Artificial Intelligence Circumvents Identity-Driven Biases in Source Selection

Laura Globig[1], Rachel Xu[2], Hamza Alshamy[3] & Jay J. Van Bavel[1,4]

[1] Department of Psychology & Center for Neural Science, New York University, NY, NY, US

[2] Jigsaw (Google LLC), Mountain View, CA, USA

[3] Center for Data Science, New York University, NY, NY, US

[4] Norwegian School of Economics in Bergen, Norway

Correspondence authors: lg3962@nyu.edu & jay.vanbavel@nyu.edu

14 October 2025

Competing Interests: R.X. is an employee at Google, Inc. This project was funded with support from Google, Inc. awarded to L.K.G., and J.J.V.B.

Funding Sources: This project was funded with support from Google, Inc. awarded to L.K.G., and J.J.V.B. In addition, J.J.V.B. were supported by a grant from the National Science Foundation (#2334148) and the Templeton World Charity Foundation (JVB, doi.org/10.54224/31570).

Abstract:
Social identity profoundly shapes whom people choose as information sources, constraining exposure to diverse perspectives. While people are motivated to seek accurate information, they systematically avoid outgroup sources even when group membership is irrelevant to the task at hand. Here we investigate whether artificial intelligence (AI) can circumvent these identity-driven biases in source selection. In Study 1, a nationally representative sample of American adults (n = 1,054) preferred AI over human sources when seeking information about political conflicts. In Study 2 (n = 284), an incentivized political fact-checking experiment revealed that participants preferred AI sources over outgroup (d = 0.470) and even ingroup (d = 0.230) partisan sources, despite recognizing they were of equal competence. In Study 3 (n = 277), using an identity-irrelevant shape categorization task, participants only preferred AI over outgroup sources (d = 0.191), with no difference between AI and ingroup sources. Computational modeling revealed that these preferences emerge through selectively accumulated evidence against partisan advisors during deliberation, rather than differences in priors. These findings suggest that AI's perceived neutrality enables it to bypass identity-based discrimination. These results highlight the potential of AI to reduce echo chambers and broaden epistemic exposure by serving as an identity-neutral conduit for information acquisition.

Keywords: Artificial Intelligence, Information-Seeking, Source Selection, Social Identity, Partisan Bias

Introduction

In today's information-rich world (Castells, 1996), access to knowledge has never been easier, yet people's actual information environments remain surprisingly narrow (Del Vicario et al., 2016; Sunstein, 2018). A central, often overlooked, determinant of what people ultimately learn is who they seek information from. As errors are costly and humans are motivated to maximize their expected utility (Mongin, 1998; Von Neumann & Morgenstern, 2007), this decision should - in principle - be governed by the (perceived) accuracy of potential information sources (Bromberg-Martin & Sharot, 2020; Madsen et al., 2024; Sharot & Sunstein, 2020). In practice, however, source selection is often also shaped by factors beyond accuracy, such as social identity (Van Bavel & Pereira, 2018). People actively discount outgroup sources (Marks et al., 2019; Zhang & Rand, 2023) and prefer those who are similar to them (Mullen et al., 1992; Zou & Xu, 2023), prestigious figures (Van Noord et al., 2023), or members of majority groups (Boorman et al., 2013; Marks et al., 2019; Otten, 2016) as information sources, even when group identity is irrelevant to the information sought (Del Vicario et al., 2016; Sunstein, 2018). Such outgroup derogation constrains information acquisition, narrows exposure, and contributes to polarization (Cikara & Van Bavel, 2014; Hobson & Inzlicht, 2016; Molenberghs, 2022), demanding urgent solutions. We examine whether Artificial Intelligence (AI) can overcome these partisan biases by offering a neutral source of information.

Even when people can accurately assess which sources are competent, they often fail to act on this knowledge. For instance, individuals sometimes share misinformation on social media despite being able to distinguish true from false content when asked directly (Pennycook et al., 2021). Similarly, people may continue to seek advice from familiar or identity-congruent sources even when these sources are demonstrably inaccurate (Kim & Kim, 2021; Metzger et al., 2020). These dissociations suggest that the processes guiding accuracy detection and source selection are partially independent: people can recognize which sources are more competent yet still choose based on identity concerns, social rewards, or motivated reasoning. As a result, identity-driven biases in source selection persist even in contexts where accuracy is incentivized and evaluative judgments are intact.

Whereas information used to be exclusively provided by other humans - be it in person or indirectly through different media - this is now being supplemented by AI. AI tools have seen explosive adoption. For example, OpenAI's ChatGPT reached 100 million users within two months of its launch, making it the fastest-growing consumer application in history (Blunt, 2025). Early research indicates that people may be especially receptive to information from AI (Costello, 2025; Stapleton et al., 2022; Xu et al., 2025). As yet, however, it remains unclear how people's information-seeking preferences differ when consulting human versus AI sources, and whether AI can reduce the identity-driven biases that constrain human-to-human advice seeking. To fill this gap, we now investigate how AI interacts with existing source selection preferences, and examine the processes underlying source selection decisions.

People actively seek information to reduce uncertainty (Hofmann et al., 2009; Schrah et al., 2006; Yaniv & Kleinberger, 2000). In doing so, they strive to improve their mental models of the world around them (Sharot & Sunstein, 2020). For instance, they might read political commentary to refine their understanding of societal values, ask a psychologist for an explanation for their own emotional responses, or seek social feedback to clarify how they are perceived by others. Beyond the specific type of information sought (Kelly & Sharot, 2021), the initial choice of who to consult is a crucial first step that profoundly shapes what people ultimately learn. Prior work illustrates that this decision is shaped by both accuracy goals (Bromberg-Martin & Sharot, 2020) and social identity goals, such as belonging, status, and moral validation (Van Bavel & Pereira, 2018). For instance, people tend to seek information from

sources that align with their partisan identity (Kim & Kim, 2021; Knobloch-Westerwick, 2012; Metzger et al., 2020; Stroud, 2010; Winter et al., 2016). Democrats often prefer CNN while Republicans favor Fox News (Hawkins & Nosek, 2012; Rosentiel, 2009).

Partisan preferences shape not only media consumption (Cikara & Van Bavel, 2014; Molenberghs, 2013, 2022; Morrison et al., 2012) but also beliefs about political facts (Kahan, 2015; Rathje et al., 2023) and policy preferences (Geerlings et al., 2017; Tribukait, 2021). By reinforcing only congenial perspectives, these habits create curation bubbles (Green et al., 2025) and even echo chambers (Del Vicario et al., 2016; Sunstein, 2018) with profound epistemic consequences. Over time, selective sampling may narrow viewpoints, reinforce prejudice and polarization (Druckman & Levy, 2022), undermine trust in democratic institutions (Pasek et al., 2022), and foster voter apathy (Ahn & Mutz, 2023; Crepaz, 1990; Fivaz & Nadig, 2010; Moral, 2017; Phillips, 2024; Snyder III, 2011).

