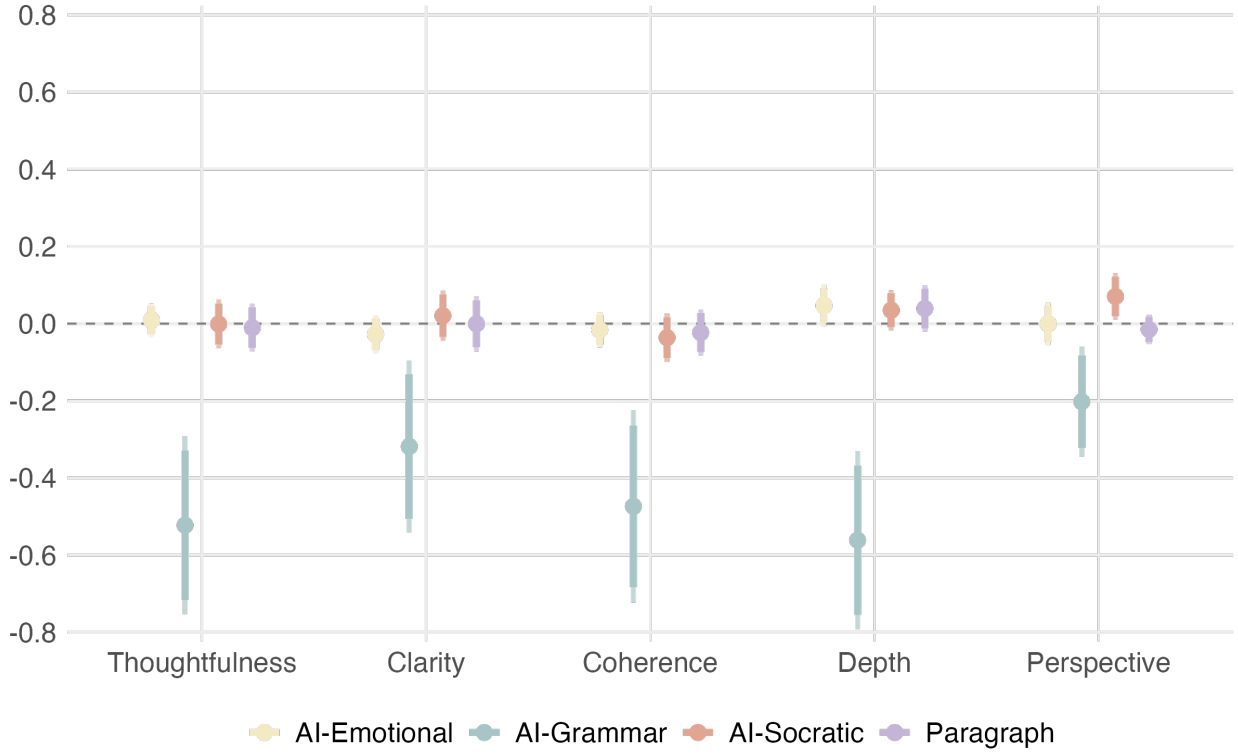
Argumentative quality scores

Table D7: Conditional average treatment effects on argumentative quality scores

	Score				
	Thoughtfulness	Clarity	Coherence	Depth	Perspective diversity
	(1)	(2)	(3)	(4)	(5)
Waiting Control (Intercept)	0.164***	0.070***	0.090***	0.135***	0.026*
	(0.030)	(0.023)	(0.025)	(0.028)	(0.014)
AI-Emotional	−0.049	0.010	0.016	−0.030	0.060
	(0.036)	(0.031)	(0.035)	(0.038)	(0.039)
AI-Grammar	−0.929***	−0.623***	−0.606***	−0.569***	−0.135
	(0.200)	(0.216)	(0.215)	(0.201)	(0.113)
Reflective Paragraph	−0.080*	0.042	−0.011	−0.051	0.031
	(0.042)	(0.052)	(0.042)	(0.041)	(0.028)
AI-Socratic	0.013	−0.039	0.047	−0.005	0.047
	(0.063)	(0.044)	(0.048)	(0.041)	(0.044)
1[Pre-treatment extreme]	−0.045	0.030	0.026	−0.060	−0.003
	(0.035)	(0.032)	(0.029)	(0.039)	(0.009)
AI-Emotional x Extreme	0.058	−0.038	−0.032	0.077	−0.060
	(0.045)	(0.040)	(0.043)	(0.054)	(0.053)
AI-Grammar x Extreme	0.406***	0.304	0.133	0.008	−0.067
	(0.282)	(0.295)	(0.305)	(0.283)	(0.165)
Reflective Paragraph x Extreme	0.069	−0.042	−0.012	0.090	−0.046
	(0.061)	(0.074)	(0.055)	(0.061)	(0.035)
AI-Socratic x Extreme	−0.014	0.060	−0.083	0.040	0.023
	(0.079)	(0.068)	(0.062)	(0.054)	(0.056)
Num. obs	1,503	1,503	1,503	1,503	1,503
Pre-treatment control mean	4.049	2.595	2.574	1.986	0.306
Pre-treatment control SD	1.004	0.596	0.655	0.600	0.577
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Test: Pre Extreme in Waiting = Pre Extreme in AI-Socratic					
(p-value)	0.975	0.539	0.256	0.191	0.028
R^2	0.724	0.706	0.693	0.726	0.853
Adjusted R^2	0.685	0.663	0.650	0.689	0.842

Notes: Scores are LLM-generated using a fine-tuned model for each of the rubrics as detailed in Section 5.3.1. Table shows standardized scores. All estimates include pre-treatment score, self-reported position and confidence in a 5-item scale, and PDS-LASSO selected covariates. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ for two-sided tests. Heteroskedasticity-robust standard errors in parentheses.

Figure D3: Experiment 1: Conditional average treatment effects on argumentative quality scores among extreme participants



Notes: The figure indicates average effects among extreme participants on standardized argumentative quality scores. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the waiting control condition. Bars indicate 90 and 95% confidence intervals, with heteroskedasticity-robust standard errors.

D.3.2 Experiment 2

Argumentative quality scores

Table D8: Experiment 2: Average treatment effects on argumentative quality scores