Identity-driven source selection is not confined to identity-relevant contexts. Even in simple perceptual judgments, people prefer politically similar advisors over dissimilar ones (Marks et al., 2019; Zhang & Rand, 2023). In educational contexts, students often resist seeking information from perceived outgroup sources, constraining academic development and engagement with diverse ideas (Dion et al., 1972; Nisbett & Wilson, 1977; Schuchart et al., 2021; Thorndike, 1920). Identity thus operates at a fundamental cognitive level, alongside accuracy. Identity value biases processing across the brain (Pereira et al., 2023), creating a systematic tilt in whom people choose to hear from by making ingroup information feels more rewarding, and outgroup information aversive (Hackel et al., 2017; Van Bavel et al., 2008).

When encountering outgroup sources, people exhibit negative affective (Iyengar et al., 2012; Iyengar & Westwood, 2015), cognitive (Cikara et al., 2014; Mullen et al., 1992; Tajfel et al., 1971) and neural responses (Cikara & Van Bavel, 2014; Falk et al., 2012; Hobson & Inzlicht, 2016; Molenberghs, 2013, 2022; Molenberghs & Louis, 2018; Morrison et al., 2012). This aversion reduces their willingness to sample from outgroup sources (Marks et al., 2019; Zhang & Rand, 2023). These findings reveal a pervasive epistemic constraint: social identity systematically biases the selection of information sources, narrowing the evidence people encounter before any belief updating can even occur.

Importantly, intergroup biases may operate automatically, making them especially resistant to correction (Devine, 1989; Greenwald & Banaji, 1995; Yudkin et al., 2016). Even explicit efforts to debias, such as incentivizing accuracy or urging impartiality, frequently fail to eliminate identity-driven avoidance, indicating that these group biases can persist as reflexive, ingrained responses (Zhang & Rand, 2023). The social reward and threat dynamics underlying group identity biases are deeply ingrained at a neural level, meaning that conventional interventions have only limited impact on these automatic tendencies (Krajbich, 2022; Lai et al., 2014).

AI has begun to supplement, and in some domains replace, human sources, opening up new possibilities for source selection. Recent work suggests that large language models (LLMs) can act as new kinds of intermediaries for information-seeking. Unlike partisan human sources, LLMs distill vast corpora of human knowledge into accessible explanations, potentially reducing epistemic fragmentation and offering individuals shared evidence that informs expert judgment (Costello, 2025). To understand the implications of this new class of sources for human decision-making, we ask: Can we utilize AI to reduce identity-driven avoidance at the very first step: source selection?

We speculate that unlike humans, AI systems are not (yet) inherently embedded in social group hierarchies and may thus be perceived as more neutral during source selection (Messeri & Crockett, 2024). Advances in large language models (LLMs) enable complex reasoning and human-like dialogue (Antikatzidis et al., 2024; Argyle et al., 2023; Bail, 2024; Costello et al., 2024). Recent studies suggest that AI usage can result in improvements to information environments: motivating news consumption (Askari et al., 2024), fostering more constructive political discussion (Argyle et al., 2023), and reducing conspiracy beliefs (Costello et al., 2024). Together, these findings suggest that AI could preserve sensitivity to accuracy while blunting identity-based avoidance in source selection.

We therefore test whether AI can mitigate information selection bias in both politically sensitive and politically neutral contexts. We hypothesize that (1) people will prefer AI over human sources when given the choice; and (2) that this preference will be particularly pronounced relative to humans that belong to a political outgroup. Computationally, these identity-driven biases could operate at different stages of the decision process. People might enter choices with an initial predisposition toward one type of source (a starting point bias), or they might selectively accumulate evidence in favor of some sources over others during deliberation (a drift rate bias). These dual pathways can yield the same observable behavior while reflecting distinct underlying mechanisms (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013). Distinguishing between them is theoretically important because it reveals whether identity shapes source selection through reflexive priors, deliberative evidence processing, or both.

Overview:
Across 3 studies, we investigated source selection preferences for information-seeking. In Study 1, we examined whether people prefer AI over human sources when choosing who should explain a political conflict to them in a representative sample of 1,054 Americans across multiple generations (18-65 year olds). We then ran two pre-registered behavioral experiments with politically-balanced American samples to assess whether people also preferred AI as an information source over partisan sources. In Study 2 (n = 284), we compared their information-seeking preferences for AI relative to ingroup and outgroup sources, using a highly identity-relevant political fact-checking task. In Study 3 (n = 277), we then assess whether the results from Study 2 generalize to an identity-irrelevant shape-categorization task (adapted from Marks et al., 2019). We used drift–diffusion modeling to decompose the cognitive processes underlying participants' choices. Pre-registrations, code, and data are all available on the Open Science Framework (OSF).

Results

Participants prefer learning about a political conflict from AI vs humans (Study 1).

To examine whether participants prefer AI rather than human sources for politically sensitive information, we first surveyed a nationally representative sample of 1,054 adults in the United States. The sample was quota-matched by age, gender, and ethnicity to reflect the broader population. Participants completed an online survey in which they indicated whether they would prefer a human or an AI system to explain a political debate to them (**see Figure 1a**). The majority of participants selected AI (55.6%) over a human (44.4%) as their preferred information source (**see Figure 1b**). An exact binomial test against a null of equal preference confirmed that this effect was statistically significant ($p < 0.001$, *95% CI* [0.530, 0.590]), suggesting that AI is viewed as a more neutral or competent information source for navigating politically charged, identity-relevant contexts.

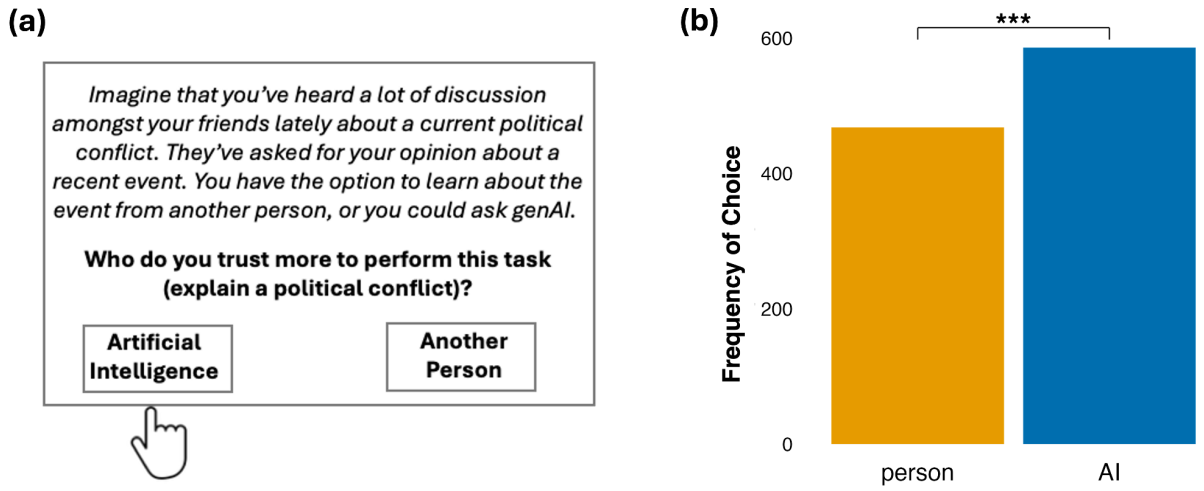

Figure 1. Participants preferred AI over a human to explain a political debate. (a) Participants were asked to indicate whether they trusted AI or another person more to explain a political conflict. **(b)** A significantly greater number of participants sought information about a political conflict from AI (blue) compared to a person (orange). Bars reflect raw choice frequencies. *** $p < 0.001$.

Social identity does not impair source competence perceptions during highly identity relevant tasks (Study 2).

Thus far our results suggest that participants prefer learning about political conflicts from AI over humans. Political conflicts are often partisan issues, and thus tend to be highly identity-relevant. It is thus plausible that whether or not participants prefer AI over humans, depends on the latter's partisan identity. To test this, we ran a second study, in which a new group of participants completed a political fact-checking task (**see Figure 2**). In this task, participants were asked to choose between advisors (=information sources) of varying identity (ingroup partisan, outgroup partisan, or AI) and demonstrated accuracy (accurate vs. random). Prior to the task, they completed an initial learning phase, in which they observed each source's responses, received feedback on whether the source's response was accurate or not, and rated each source's competence. Later, in the decision-making phase, participants made their own judgments, reported their confidence, and then chose which source to consult before making a final decision. They were incentivized for accuracy and could receive a performance-related bonus payment of up to $2. This setup allowed us to test whether participants could distinguish competent from incompetent sources, and whether identity biases still guided their choices even when accuracy had tangible consequences.

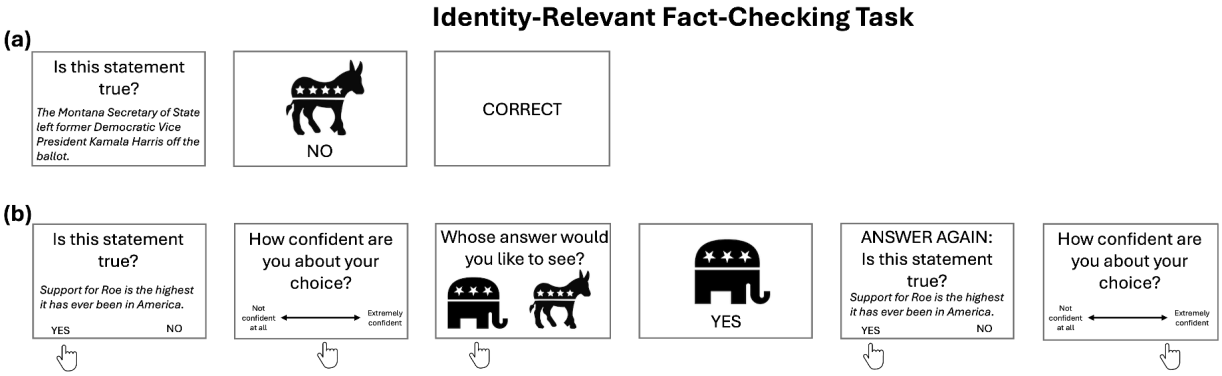

Figure 2. Example Trials of Identity-Relevant Political Fact-Checking Task. In Study 2 participants completed a political fact-checking task in which they had to determine whether statements were true or not. This constituted a learning phase and a test phase. **(a)** In the learning phase, participants learned about each source's task performance. The learning phase consisted of 6 blocks (one for each source), with 10 trials each. In each trial, a new political statement was presented, for which participants would observe the source's response and receive feedback about whether that response was correct or not. The order of the blocks was randomized. **(b)** After the learning phase, participants completed the test phase. This consisted of four blocks with 30 trials each. In each trial, a new political statement was presented, for which participants then had to indicate whether the statement was correct or not, and indicate how confident they were in their response. They were then given the choice to ask one of two randomly assigned sources for advice and see their chosen source's response. Afterwards they were able to update their response and indicate their revised confidence rating. Sources were counterbalanced across trials and blocks.

In a first step, we examined whether participants accurately evaluated source performance. To that end we entered task performance for each source into a 3 (source identity: ingroup, outgroup, AI) × 2 (source accuracy: accurate, random) repeated-measures ANOVA. This revealed that participants accurately estimated source competence. Participants rated accurate sources (M = 77.600, SEM = 0.776) significantly more competent than random-performing sources (M = 54.400, SEM = 0.9300; (F(1, 282) = 528.800, p < 0.001, ηp^2 = 0.650). Importantly, their perceptions of source competence did not vary as a function of source identity (F(1.990, 561.080) = 2.240, p = 0.108, ηp^2 = 0.008). They did not assign higher or lower competence ratings to ingroup, outgroup, or AI sources based on their identity alone (ingroup: M = 65.600, SEM = 0.929; outgroup: M = 65.200, SEM = 0.856; AI: M = 67.200, SEM = 0.903). We also did not observe a significant interaction effect between accuracy and identity (F(1.970, 554.520) = 1.220, p = 0.295, ηp^2 = 0.004). Thus, social identity did not impede participants' perceptions of source competence. The results hold even when controlling for political orientation and political sectarianism (**see Supplementary Tables 1-2**).