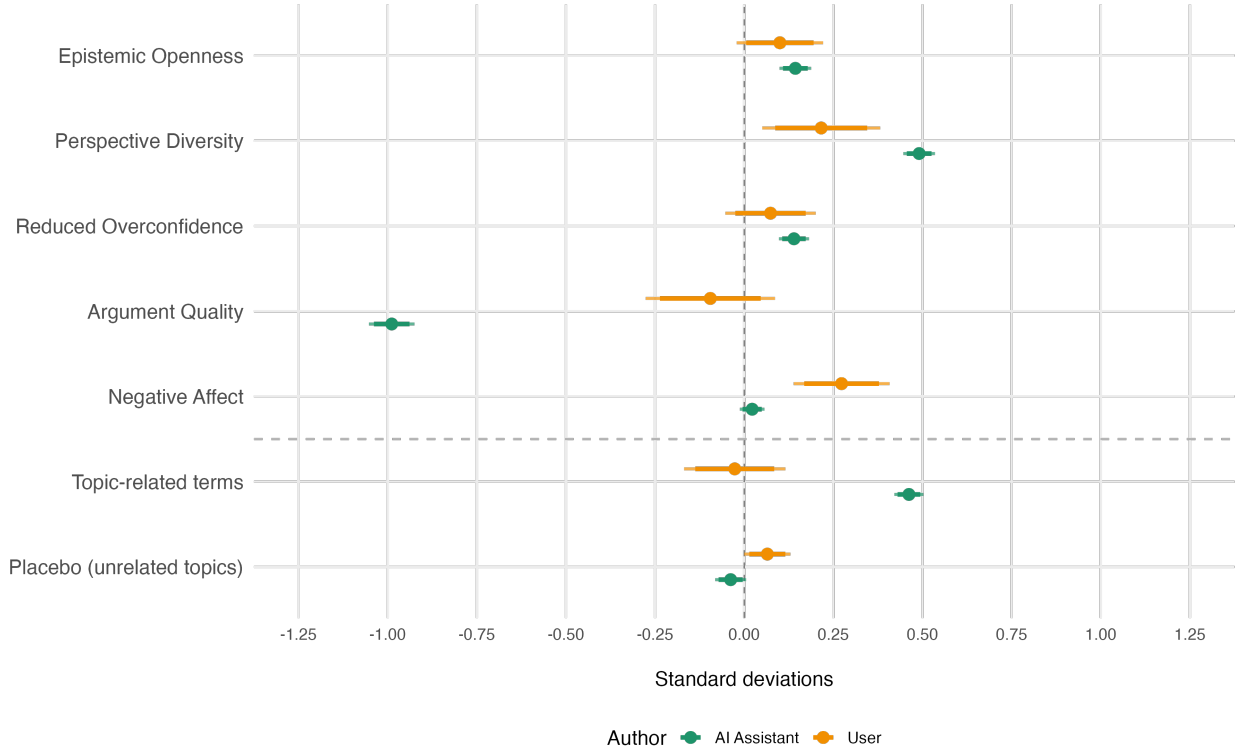
	Score				
	Thoughtfulness	Clarity	Coherence	Depth	Perspective diversity
	(1)	(2)	(3)	(4)	(5)
Waiting Control (Intercept)	-0.023	0.071**	-0.195***	0.059**	-0.062
	(0.027)	(0.032)	(0.026)	(0.027)	(0.051)
1[Pre-treatment extreme]	-0.017	-0.095***	0.044**	-0.010	0.126***
	(0.022)	(0.026)	(0.020)	(0.023)	(0.049)
AI-Socratic	-0.035*	0.011	-0.002	-0.013	0.018
	(0.020)	(0.025)	(0.018)	(0.022)	(0.045)
AI-Socratic x 1[Pre-treatment extreme]	0.021	0.024	-0.047*	-0.003	-0.048
	(0.027)	(0.034)	(0.026)	(0.030)	(0.032)
Num. obs	7,036	7,036	7,036	7,036	7,036
Pre-treatment Control Mean	2.420	1.922	0.905	0.942	0.065
Pre-treatment Control SD	1.596	1.038	1.096	0.988	0.286
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order	Topic-order
Fixed effects	Topic	Topic	Topic	Topic	Topic
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic					
(p-value)	0.819	0.001	0.883	0.485	0.045
R^2	0.706	0.538	0.732	0.692	0.616
Adjusted R^2	0.705	0.536	0.731	0.691	0.615

Notes: Scores are LLM-generated using a fine-tuned model for each of the rubrics as detailed in Section 5.3.1. While thoughtfulness score ranges from 0 to 6, the rest of the scores range from 0 to 3. All estimates include pre-treatment score, self-reported position and confidence in a 5-item scale, and PDS-LASSO selected covariates. All estimates include topic and topic-order fixed effects. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ for one-sided tests. Standard errors clustered at the participant level in parentheses.

Conversational text analysis

The dictionaries for each index were created to test the theoretical predictions of the epistemic openness framework. Each index is constructed by normalizing all text (lowercase, removing special characters); tokenizing into individual words; stemming using the Snowball algorithm; removing stopwords from a standard English stopword list (i.e., SMART, onix, and Snowball); counting matches to each dictionary; and normalizing by total tokens per document.

Figure D4: Linguistic differences between AI-Socratic and AI-Grammar treatments among non-extreme participants



Notes: Figure indicates average effects among non-extreme participants on standardized linguistic indices. The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar condition. Bars indicate 90 and 95% confidence intervals for one-sided tests, with standard errors clustered at the participant level.

Table D9: Dictionary terms for text analysis indices

Index	Terms (unstemmed)
Epistemic Openness (Composite)	
Uncertainty	unsure, uncertain, not sure, don't know, unclear, hard to say, difficult to say, depends, it depends, varies, case by case, situation, circumstance
Modal softening	might, may, could, seems, appear, probably, possibly, perhaps, maybe
Receptivity	fair point, valid point, see your point, make sense, hadn't thought, hadn't considered, didn't realize, hadn't known, didn't think of, changed my mind, reconsidered, rethinking
Perspective Diversity (Composite)	

Tradeoff recognition	tradeoff, trade-off, tradeoffs, trade-offs, balance, balanced, weigh, weighing, nuanced, nuance, complicated, complex, complexity, gray area, grey area, not simple, not straightforward, both, also, too, however, but, yet, although, though, while, whereas, on one hand, on the other hand, at the same time, competing, tension, conflict, dilemma, paradox
Perspective-taking	perspective, perspectives, viewpoint, viewpoints, view, views, standpoint, angle, lens, position, positions, others, other, those who, people who, someone, someone who, opponents, supporters, critics, advocates, proponents, side, sides, other side, both sides, either side, different views, various perspectives, alternative view, opposing view, different perspective, another perspective, imagine, consider, think about, from their, in their, put myself, step back, see from, look from
Group considerations	women, woman, men, man, children, child, families, family, parents, parent, mothers, mother, fathers, father, workers, worker, employees, employee, businesses, business, communities, community, people, person, individuals, individual, those affected, people affected, those who, people who, vulnerable, disadvantaged, marginalized, affect, affects, affected, impact, impacts, impacted, harm, harms, harmed, hurt, hurts, suffer, suffers, benefit, benefits, help, helps, protect, protects, burden, hardship, struggle, consequences, consequence

Reduced overconfidence

Reduce overconfidence	somewhat, fairly, rather, quite, relatively, reasonably, moderately, slightly, partially, partly, kind of, sort of, a bit, a little, more or less, to some extent, to a degree, I think, I believe, I would say, I'd say, in my view, in my opinion, personally, it seems, it appears, it looks like, around, about, approximately, roughly, generally, typically, often, usually, sometimes

Argument quality (Composite)

Evidence	evidence, data, study, studies, research, statistics, statistic, survey, poll, polls, report, reports, finding, findings, shows, show, showed, indicates, indicate, demonstrates, demonstrate, suggests, suggest, proves, prove, proven, confirms, confirm
Reasoning	because, since, cause, causes, caused, effect, effects, leads to, led to, results in, due to, therefore, thus, hence, consequently, so, reason, reasons, if then, logically, logical, follows that, implies, imply, means that

Negative affect

Negative affect	Terms from NRC Emotion Lexicon (Mohammad and Turney, 2013): anger and fear categories (140 total unique terms)
Policy-topic-related terms	
Topic-related terms	abortion, pregnancy, pregnant, gestation, handgun, handguns, firearm, firearms, weapon, weapons, ammunition, election, elections, ballot, identification, ID, register, registration
Placebo	
Unrelated policy topics	migration, asylum, visa, climate change, global warming, carbon, emissions, renewable energy, fossil fuel, greenhouse, medicare, medicaid, deductible, healthcare

D.4 Additional results

D.4.1 Experiment 2

Table D10: Experiment 2: Perceived bias by treatment group and extremity

	AI is neutral		Research team is neutral	
	Main	Follow-up	Main	Follow-up
	(1)	(2)	(3)	(4)
AI-Socratic	-0.011	-0.001	-0.023	-0.044
	(0.061)	(0.066)	(0.048)	(0.066)
1[Pre-extreme extreme]	0.009	-0.041*	-0.023	-0.028
	(0.022)	(0.025)	(0.022)	(0.027)
AI-Socratic × 1[Pre-extreme extreme]	-0.049*	-0.031	-0.021	0.024
	(0.032)	(0.038)	(0.033)	(0.040)
Num obs.	3,518	2,369	3,518	2,369
Control mean	0.816	0.850	0.764	0.761
Control SD	0.387	0.357	0.425	0.427
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	-0.060	-0.032	-0.044	-0.020
	(0.062)	(0.066)	(0.049)	(0.066)
p-value (two-sided)	0.335	0.624	0.363	0.761
R^2	0.050	0.063	0.032	0.029
Adjusted R^2	0.041	0.048	0.023	0.014

Notes: Outcomes are binary indicators equal to 1 if the participant responded "I did not think the AI chatbot [research team] was biased" when asked about the neutrality of the AI conversational agent and the research team. All estimates include self-reported position, confidence in a 5-item scale, partisanship, and PDS-LASSO selected covariates. Follow-up estimates also include the predicted probability of participating in the follow-up survey based on all the set of covariates. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ for two-sided tests. Heteroskedasticity-robust standard errors in parentheses.