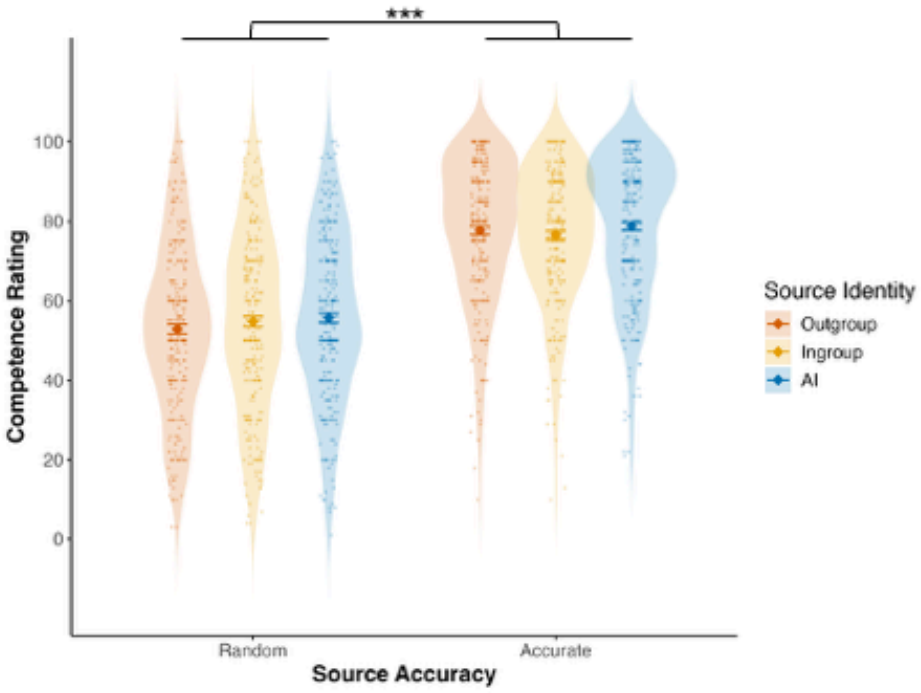

Figure 3. Social Identity did not impair participants' perceptions of source competence. Participants rated accurate sources as more competent than random sources, regardless of source identity. Y axis shows perceived competence ratings on a 0–100 scale. X axis shows the accuracy level of the source during the learning phase (random = 50%, accurate = 80%). Each violin plot shows the distribution of competence ratings for outgroup (dark-orange), ingroup (light-orange), and AI (blue) sources. Individual participant responses are plotted as dots. Diamond shapes represent the mean rating for each group, and vertical error bars indicate the standard error of the mean (SEM). Violin width reflects the density of responses.***$p < 0.001$.

Participants prefer seeking information from AI over partisan sources (Study 2).
Since participants accurately assessed source competence, they could in principle maximize their likelihood of receiving accurate advice - and thus increase their bonus payment - by basing their choices solely on competence, irrespective of source identity. However, prior work shows that people often fail to act on competence alone, instead privileging identity-congruent sources even when these are no more accurate than alternatives (Marks et al., 2019). To test this, we assessed how frequently participants chose each source to give them advice during the test phase of the task. Across trials, each source was presented as an option an equal number of times, allowing us to compute the proportion of times each was selected.

We hypothesized that participants would prefer accurate over random sources, ingroup over outgroup sources, and AI over partisan sources. Indeed, a 3 (source identity: ingroup, outgroup, AI) × 2 (accuracy: accurate, random) repeated-measures ANOVA on proportion of times a source was selected revealed a significant main effect of accuracy ($F(1, 282) = 161.484$, $p < 0.001$, $\eta p^2 = 0.360$), with accurate sources ($M = 0.596$, $SEM = 0.008$) chosen more often than random ones ($M = 0.404$, $SEM = 0.008$; **see Figure 4**). We also observed a significant main effect of source identity ($F(1.713, 483.02) = 39.301$, $p < 0.001$, $\eta p^2 = 0.120$), such that participants preferred seeking information from AI ($M = 0.575$, $SEM = 0.012$) over outgroup sources ($M = 0.422$, $SEM = 0.009$; $t(282) = 7.913$, $p < 0.001$, $d = 0.470$), as well as over ingroup

sources (*M* = 0.503, *SEM* = 0.009; t(282) = 3.875, *p* < 0.001, *d* = 0.230). Moreover, participants preferred ingroup sources over outgroup sources (*t*(282) = 6.097, *p* < 0.001, *d* = 0.362). There was no significant interaction between source identity and accuracy (*F*(1.906, 537.560) = 0.583, *p* =0.550, ηp^2 = 0.002).

Together, these findings reveal that participants balance both accuracy and identity considerations during source selection. As expected, they favored ingroup over outgroup sources, replicating prior work (Marks et al., 2019). However, participants consistently preferred AI over both ingroup and outgroup sources, even when controlling for political orientation and political sectarianism (**see Supplementary Tables 3–4**). We found no systematic differences in belief updating across source types (**see Supplementary Table 5-7**), indicating that AI's overcome intergroup bias at the stage of source selection. Because information-seeking is the first step in the cascade that can produce curation bubbles and echo chambers, these results suggest that AI may mitigate identity-driven avoidance.

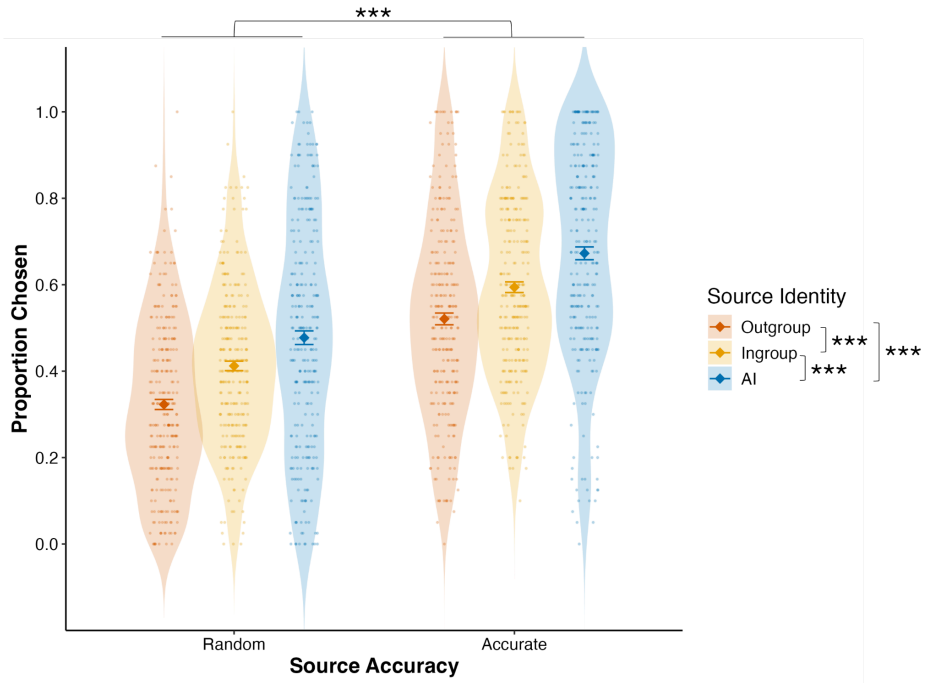

Figure 4. Participants preferred seeking advice from accurate and from politically aligned sources. A main effect of accuracy emerged such that accurate sources were chosen more often than random sources. Participants also preferred AI over both ingroup and outgroup sources, as well as ingroup over outgroup sources. Y axis shows the proportion of times each source was selected during the decision-making stage. X axis shows source accuracy (random = 50%, accurate = 80%). Each violin plot shows the distribution of selection proportions for ingroup (light orange), outgroup (dark orange), and AI (blue) source. Individual participant responses are plotted as dots. Diamond shapes represent the mean proportion chosen for each group, and vertical error bars indicate the standard error of the mean (SEM). Violin width reflects the density of responses. Significance brackets indicate pairwise comparisons. ****p < 0.001.*

Identity biases operate through selective derogation of partisan sources during deliberation (Study 2).

Importantly, the same behavioral outcome, preferring AI sources over partisan sources, may arise from different underlying processes. First, participants may have a prior (= starting point

bias) towards AI, entering the decision predisposed to choose AI before seeing what the other option even is. Alternatively, they could also have a process bias (= drift rate bias) towards AI, selectively discounting partisan sources during the source deliberation process, once both options are presented. This again will lead to preferential source selection of AI sources. To tease apart these possibilities and better characterize how identity and accuracy shape the decision process, responses were modeled with a drift–diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013) with the following parameters: (1) *t0*—amount of non-accumulation time; (2) *a*—distance between decision thresholds; (3) *z*—starting point of the accumulation process; and (4) *v*–drift rate. Source identity was coded with two orthogonal contrasts (ingroup vs. AI; outgroup vs. AI), and accuracy difference between both source options was entered as a trial-wise predictor; these variables modulated the parameters depending on the model specification.

We compared 16 models in which parameters were either fixed or allowed to vary by source identity and accuracy (**see Supplementary Table 8**). The simplest model assumed all parameters were fixed, while successive models allowed variation in either z or v. We then estimated models in which two parameters varied simultaneously. Finally, the full model allowed both parameters to vary as a function of accuracy and source identity. This model space enabled us to ask whether identity concerns manifest as a starting-point bias, a process bias in drift rate, or a combination of both. The Watanabe-Aikaike Information Criterion (WAIC) was calculated for each model. The best fitting yet simplest model was one in which the drift rate (v) varied as a function of source identity and accuracy. While more complex models including starting-point bias (z) parameters showed numerically similar fit, the simpler model with only drift rate variation was preferred based on parsimony (**see Supplementary Table 8**).

We observed that accuracy differences between sources influenced the drift rate, with faster evidence accumulation rates towards more accurate sources (β = 0.299, *95% HDI* [0.288, 0.311]). Participants also exhibited systematic biases in drift rate parameters as a function of source identity. Participants demonstrated large negative drift rates for both ingroup (β = -0.118, *95% HDI* [-0.132,-0.104]) and outgroup sources (β = -0.241, 95% HDI [-0.255,-0.227]) relative to AI. This indicates that during deliberation, participants systematically accumulated evidence against human partisan sources and toward AI sources, with the bias being particularly pronounced against outgroup sources. See **Table 1** for complete results.

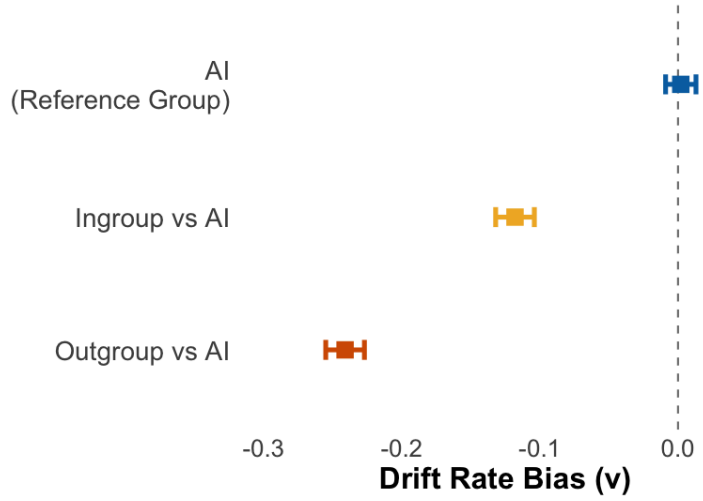

Figure 5. Source selection preferences emerge through systematic devaluation during deliberation. Drift-diffusion model parameter estimates reveal systematic biases in how participants processed information from different sources. Participants accumulated evidence more slowly from both human sources (ingroup = yellow, outgroup = orange) compared to AI (=blue). AI serves as the reference category (β = 0) for both parameters. Error bars represent 95% HDIs.

Table 1. Parameter estimates of evidence accumulation process (Study 2).

Estimates	**Mean** [95% HDI]
Decision threshold (a)	1.244 [1.237,1.251]
Non-decision time (t0)	0.188 [0.182,0.194]
Starting point (z)	0.526 [0.522,0.530]
v Intercept	0.002 [-0.009,0.013]
v Ingroup vs AI	-0.188 [-0.132,-0.104]
v Outgroup vs AI	-0.241 [-0.255,-0.227]
v Accuracy Difference	0.299 [0.288,0.311]

Context-dependent Source Selection Biases: Only Outgroup-Derogation in Source Selection persists when the task is identity-irrelevant (Study 3).

Thus far, our results reveal that people prefer learning from AI over human sources for identity-relevant topics. However, prior work suggests that identity-biases also drive information-seeking preferences in identity-irrelevant contexts (Marks et al., 2019). To determine if preferential source selection of AI extends beyond identity-relevant tasks, such as political conflicts (Study 1) or the political fact-checking task (Study 2), we next sought to determine if this preference holds for tasks that are completely unrelated to social identity. To that end, we ran a third study with a new group of participants. The task was identical to Study 2 (**see Figure 6**), but this time instead of a political fact-checking task, we used a shape-categorization task (adapted from Marks et al., 2019), in which participants had to determine whether a shape was a "blip" or not. In reality, the category was randomly assigned. This task has previously been shown to elicit partisan bias in source selection.