Table D11: Experiment 2: Perceived agency by treatment group and extremity

	AI pressure		Research team pressure	
	Main	Follow-up	Main	Follow-up
	(1)	(2)	(3)	(4)
AI-Socratic	-0.026	-0.127	-0.024	0.076
	(0.120)	(0.177)	(0.102)	(0.152)
1[Pre extreme]	-0.108**	-0.127**	-0.163***	-0.090*
	(0.050)	(0.061)	(0.052)	(0.061)
AI-Socratic × 1[Pre extreme]	0.066	0.085	0.039	0.017
	(0.073)	(0.091)	(0.075)	(0.092)
Num obs.	3,518	2,369	3,518	2,369
Control mean	-0.172	-0.142	-0.081	-0.094
Control SD	0.870	0.871	0.950	0.911
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	0.040	-0.042	0.014	0.093
	(0.120)	(0.174)	(0.103)	(0.154)
p-value (two-sided)	0.737	0.812	0.891	0.547
R^2	0.064	0.053	0.035	0.029
Adjusted R^2	0.054	0.036	0.027	0.016

Notes: The outcome is a standardized 5-item Likert scale (1 to 5) indicating whether the participant felt pressured by the AI conversational agent or the research team when expressing their views. All estimates include self-reported position, confidence in a 5-item scale, partisanship, and PDS-LASSO selected covariates. Follow-up estimates include the predicted probability of participating in the follow-up survey based on all the set of covariates. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ for two-sided tests. Heteroskedasticity-robust standard errors in parentheses.

Table D12: Experiment 2: Openness and information seeking by treatment group and extremity

	Openness		Information seeking	
	Main	Follow-up	Deliberation topics	Other polarized topics
	(1)	(2)	(3)	(4)
AI-Socratic	0.056	0.051	-0.029	-0.009
	(0.066)	(0.115)	(0.072)	(0.081)
1[Pre-treatment extreme]	-0.039**	-0.045*	-0.012	-0.051**
	(0.023)	(0.030)	(0.025)	(0.028)
AI-Socratic × 1[Pre-treatment extreme]	0.030	0.049	-0.011	0.094**
	(0.033)	(0.045)	(0.037)	(0.042)
Num obs.	3,518	2,369	2,369	2,369
Control mean	0.486	0.579	0.166	0.391
Control SD	0.500	0.494	0.419	0.488
Covariates	PDS-LASSO	PDS-LASSO	PDS-LASSO	PDS-LASSO
Fixed effects	Topic-order	Topic-order	Topic-order	Topic-order
Test: Pre Extreme in AI-Grammar = Pre Extreme in AI-Socratic	0.086	0.100	-0.040	0.084
	(0.066)	(0.112)	(0.072)	(0.081)
p-value (one-sided)	0.097	0.186	0.288	0.147
R^2	0.187	0.045	0.091	0.126
Adjusted R^2	0.176	0.028	0.074	0.109

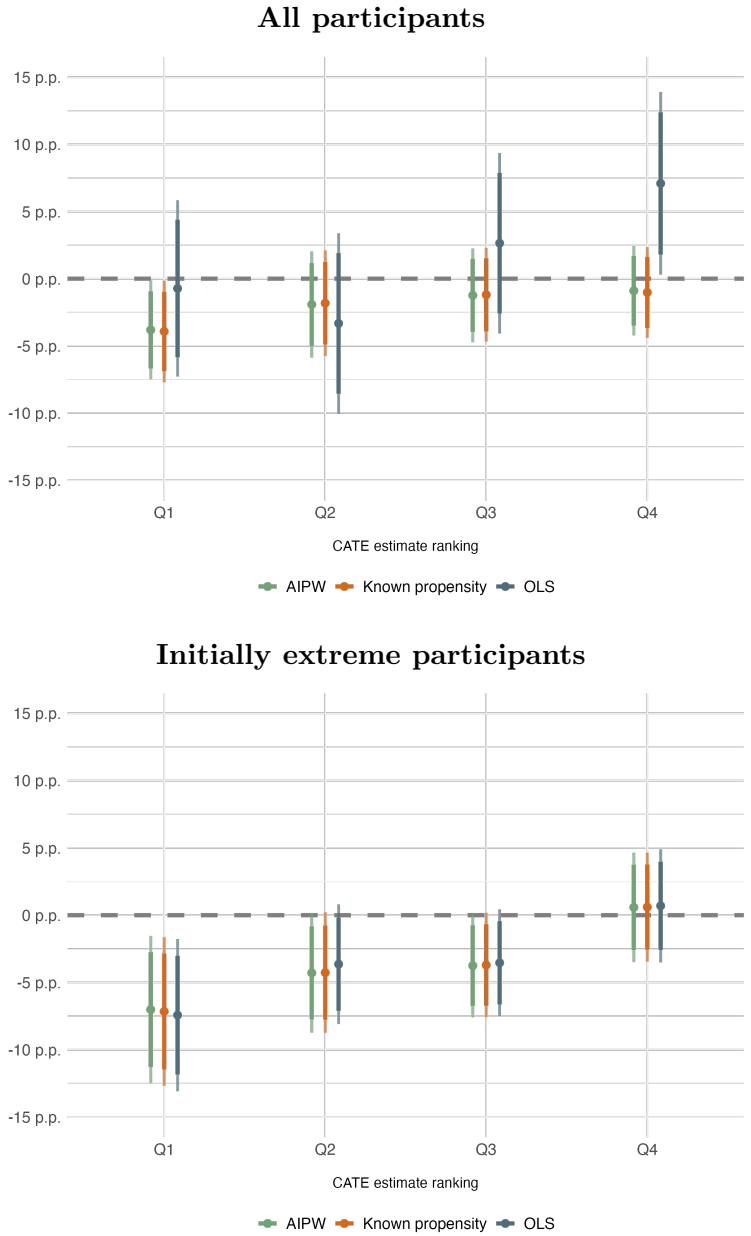
Notes: Columns (1) and (2) outcomes are binary indicators equal to 1 if the participant responded "Even if I don't agree with them, I understand people have good reasons for disagreeing with me" or "It is important to understand people who disagree with me by imagining how things look from their perspective" when asked about understanding, recognizing, or being open to the perspectives of others. Information seeking in columns (3) and (4) refers to binary indicators equal to 1 for self-reported information seeking behavior on the two deliberation topics covered in the main survey or three other contentious topics (police funding, climate and energy policy, immigration policy), respectively. All estimates include self-reported position, confidence on a 5-item scale, partisanship, and PDS-LASSO selected covariates. Follow-up estimates include the predicted probability of participating in the follow-up survey based on the full set of pre-treatment covariates. Estimates in columns (3) and (4) also include the (demeaned) number of topics searched among non-contentious topics: household recycling, local public library services and hours, and public transportation routes and schedules. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$ for one-sided tests. Heteroskedasticity-robust standard errors in parentheses.

D.5 Heterogeneous effects

D.5.1 Experiment 2

Position extremity at post-treatment

Figure D5: Average CATE within each ranking as defined by predicted CATE



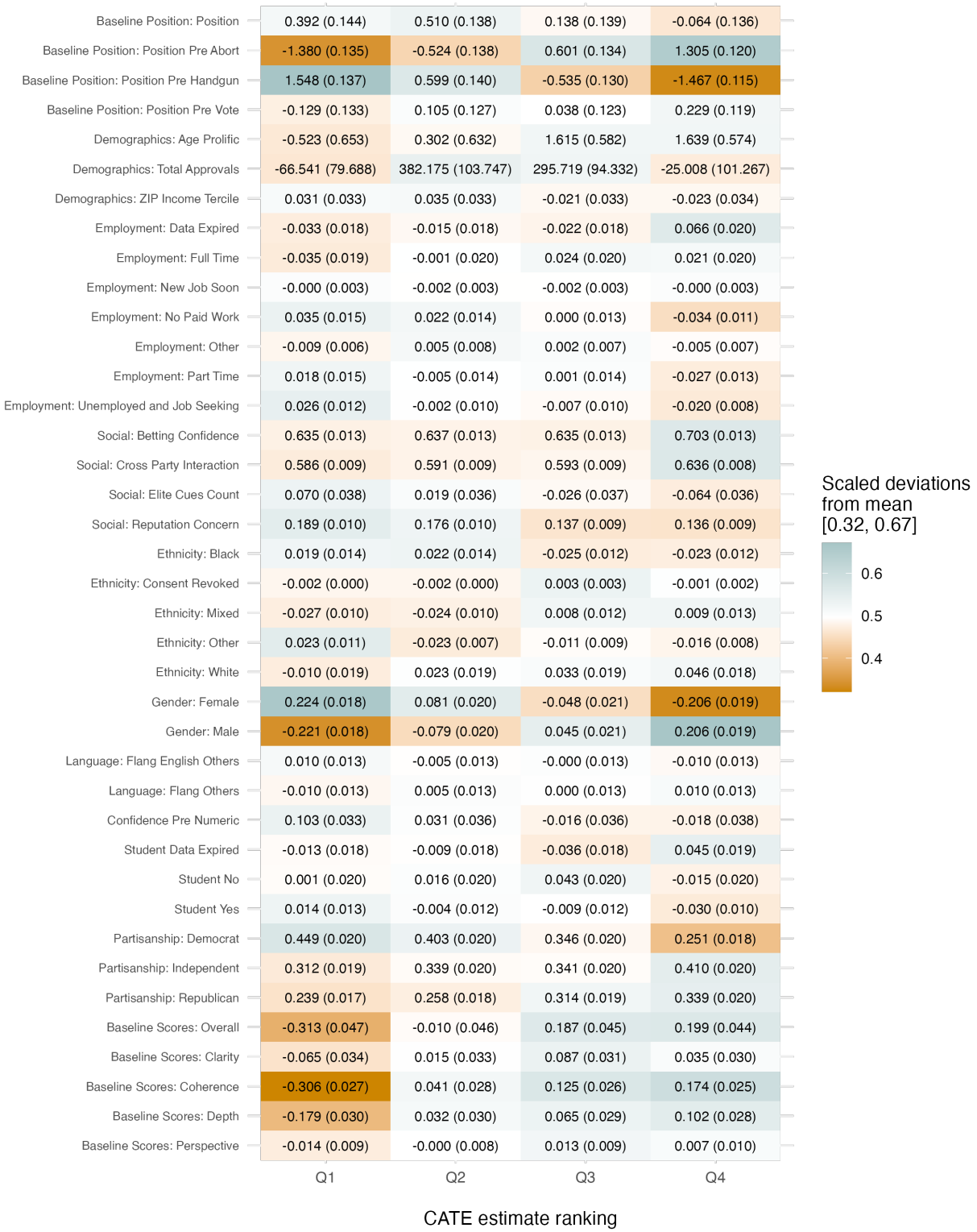
Notes: The figure displays conditional average treatment effects by quartile of predicted CATE on post-treatment extremity. Rankings derive from causal forest estimates with cross validation. Q1 denotes the highest predicted CATE quartile and Q4 the lowest. The top panel reports results for all participants; the bottom panel restricts to those extreme at baseline. Each panel compares estimates from OLS, AIPW, and AIPW with known propensity, calculated using one-sided tests with standard errors clustered at the individual participant level.

Table D13: Conditional average treatment effects by predicted CATE quartile

Quartile	Estimate	Std. Error	p-value	
			Unadjusted	RW-adjusted
Panel A: All participants				
OLS				
Q1	−0.007	0.040	0.429	0.812
Q2	−0.033	0.041	0.208	0.606
Q3	0.026	0.041	0.742	0.939
Q4	0.071	0.041	0.959	0.955
AIPW				
Q1	−0.038	0.022	0.041	0.164
Q2	−0.019	0.024	0.192	0.469
Q3	−0.012	0.021	0.288	0.472
Q4	−0.009	0.020	0.342	0.472
Known propensity				
Q1	−0.039	0.023	0.038	0.130
Q2	−0.018	0.024	0.207	0.510
Q3	−0.012	0.021	0.296	0.520
Q4	−0.010	0.021	0.323	0.520
Panel B: Initially extreme participants				
OLS				
Q1	−0.074**	0.034	0.004	0.017
Q2	−0.036	0.027	0.096	0.260
Q3	−0.035	0.024	0.102	0.260
Q4	0.007	0.026	0.598	0.607
AIPW				
Q1	−0.070**	0.033	0.005	0.021
Q2	−0.043	0.027	0.059	0.170
Q3	−0.038	0.023	0.086	0.170
Q4	0.006	0.025	0.583	0.564
Known propensity				
Q1	−0.072**	0.034	0.005	0.019
Q2	−0.043	0.027	0.061	0.163
Q3	−0.037	0.024	0.090	0.178
Q4	0.006	0.025	0.586	0.604

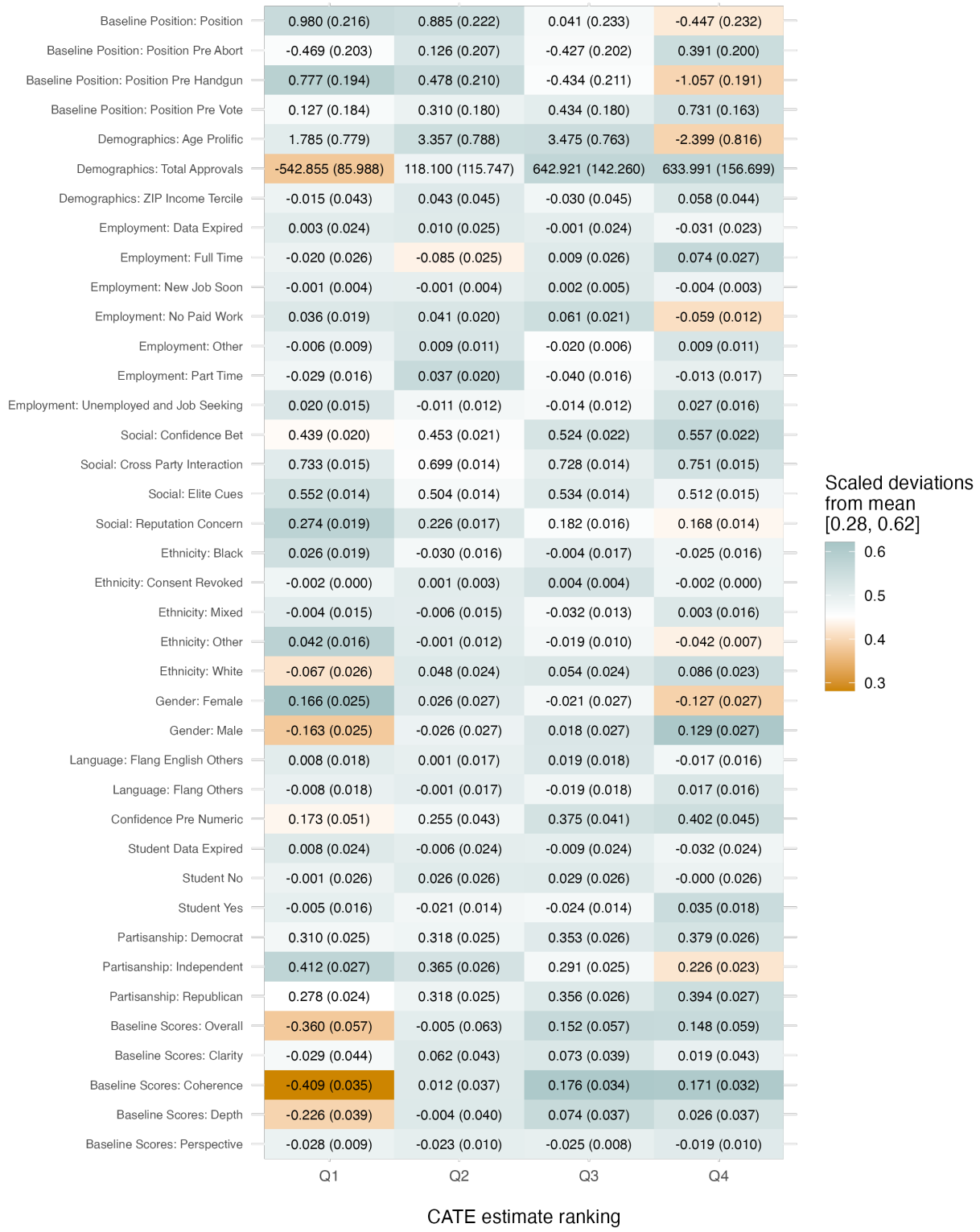
Notes: The table reports conditional average treatment effects on post-treatment position extremity by quartile of predicted CATE. Rankings are constructed using honest five-fold cross-validation with causal forests: within each fold, observations are ranked into quartiles based on out-of-bag CATE predictions from forests trained on remaining folds. Q1 denotes the highest predicted CATE quartile (strongest estimated treatment effect) and Q4 the lowest. Panel A presents results for all participants. Panel B restricts the sample to participants who held extreme positions at baseline (1, 2, 9 or 10 on the 10-point scale), with rankings constructed separately within this subsample. Within each panel, estimates are reported using three approaches: difference-in-means from OLS regressions, and augmented inverse propensity weighted (AIPW) estimates using forest-estimated and known (0.5) propensity scores, respectively. Standard errors are heteroskedasticity-consistent and clustered at the participant level. Unadjusted p-values are from one-sided tests. RW-adjusted p-values control the family-wise error rate across quartiles using the Romano–Wolf stepdown procedure with 2,000 bootstrap replications. * denotes RW-adjusted $p < 0.10$; ** denotes RW-adjusted $p < 0.05$; *** denotes RW-adjusted $p < 0.01$.