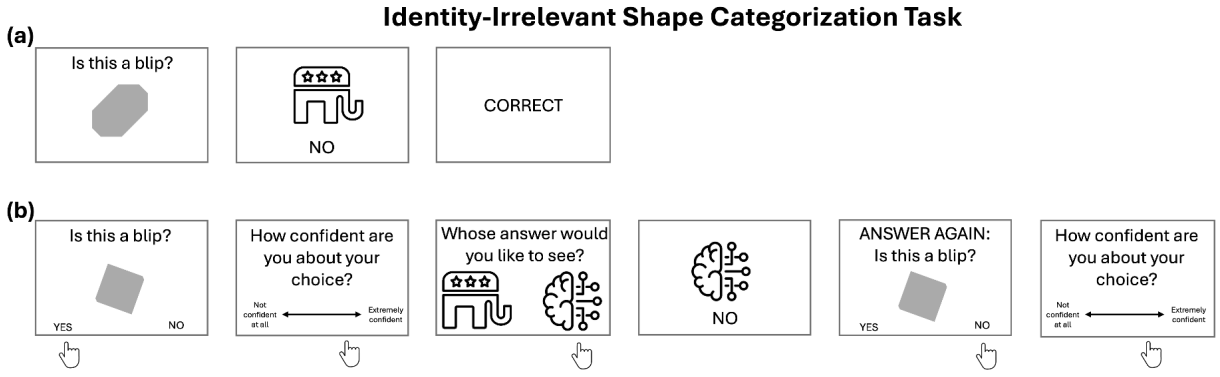

Figure 6. Example Trials of Identity-Irrelevant Shape Categorization Task. In Study 3 participants completed a shape categorization task, in which they had to determine whether the shape was a "blip" or not. This constituted a learning phase and a test phase. **(a)** In the learning phase, participants learned about each source's task performance. The learning phase consisted of 6 blocks (one for each source), with 10 trials each. In each trial, a new shape was presented, for which participants would observe the source's response and receive feedback about whether that response was correct or not. The order of the blocks was randomized. **(b)** After the learning phase, participants completed the test phase. This consisted of four blocks with 30 trials each. In each trial, a new shape was presented, for which participants then had to indicate whether the shape was a "blip" or not, and indicate how confident they were in their response. They were then given the choice to ask one of two randomly assigned sources for advice and see their chosen source's response. Afterwards they were able to update their response and indicate their revised confidence rating. Sources were counterbalanced across trials and blocks.

Replicating the results from Study 2, participants rated accurate sources (M = 76.700, SEM = 0.783) significantly more competent than random-performing sources (M = 57.900, SEM = 0.993; (F(1, 271) = 362.523, p < 0.001, ηp^2 = 0.570; **see Figure 7**), indicating that they successfully tracked performance. By contrast, there was no main effect of source identity (F(1.953, 529.604) = 0.087, p = 0.912, ηp^2 = 0.0003). That is, participants did not assign higher or lower competence ratings to ingroup, outgroup, or AI sources based on their identity alone (ingroup: M = 67.200, SEM = 0.924; outgroup: M = 67.300, SEM = 0.991; AI: M = 67.600, SEM = 0.951). We also did not observe a significant interaction effect between accuracy and identity (F(1.994, 539.637) = 1.296, p = 0.276, ηp^2 = 0.004). Thus participants' perceptions of information source competence were primarily driven by observed performance, not by identity. The results hold even when controlling for political orientation and political sectarianism (**see Supplementary Tables 9–10)**.

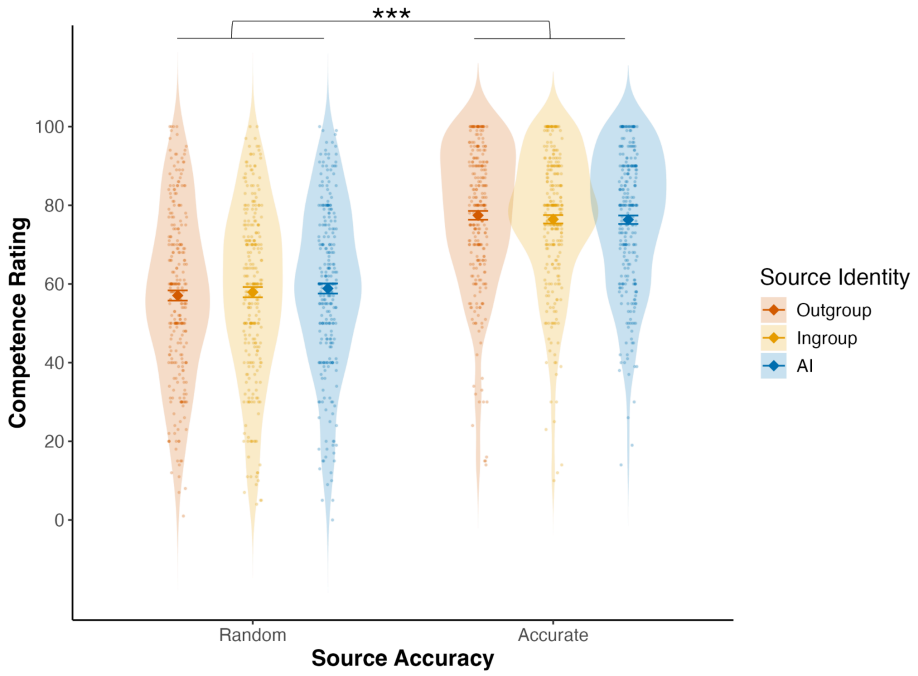

Figure 7. Social Identity did not impair participants' perceptions of source competence for identity-irrelevant tasks. Participants rated accurate sources as more competent than random sources, regardless of source identity. Y axis shows perceived competence ratings on a 0–100 scale. X axis shows the accuracy level of the source during the learning phase (random = 50%, accurate = 80%). Each violin plot shows the distribution of competence ratings for outgroup (dark-orange), ingroup (light-orange), and AI (blue) sources. Individual participant responses are plotted as dots. Diamond shapes represent the mean rating for each group, and vertical error bars indicate the standard error of the mean (SEM). Violin width reflects the density of responses.***$p < 0.001$.