Figure D6: Experiment 2: Average covariate values within each CATE estimate ranking group



Notes: The heatmap presents average covariate values within each conditional average treatment effect (CATE) ranking group on post-treatment extremity. Rankings come from a causal forest fit with honest sample splitting (Athey and Imbens, 2016). Q1 denotes the highest predicted CATE quartile and Q4 the lowest. Point estimates and clustered standard errors are shown within every tile. Stronger colors indicate larger deviations from the sample mean.

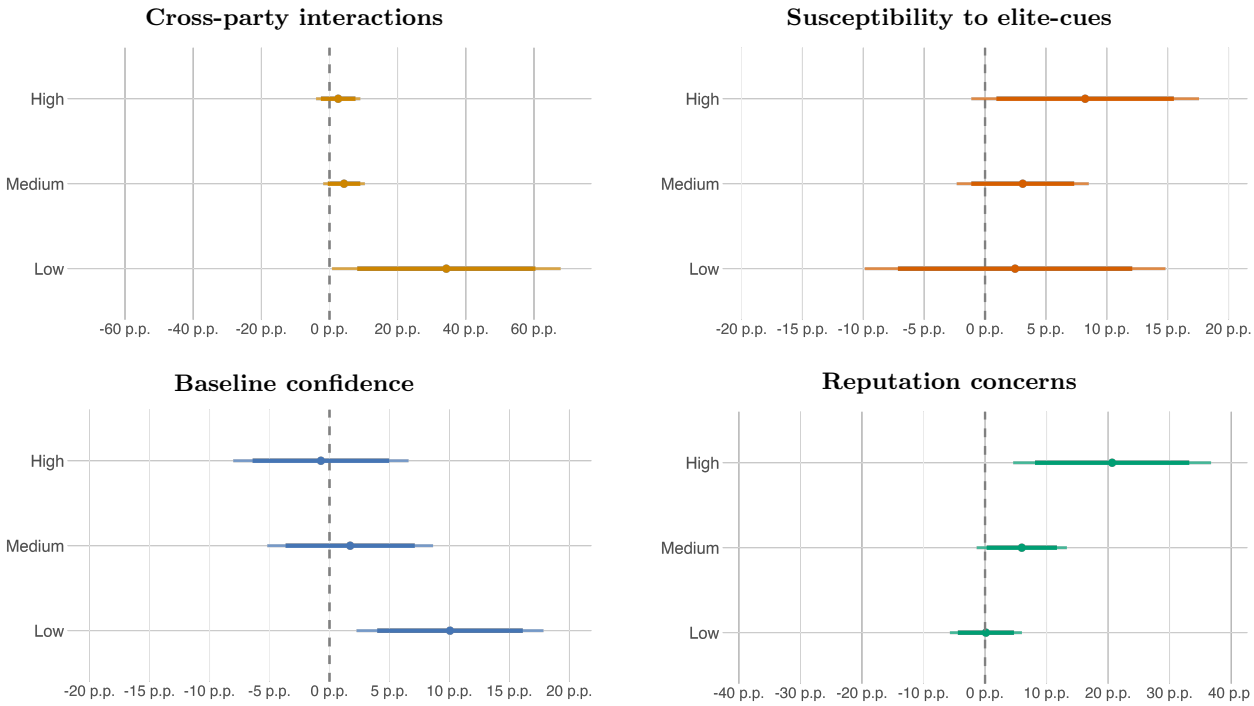
Figure D7: Experiment 2: Average covariate values within each CATE estimate ranking group among initially extreme participants



Notes: The heatmap presents average covariate values within each conditional average treatment effect (CATE) ranking group on post-treatment extremity among extreme participants at pre-treatment. Rankings come from a causal forest fit with five-way honest sample splitting (Athey and Imbens, 2016). Q1 denotes the highest predicted CATE quartile and Q4 the lowest. Point estimates and clustered standard errors are shown within every tile. Stronger colors indicate larger deviations from the sample mean.

Position moderation at post-treatment

Figure D8: Experiment 2: Heterogeneous treatment effects on position moderation



Notes: Each panel presents conditional average treatment effects by baseline participant characteristics, presented as the difference between extreme AI-Socratic and extreme AI-Grammar on position moderation at post-treatment (i.e., distance to the baseline median is smaller than at pre-treatment). Top-left: cross-party interaction frequency (Low: Never; Medium: Rarely/Sometimes; High: Often/Always). Top-right: susceptibility to elite cues measured by agreement with randomly labeled policy proposals (Low: 0/3; Medium: 1/3 and 2/3; High: 3/3). Bottom-left: belief confidence assessed through willingness to wager on factual claims from an initial $100 endowment (Low: $0-$40; Medium: $41-$80; High: $81-$100). Bottom-right: social conformity concerns regarding opinion changes (Low: Not at all; Medium: Slightly/Moderately; High: Very/Extremely). The vertical dashed line indicates no difference with respect to the average of pre-treatment extreme participants in the AI-Grammar control condition. Bars indicate 90 and 95% confidence intervals, calculated using one-sided tests with standard errors clustered at the individual participant level.

E Experiment 1 survey instruments

E.1 Baseline survey

Participants begin the study by completing a baseline survey. The order of question blocks in this stage is randomized, except for blocks 1, 2, and 3. Within each block, the order of questions and the order of multiple-choice options are also randomized to reduce ordering bias.

Table E1: Survey Questions and motivation

Question Block	**Element**	**Question**	**Explanation**	**Range**
1	Prolific ID	What is your prolific ID?	Needed to pay participant.	
2	position	[`Would you favor or oppose allowing people with mental illnesses to purchase guns?`] What is your position on the issue?	User's self-reported position on the issue.	{-5,5}
3	confidence	How confident do you feel about your position?		{-5,5}
4	original argument	[`Would you favor or oppose allowing people with mental illnesses to purchase guns?`] [text box]	User's original argument.	String
5	prior knowledge	Under current federal law, which of the following mental health records are required to be reported to the firearm background check system (NICS)? a) All mental health treatment records b) **Only involuntary mental health commitments and legal determinations of mental incompetency** c) Any history of anxiety or depression treatment d) Only mental health records from the past 5 years	Evaluate baseline level of information.	{0,1}
	prior knowledge	What is Form 4473? a) A tax document for ammunition purchases. b) A firearm registration renewal form. c) A federal firearms license application. d) **A firearm transaction record for purchases.**	Evaluate baseline level of information.	{0,1}
	prior knowledge	When someone tries to buy a gun from a licensed dealer, what type of check is required? a) No check is required. b) Only a criminal record check. c) **A background check that includes both criminal and certain mental health records.** d) Only a check of their current address.	Evaluate baseline level of information.	{0,1}

Table E1: (Continued)

Question Block	Element	Question	Explanation	Range
6	second-order beliefs	In a representative sample of US respondents, what proportion of users do you think support allowing people with mental illnesses to purchase guns?	Evaluate whether user thinks this is a polarized or well-established topic.	{0,100}
7	comfort	Before we continue with the survey, which of these best describes your current situation? a) I'm in a private space with time to think and respond. b) I'm in a shared space but can focus on the questions. c) I'm multitasking or have limited time right now. d) I'm in a busy or public environment.	Evaluate whether the participant is in a physical space conducive for thoughtful responses.	
8	cognitive reflection test	If it takes five machines seven minutes to make five cars, how long would it take 100 machines to make 100 cars? **Seven minutes**	Evaluate cognitive reflection (Frederick, 2005).	{0,1}
	cognitive reflection test	Zach is taller than Matt and Richard is shorter than Zach. Which of the following statements would be most accurate? a) Richard is taller than Matt. b) Richard is shorter than Matt. c) Richard is as tall as Matt. d) **It's impossible to tell.**	Evaluate cognitive reflection (Erceg et al., 2020).	{0,1}
9	emotional state survey	How do you feel? (based on Larsen and Diener's circumplex model of affect (1992))	Evaluate the emotional state of the participant.	x = {0,10}; y = {0,10}

E.2 Endline survey, Part 1

After participants receive their assigned treatment and had the option to revise their initial argument, they proceed to the first endline survey. As before, the order of question blocks is randomized except for blocks 1 and 2, and question order within blocks, as well as multiple-choice options, is also randomized.