Having established that participants could accurately assess source competence for both identity-relevant (Study 2) and identity-irrelevant tasks (Study 3), we next assessed whether their preferences for AI over human sources also extended to identity-irrelevant tasks. As for the identity-relevant task, a 3 (source identity: ingroup, outgroup, AI) × 2 (source accuracy: accurate, random) repeated-measures ANOVA on proportion of times a source was selected revealed a significant main effect of accuracy ($F(1, 271) = 101.091$, $p < 0.001$, $\eta p^2 = 0.104$), with accurate sources ($M = 0.577$, $SEM = 0.008$) chosen more often than random ones ($M = 0.423$, $SEM = 0.008$). We also once again observed a significant main effect of source identity ($F(1.813, 491.413) = 6.666$, $p = 0.002$, $\eta p^2 = 0.014$), such that participants preferred seeking information from AI ($M = 0.516$, $SEM = 0.012$) over outgroup sources ($M = 0.462$, $SEM = 0.010$; $t(271) = 2.728$, $p = 0.019$, $d = 0.191$) and ingroup sources ($M = 0.522$, $SEM = 0.009$) over outgroup sources, respectively ($t(271) = 3.998$, $p < 0.001$, $d = 0.173$). However, there was no difference in their information-seeking preferences between ingroup and AI sources ($t(271) = 0.295$, $p = 0.953$, $d = 0.018$), suggesting that perhaps with the political fact-checking task, participants were cautious of potential political biases in partisan sources. There was no significant interaction between source identity and accuracy ($F(1.953, 529.210) = 1.596$, $p = 0.204$, $\eta p^2 = 0.006$). The results hold even when controlling for political orientation and political sectarianism (**see Supplementary Tables 11-12**). We did not find significant differences in belief-updating tendencies (**see Supplementary Tables 13-15**). Taken together, these results suggest that AI may provide a novel pathway to broaden information exposure by reducing identity-driven barriers to source selection.

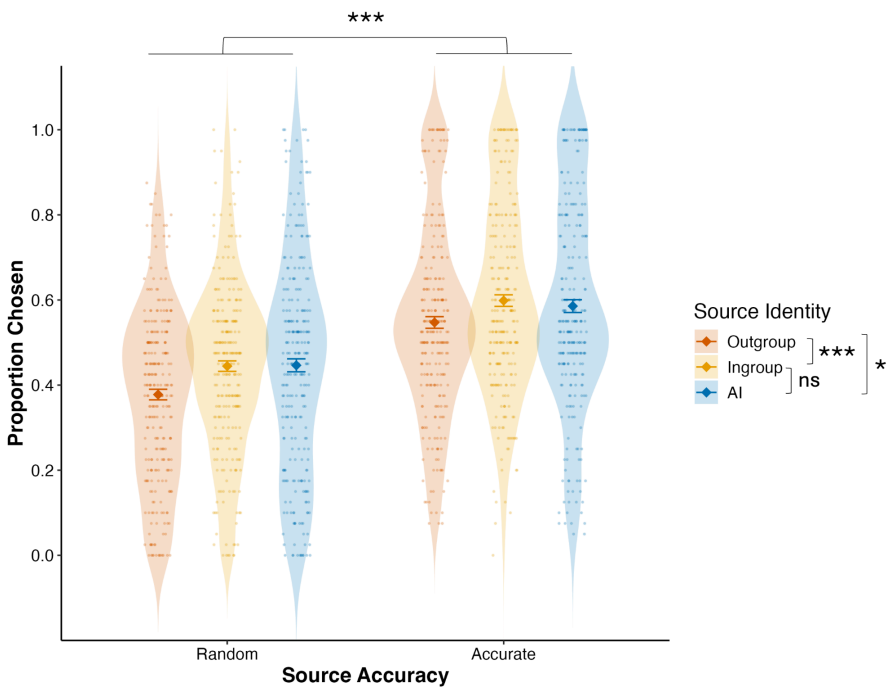

Figure 8. Participants preferred seeking advice from accurate and from politically aligned sources. A main effect of accuracy emerged such that accurate sources were chosen more often than random sources. Participants also preferred AI over outgroup sources, and ingroup over outgroup sources. No difference was observed between ingroup and AI sources. Y axis shows the proportion of times each source was selected during the decision-making stage. X axis shows source accuracy (random = 50%, accurate = 80%). Each violin plot shows the distribution of selection proportions for ingroup (light orange), outgroup (dark orange), and AI (blue) source. Individual participant responses are plotted as jittered dots. Diamond shapes represent the mean proportion chosen for each group, and vertical error bars indicate the standard error of the mean (SEM). Violin width reflects the density of responses. Significance brackets indicate pairwise comparisons. *t* = p < 0.10, ***p* < 0.01, ***p* < *0.001.*

Ingroup sources achieve computational parity with AI when source identity becomes task-irrelevant (Study 3).

We then once again thought to tease apart the processes underlying our behavioral findings using DDM. As for Study 2, we observed that a model which allowed the drift rate (v) to vary as a function of source identity and accuracy provided the best model fit while accounting for parsimony (**see Supplementary Table 16**).

Participants in Study 3 also showed negative drift rates only for outgroup sources (*β* = -0.077, *95% HDI* [-0.091, -0.063]). The ingroup penalty observed in political contexts completely disappeared in the identity-irrelevant task (*β* = 0.008, *95% HDI* [-0.006, 0.0222]), suggesting that ingroup sources achieved computational parity with AI when political identity became irrelevant. This is consistent with our behavioral results for both studies, where participants preferred AI sources over human sources for the political fact-checking task, but only discounted outgroup sources for the identity-irrelevant shape categorization task.

Accuracy effects remained robust with participants having faster evidence accumulation rates (*β* = 0.220, *95% HDI* [0.209, 0.232]) towards accurate sources relative to random sources. **See Table 2** for full parameter estimates.

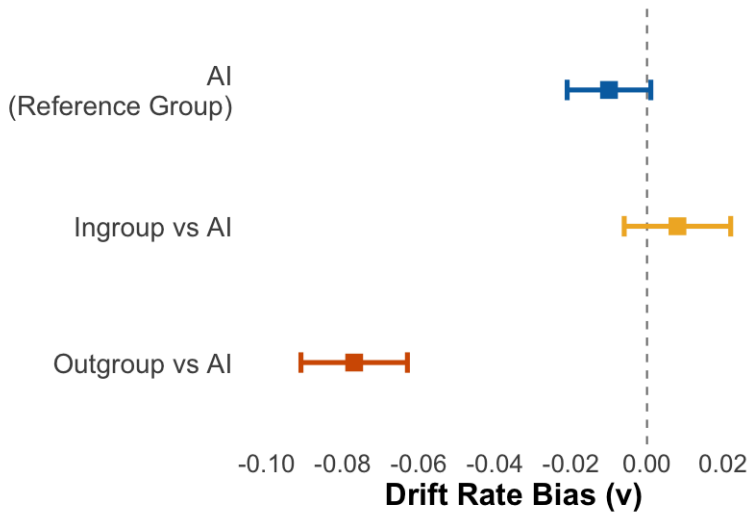

Figure 9. Participants accumulate evidence more slowly from outgroup sources compared to AI. There was no difference between ingroup and AI. AI is the reference category ($\beta = 0$) for both parameters. Error bars represent *95% HDIs*.