Table E2: Survey questions and motivation

Question Block	**Element**	**Question**	**Explanation**	**Range**
1	position	What is your position on the issue?	User's self-reported position on the issue.	{-5,5}
2	confidence	How confident do you feel about your position?		{-5,5}
3	cognitive reflection test 1	A man is 30 years old and his sister is half his age. How old will the man be when his sister is 35 years old?	Proxy for system-2 thinking.	{0,1}
	cognitive reflection test 2	If the day after tomorrow is two days before Thursday, then what day is today?	Proxy for system-2 thinking.	{0,1}
4	emotional state survey	How do you feel? How do you feel? (based on Larsen and Diener's circumplex model of affect (1992))	Proxy for emotional state.	x = {0,10}; y = {0,10}
5	sign-up	Would you like to sign up to receive more information about the topic? If you do, we will subscribe you to the digital newsletter of the National Alliance on Mental Illness (NAMI). a) Yes b) No	Interest/willingness to engage more.	{0,1}
6	perception conversation	How would you grade the conversation you just had? (1 for "very poor" to 5 "excellent").	Perception of conversation quality.	{1,5}
	perception respect	Do you feel that your comments and policy positions were respected in this process? (1 for "definitely yes" to 5 for "definitely no").	Mutual respect.	{1,5}

Table E2: (Continued)

Question Block	Element	Question	Explanation	Range
7	mutual understanding	Do you understand, recognize, or are open to the perspectives of others' comments, even if they are not like your own? a) I respect the opinions of people who disagree. b) Even if I don't agree with them, I understand people have good reasons for disagreeing with me. c) It is important to understand people who disagree with me by imagining how things look from their perspective. d) I find it difficult to see things from the point of view of people who disagree with me.	Empathy and openness to others' perspective (adapted from Argyle et al., 2023).	{a-d}
8	future participation	Would you be willing to participate in a deliberation process with the same group of participants again? (1 for "definitely yes" and 5 for "definitely no").	Perceptions of legitimacy.	{1,5}
9	agency	From 1 to 5, did you feel pressured in some way or another when expressing your views? (1 for "definitely yes" and 5 for "definitely no")	Perception of agency.	{1,5}

E.3 Endline survey, Part 2

The study concludes with the second part of the endline survey. Similar as before, the order of question blocks is randomized. The order of questions within blocks and the choices in multiple-choice questions are also randomized.

Table E3: Survey questions and motivation

Question Block	Element	Question	Explanation	Range
1	outcome acceptance	Would you like to endorse the top comments? a) Yes b) No	Acceptance of outcome	{0,1}

Table E3: (Continued)

Question Block	Element	Question	Explanation	Range
	willingness to take collective action	From 1 to 5, would you be willing to sign the following petition based on the comments you read today? Motivation: Keeping firearms away from those with severely impaired judgment due to mental illness can reduce both self-harm and public harm. Proposal: Implement legislation restricting firearm access for individuals diagnosed with serious mental health conditions that compromise decision-making, based on standardized psychiatric evaluations. Rationale: 1. Judgment: Reduced ability to determine lawful firearm use. 2. Rationality: Greater risk of self-harm or violence toward others. 3. Clear Thinking: Stable mental states are essential for responsible gun handling. Call to Action: Endorse or oppose this proposal, guiding policymakers to balance public safety with individual rights.	Willingness to compromise	{1-5}
2	costly compromise	Do you want us to donate half of your potential bonus payment of {$1/$10} to the National Alliance on Mental Illness organization (NAMI)? NAMI provides advocacy, education, support and public awareness so that all individuals and families affected by mental illness can build better lives. a) Yes b) No	Willingness to compromise through a donation	{0,1}
3	agency with respect to AI	Did you feel pressured by the AI chatbot to change your opinions? (1 for "definitely yes" and 5 for "definitely no")	Perception of agency	{1-5}
	persuasion	Did you feel persuaded by the AI chatbot? (1 for "definitely yes" and 5 for "definitely no")	Persuasion	{1-5}
4	zipcode	What is your zipcode?	Proxy for income	Numeric

Table E3: (Continued)

Question Block	Element	Question	Explanation	Range
	income	What was your total household income before taxes during the past 12 months? Multiple choice. Options: Less than $25,000 $25,000-$49,999 $50,000-$74,999 $75,000-$99,999 $100,000-$149,999 $150,000 or more	Proxy for income	Pre-determined buckets
5	profession	Select your current profession from the list below. Multiple choice. Options: Healthcare / Medical (e.g., doctor, nurse, pharmacist) Education (e.g., teacher, professor, school administrator) Finance / Banking / Accounting Real Estate / Insurance Information Technology (IT) Construction / Skilled Trades (e.g., electrician, plumber) Manufacturing / Production Agriculture / Farming / Forestry Retail / Sales Hospitality / Food Service (e.g., hotel, restaurant) Transportation / Logistics (e.g., truck driver, pilot) Government / Public Administration Nonprofit / Social Services Arts / Entertainment / Media Legal Services (e.g., attorney, paralegal) Marketing / Advertising / Public Relations Engineering / Research & Development Customer Service / Call Center Military / Defense / Armed Forces Freelance / Self-Employed / Consulting Homemaker / Stay-at-Home Parent Religious / Clergy Student Retired Unemployed / Looking for Work Other	Proxy for expertise	Pre-determined list of professions

Table E3: (Continued)

Question Block	**Element**	**Question**	**Explanation**	**Range**
6	prior exposure to AI	Please respond to the following questions regarding your experience with AI products or services. Two sliders: 1. (First Slider) How many of the following technologies have you used? 2. (Second Slider) On a scale from 0 to 5, how would you rate your experience with the technology? 1. OpenAI (GPT models, Sora) 2. Co-pilot 3. Llama 4. Microsoft Azure 5. Google Gemini 6. Claude 7. NotebookLM 8. Grok 9. Midjourney	Proxy for experience with AI	Predetermined list of AI services

F Experiment 2 survey instrument

F.1 Main survey

Participants begin the study by completing the baseline survey. The survey covers two of three policy topics, in random order: handgun regulation ("`Would you favor or oppose banning the possession of handguns, except by the police and other authorized persons?`"), voter identification requirements ("`Would you favor or oppose requiring all voters to show government-issued photo identification to vote?`"), and abortion regulations based on gestational age ("`Would you favor or oppose considering how long a woman has been pregnant in determining whether it is legal or illegal to have an abortion?`"). For each topic, participants provide their initial position, engage with an AI conversational agent, optionally revise their position, and vote on comments from other participants. Within each block, the order of questions and the order of multiple-choice options are also randomized to reduce ordering bias.

Table F1: Survey Questions and motivation

Question Block	**Element**	**Question**	**Explanation**	**Range**
`topic 1`				
1	initial position	[`Topic 1 position question`]	User's self-reported position.	{1,10}
	initial confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Confidence in position.	{a-e}
2	initial comment	[`Topic 1 discussion question`]	User's initial argument.	String (min 5 chars)
3	AI chatbot	[See Section A.2]	Treatment	1-5 rounds
4	revised comment	Would you like to update your previous comment? Your previous comment is below. You may update it at this time.	User's revised argument after AI interaction.	String (min 5 chars)

Table F1: (Continued)

Question Block	Element	Question	Explanation	Range
5	final position	[Topic 1 position question]	User's final position requirements.	{1,10}
	final confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Final confidence in position.	{a-e}
6	comment voting	See what others are saying Below are comments from others about this issue. Please vote on these comments. [9 seeded comments distributed in 2 pages in random order]	Evaluate other participants' comments.	Approve/ Disapprove/ Pass (min 5 comments)
7	transition	Thank you for your insight and comments. We will now move on to a new topic.	Transition between topics.	
topic 2				
8	initial position	[Topic 2 position question]	User's self-reported position on gestational limits for abortion.	{1,10}
	initial confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Confidence in abortion position.	{a-e}
9	initial comment	[Topic 2 discussion question]	User's initial argument on abortion.	String (min 5 chars)

Table F1: (Continued)

Question Block	Element	Question	Explanation	Range
10	AI chatbot	[See Section A.2]	Socratic dialogue with AI to encourage reflection on abortion position.	1-5 rounds
11	revised comment	Would you like to update your previous comment? Your previous comment is below. You may update it at this time.	User's revised argument on abortion after AI interaction.	String (min 5 chars)
12	final position	[`Topic 2 position question`]	User's final position on gestational limits for abortion.	{1,10}
	final confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Final confidence in abortion position.	{a-e}
13	comment voting	See what others are saying [`9 seeded comments distributed in 2 pages in random order`]	Evaluate other participants' comments on abortion.	Approve/ Disapprove/ Pass (min 5 comments)
14	transition	Thank you for your insight and comments! We have a few more questions to conclude the exercise.	Transition to closing questions.	

Table F1: (Continued)

Question Block	Element	Question	Explanation	Range
15	donation	Would you be willing to donate to Braver Angels? We will give you an extra $1. You may keep it or you may donate it to Braver Angels. Braver Angels is a leading cross-partisan, volunteer-led movement in the United States, with the goal of bridging the partisan divide and strengthening democracy through education, media, events, conversations, and so on. Every event held by Braver Angels includes liberal and conservative representatives. You can find more information about Braver Angels here: https://braverangels.org/ a) I would like to keep my extra $1. b) I would like to donate my extra $1 to Braver Angels.	Willingness to engage in cross-partisan dialogue.	{0,1}
16	conversation quality	How would you grade the conversation you just had? a) Very poor b) Poor c) Acceptable d) Good e) Very good	Perception of conversation quality.	{a-e}
	respect	Do you feel that your comments and policy positions were respected in this process? a) Definitely no b) Probably no c) Possibly d) Possibly yes e) Definitely yes	Perception of respect.	{a-e}
	mutual understanding	Do you understand, recognize, or are open to the perspectives of others' comments, even if they are not like your own? a) I respect the opinions of people who disagree. b) Even if I don't agree with them, I understand people have good reasons for disagreeing with me. c) It is important to understand people who disagree with me by imagining how things look from their perspective. d) I find it difficult to see things from the point of view of people who disagree with me.	Empathy and openness to others' perspective.	{a-d}

Table F1: (Continued)

Question Block	Element	Question	Explanation	Range
	human agency	From 1 to 5, did you feel pressured in some way or another when expressing your views?	Perception of agency.	{1,5}
	AI agency	From 1 to 5, did you feel pressured by the AI chatbot to change your opinions?	Perception of AI influence.	{1,5}
	AI bias	Do you think the AI chatbot was biased? If so, in what direction? a) Liberal b) Conservative c) I did not think the AI chatbot was biased. d) Unsure	Perceived bias of AI chatbot.	{a-d}
	team bias	Do you think the research team was biased? If so, in what direction? a) Conservative b) Liberal c) I do not think the research team was biased. d) Unsure	Perceived bias of research team.	{a-d}
	zipcode	What is your zip code?	Geographic location.	Numeric (5 digits)
17	Prolific ID	What is your Prolific ID?	Needed to pay participant.	String (24 chars)

F.2 Follow-up survey

Participants in the follow-up survey complete questions on two policy topics, starting with topic 1 of the main survey followed by a new topic. As previously, participants provide their initial position, engage with an AI conversational agent for socratic dialogue or grammar and spellcheck assistance (first topic only), optionally revise their position, and vote on comments from other participants. The survey concludes with questions measuring perceptions, information-seeking behavior, cross-partisan interaction, and elite cue effects. The order of questions and multiple-choice options is randomized within the page. In question block 18, participants are randomly assigned to see elite cues on options (a) and (b) or on options (c) and (d) for each policy proposal.

Table F2: Survey Questions and motivation

Question Block	Element	Question	Explanation	Range
`Topic 1`				
1	initial position	[`Topic 1 position question`]	User's self-reported position.	{1,10}
	initial confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Confidence in position.	{a-e}
2	initial comment	[`Topic 1 discussion question`]	User's initial argument.	String (min 5 chars)
3	AI chatbot	[See Section A.2]	1-5 rounds	
4	revised comment	Would you like to update your previous comment? Your previous comment is below. You may update it at this time.	User's revised argument after AI interaction.	String (min 5 chars)
5	comment voting	See what others are saying Below are comments from others about this issue. Please vote on these comments. [`9 seeded comments distributed in 2 pages in random order`]	Evaluate other participants' comments.	Approve/ Disapprove/ Pass (min 5 comments)
6	final position	[`Topic 1 position question`]	User's final position.	{1,10}
	final confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Final confidence.	{a-e}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
7 `topic 3`	transition	Thank you for your insight and comments. We will now move on to a new topic.	Transition between topics.	
8	initial position	[`Topic 3 position question`]	User's self-reported position on handgun regulation.	{1,10}
	initial confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Confidence in handgun position.	{a-e}
9	initial comment	[`Topic 3 discussion question`]	User's initial argument on handguns.	String (min 5 chars)
10	comment voting	See what others are saying Below are comments from others about this issue. Please vote on these comments. [`9 seeded comments distributed in 2 pages in random order`]	Evaluate other participants' comments on handguns.	Approve/ Disapprove/ Pass (min 5 comments)
11	final position	[`Topic 3 position question`]	User's final position on handgun regulation.	{1,10}
	revised comment	Would you like to update your previous comment? Your previous comment is below. You may update it at this time.	User's revised argument on handguns.	String (min 5 chars)

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
	final confidence	How confident are you in your position? a) Very unconfident b) Unconfident c) Neither confident nor unconfident d) Confident e) Very confident	Final confidence in handgun position.	{a-e}
12	transition	Thank you for your insight and comments! We have a few more questions to conclude the exercise.	Transition to closing questions.	
13	conversation quality	How would you grade the conversation you just had? a) Very poor b) Poor c) Acceptable d) Good e) Very good	Perception of conversation quality.	{a-e}
	AI agency	From 1 to 5, did you feel pressured by the AI chatbot to change your opinions?	Perception of AI influence.	{1,5}
	AI bias	Do you think the AI chatbot was biased? If so, in what direction? a) Liberal b) Conservative c) I did not think the AI chatbot was biased. d) Unsure	Perceived bias of AI chatbot.	{a-d}
	human agency	From 1 to 5, did you feel pressured in some way or another when expressing your views?	Perception of agency.	{1,5}
	team bias	Do you think the research team was biased? If so, in what direction? a) Conservative b) Liberal c) I do not think the research team was biased. d) Unsure	Perceived bias of research team.	{a-d}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
15	respect	Do you feel that your comments and policy positions were respected in this process? a) Definitely no b) Probably no c) Possibly d) Possibly yes e) Definitely yes	Perception of respect.	{a-e}
	mutual understanding	Do you understand, recognize, or are open to the perspectives of others' comments, even if they are not like your own? a) I respect the opinions of people who disagree. b) Even if I don't agree with them, I understand people have good reasons for disagreeing with me. c) It is important to understand people who disagree with me by imagining how things look from their perspective. d) I find it difficult to see things from the point of view of people who disagree with me.	Empathy and openness to others' perspective.	{a-d}
16	donation	Would you be willing to donate to Braver Angels? We will give you an extra $1. You may keep it or you may donate it to Braver Angels. Braver Angels is a leading cross-partisan, volunteer-led movement in the United States, with the goal of bridging the partisan divide and strengthening democracy through education, media, events, conversations, and so on. Every event held by Braver Angels includes liberal and conservative representatives. You can find more information about Braver Angels here: https://braverangels.org/ a) I would like to keep my extra $1. b) I would like to donate my extra $1 to Braver Angels.	Willingness to engage in cross-partisan dialogue.	{0,1}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
17	search behavior	Have you searched for information (online or offline) about any of the following topics in the past 4 weeks? Select all that apply. • Regulation of handgun possession and use • How gestational age factors into abortion laws • Requirements to use a government-issued photo ID to vote • Local public library services and hours • Household recycling and trash collection guidelines • Public transportation routes and schedules in your area • Immigration policy (e.g., border enforcement and pathways to legal status) • Climate and energy policy (e.g., emissions regulations, renewable energy) • Police funding and public safety policies • None of the above • Prefer not to say	Prior information-seeking behavior on policy topics.	Multiple select
17	cross-partisan interaction	How often do you interact with people who identify with a different political party than you (including independents or nonpartisans)? a) Never b) Rarely c) Sometimes d) Often e) Always	Frequency of cross-partisan engagement.	{a-e}
	error probabilities	Imagine we randomly pick one factual claim you believe about public issues and verify it. You have $100 to wager at even odds (if the claim is correct, you win the amount you stake; if it's incorrect, you lose it). How much of the $100 would you bet that your belief is correct? a) 0 b) 1-10 c) 11-20 d) 21-40 e) 41-60 f) 61-80 g) 81-100	Confidence in factual beliefs about public issues.	{a-g}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
18	reputation costs	If you changed your view after seeing new evidence, how concerned would you be about reactions from people important to you (friends, family, colleagues)? a) Not at all concerned b) Slightly concerned c) Moderately concerned d) Very concerned e) Extremely concerned	Social costs of belief revision.	{a-e}
	elite cue (handguns, version A)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Eliminating permit requirements for law-abiding citizens to carry handguns in public b) Requiring universal background checks for all gun sales, including at gun shows and online purchases c) Allowing qualified school personnel to carry firearms on campus after proper training (supported by Governor Greg Abbott and the NRA) d) Implementing a mandatory 72-hour waiting period for all firearm purchases (supported by Senator Chris Murphy and Everytown for Gun Safety)	Elite cues randomized (50% probability): options (a) and (b) without cues; options (c) and (d) with cues.	{a-d}
	elite cue (handguns, version B)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Eliminating permit requirements for law-abiding citizens to carry handguns in public (supported by Governor Greg Abbott and the NRA) b) Requiring universal background checks for all gun sales, including at gun shows and online purchases (supported by Senator Chris Murphy and Everytown for Gun Safety) c) Allowing qualified school personnel to carry firearms on campus after proper training d) Implementing a mandatory 72-hour waiting period for all firearm purchases	Elite cues randomized (50% probability): options (a) and (b) with cues; options (c) and (d) without cues. Shown if version A is hidden.	{a-d}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
	elite cue (abortion, version A)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Protecting health care providers from being compelled to participate in abortions through a Conscience Protection Act with a private right of action b) Ensuring over-the-counter birth control pills are covered by insurance without cost-sharing c) Permitting pharmacists to decline dispensing medications that conflict with their religious convictions (supported by Senator Marco Rubio and the Family Research Council) d) Protecting access to fertility treatments and contraception as fundamental healthcare rights (supported by Senator Elizabeth Warren and Planned Parenthood)	Elite cues randomized (50% probability): options (a) and (b) without cues; options (c) and (d) with cues.	{a-d}
	elite cue (abortion, version B)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Protecting health care providers from being compelled to participate in abortions through a Conscience Protection Act with a private right of action (supported by Senator Marco Rubio and the Family Research Council) b) Ensuring over-the-counter birth control pills are covered by insurance without cost-sharing (supported by Senator Elizabeth Warren and Planned Parenthood) c) Permitting pharmacists to decline dispensing medications that conflict with their religious convictions d) Protecting access to fertility treatments and contraception as fundamental healthcare rights	Elite cues randomized (50% probability): options (a) and (b) with cues; options (c) and (d) without cues. Shown if version A is hidden.	{a-d}

Table F2: (Continued)

Question Block	Element	Question	Explanation	Range
	elite cue (voter ID, version A)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Requiring driver's license or state ID numbers on absentee ballot applications and envelopes for verification b) Expanding early voting periods to include additional weekends and evenings c) Mandating signature verification and witness requirements for absentee ballots (supported by Governor Brian Kemp and the Heritage Foundation) d) Implementing automatic voter registration when citizens interact with government agencies (supported by Stacey Abrams and the Democratic National Committee)	Elite cues randomized (50% probability): options (a) and (b) without cues; options (c) and (d) with cues.	{a-d}
	elite cue (voter ID, version B)	Below are various policy proposals currently being discussed. Please select the policy you most prefer from the list below. a) Requiring driver's license or state ID numbers on absentee ballot applications and envelopes for verification (supported by Governor Brian Kemp and the Heritage Foundation) b) Expanding early voting periods to include additional weekends and evenings (supported by Stacey Abrams and the Democratic National Committee) c) Mandating signature verification and witness requirements for absentee ballots d) Implementing automatic voter registration when citizens interact with government agencies	Elite cues randomized (50% probability): options (a) and (b) with cues; options (c) and (d) without cues. Shown if version A is hidden.	{a-d}
19	partisanship	How would you describe your political preferences? a) Strong Republican b) Moderate Republican c) Moderate Democrat d) Strong Democrat e) Independent	Political party identification.	{a-e}
	Prolific ID	What is your Prolific ID?	Needed to pay participant.	String (24 chars)

G Rubrics for argument scoring

This section details the criteria used to rate the quality of participants' arguments along five dimensions: *Thoughtfulness (0-6)*, *clarity (0-3)*, *coherence (0-3)*, *depth (0-3)*, and *perspective diversity (0-3)* (see Section 5.3). Each dimension has a discrete scoring scale with corresponding definitions and illustrative examples. Human and LLM annotators will independently apply these rubrics to each participant's written response.

Table F1: Rubrics for argument scoring

Dimension	Score	Definition / Description
Thoughtfulness score (0–6)		
Thoughtfulness	0	Statement is largely incomprehensible or not self-contained.
Thoughtfulness	1	Blunt position, but unrelated or only tangentially relevant to the prompt.
Thoughtfulness	2	States a stance with no meaningful argument or justification.
Thoughtfulness	3	Argument is present but remains incomplete or inconclusive.
Thoughtfulness	4	A clear, reasoned argument is presented.
Thoughtfulness	5	Argument elaborates potential nuances or conditions.
Thoughtfulness	6	Argument is thoroughly developed, with scope and limits articulated.
Clarity score (0–3)		
Clarity	0	Statement is unclear or incomprehensible.
Clarity	1	Main idea is identifiable but requires rewriting for clarity.
Clarity	2	Mostly clear, unambiguous expression.
Clarity	3	Completely clear and self-contained.
Coherence score (0–3)		
Coherence	0	Entirely incoherent; no logical flow.
Coherence	1	Main point is present but argument lacks consistent progression.
Coherence	2	Argument is coherent and relatively straightforward to follow.
Coherence	3	All parts of the argument fit together logically.
Depth score (0–3)		
Depth	0	Extremely shallow or cursory.
Depth	1	Briefly touches on the issue with minimal detail.

Continued on next page...

Table F1 – continued from previous page

Dimension	Score	Definition / Description
Depth	2	Engages the core of the topic but with limited elaboration.
Depth	3	Demonstrates thorough and insightful understanding of the issue.
Perspective diversity score (0–3)		
Perspective Diversity	0	No recognition of alternative viewpoints.
Perspective Diversity	1	Acknowledges alternative perspectives without further discussion.
Perspective Diversity	2	Identifies and discusses alternative viewpoints but only partially.
Perspective Diversity	3	Comprehensively contrasts and integrates diverse viewpoints.

H Ethical considerations

H.1 Informed consent

All participants provided informed consent before participating in the study. They were informed about the nature of the tasks, the use of their data, and their right to withdraw at any time.

H.2 Confidentiality

Participant data are anonymized and stored securely. Personally identifiable information was not collected and only aggregated results are reported.

H.3 Risk sssessment

The study posed minimal risk to participants. Potential discomfort from engaging with the LLM or reflecting on personal beliefs was expected to be low.

H.4 Institutional Review Board (IRB) approval

The study protocol falls under Stanford IRB protocol #75616 and MIT IRB protocol #5798.

I Additional notes

OpenAI's ChatGPT and Claude's Sonnet were used to copy edit this document.