Table 2. Parameter estimates of evidence accumulation process (Study 3).

Estimates	Mean [*95% HDI*]
Decision threshold (a)	1.284 [1.278,1.290]
Non-decision time (t(0))	0.000 [0.000,0.001]
v Intercept	-0.010 [-0.021,0.001]
v Ingroup vs AI	0.008 [-0.006,0.022]
v Outgroup vs AI	-0.077 [-0.091, -0.063]
v Accuracy Difference	0.220 [0.209,0.232]

Discussion.

Across three studies, we find that AI can serve as a bypass for identity-driven source selection biases. When deciding whom to consult, people preferred ingroup over outgroup sources–but they consistently preferred AI over partisan human sources. This was the case despite participants accurately recognizing that ingroup, outgroup, and AI sources were equally competent. These findings suggest that AI is perceived as a comparatively neutral alternative when social identity threatens to constrain information seeking. They extend prior work showing that identity systematically biases source choice (Marks et al., 2019; Zhang & Rand, 2023) by identifying a technological pathway through which such biases can be bypassed.

Computational modeling points to differential deliberation underlying the effect. Once specific options were presented, they accumulated preference evidence in a way that penalized partisan sources - especially outgroup advisors - relative to AI (lower drift rates).

Notably, this effect was context-dependent. In politically charged, identity-relevant tasks (Study 2), people preferred AI over both ingroup and outgroup sources. In contrast, in an identity-irrelevant shape categorization task (Study 3), ingroup advisors achieved parity with AI, while only outgroup advisors continued to be disadvantaged. This asymmetry highlights different forms of bias. Outgroup discounting persisted across contexts, suggesting a more automatic and generalized form of discrimination (Iyengar & Westwood, 2015; Molenberghs, 2013). By contrast, ingroup discounting appeared specific to political contexts, possibly reflecting people's recognition that even their own partisan allies may introduce bias when the stakes are political. Importantly, these patterns held regardless of political orientation or levels of political sectarianism, indicating that the preference for AI over partisan advisors was not confined to a particular partisan subgroup. Moreover, actual belief-updating remained stable across source identities across both contexts, indicating that context shaped selection while integration stayed comparatively constant.

The observed preference for AI over partisan human sources did not override participants' fundamental concern with accuracy. Participants reliably distinguished competent from incompetent sources and chose accurate advisors more often, regardless of whether the source was AI or human. This indicates that the preference for ingroup members as well as AI reflects strategic information seeking rather than indiscriminate bias. In addition to accuracy goals, AI served as a trusted alternative when human sources were perceived as politically biased, preserving sensitivity to competence while circumventing identity-driven avoidance.

These findings carry practical implications for the design of information environments. Algorithmic mediation represents a viable strategy for combating political polarization in digital environments, particularly when AI systems are seen as neutral arbiters rather than partisan actors. By offering a non-identity-threatening alternative to partisan advisors, AI may reduce the formation of epistemically narrow "echo chambers" (Del Vicario et al., 2016; Sunstein, 2018). A broader body of research reinforces this cautious optimism. It has been argued that LLMs, unlike earlier intermediaries such as media outlets or search engines, have the potential to democratize the synthesis of knowledge and blunt the epistemic fragmentation that drives polarization (Costello, 2025). Moreover, AI can do more than transmit information: it can actively mediate deliberation, producing group statements that participants prefer to human-mediated outcomes and reducing division within citizens' assemblies (Tessler et al., 2024). These findings suggest that AI may circumvent identity-driven biases not only by providing information perceived as neutral but also by structuring collective reasoning in ways that foster consensus. Taken together, these insights situate our findings within a growing recognition that AI may reshape information environments not only by bypassing identity-based avoidance at the level of source selection but also by shaping how groups negotiate disagreement.

At the same time, our results caution that AI's advantage is contingent on the public's perceptions of its neutrality (Stoyanovich et al., 2020), which comes with its own risks. Young adults, for instance, often perceive information from AI chatbots as requiring less scrutiny than human sources, which may lower evaluative standards (Xu et al., 2025). Moreover, evidence of partisan bias in LLMs (Rozado, 2023) and the rise of explicitly branded systems (e.g., Grok) suggest that over time users may come to select into ideologically congenial AIs much as they do partisan media. If AI systems become strongly associated with particular institutions, political orientations, or demographic groups (Messeri & Crockett, 2024), they may cease to circumvent identity biases and instead reproduce them. AI's potential benefits therefore depend on systems being explicitly designed to prioritize veracity and transparency rather than engagement (Costello, 2025). Who programs these systems—and the values embedded in them—will

therefore critically shape whether AI broadens information exposure and reproduces polarization.

While this research focused specifically on political identity as a driver of source bias, the observed processes likely extend to other identity dimensions that shape information processing. Future research should investigate whether AI similarly circumvents biases based on race, gender, religious affiliation, or other social categories. Given that identity-protective cognition operates across multiple group memberships through the same ventromedial prefrontal and striatal circuitry (Van Bavel et al., 2008), AI may provide a generalizable solution for reducing various forms of intergroup bias in information consumption. Research should also examine the temporal stability of AI's bias-circumventing effects as public familiarity with AI systems increases and institutional associations become more salient.

Conclusion.
Together, these findings demonstrate that AI can circumvent the deeply ingrained identity biases that constrain human-to-human source selection. Across three studies, participants consistently preferred AI over outgroup advisors, even when they recognized all sources as equally competent. Computational modeling revealed that this preference reflects a deliberative devaluation of partisan input, particularly from outgroups. These results highlight both the promise and the boundary conditions of AI as an epistemic tool. On the one hand, AI's perceived neutrality enables it to broaden exposure to diverse information, reducing the formation of epistemically narrow echo chambers. On the other hand, this advantage depends on maintaining perceptions of impartiality and embedding systems within accountability structures that prioritize accuracy and transparency. By uncovering the cognitive mechanisms through which AI reshapes source selection, our work provides both theoretical insight into identity-driven biases and practical guidance for designing AI systems that promote healthier information environments.

Methods.
We report how we determined our sample size, all data exclusions, all manipulations and all measures in the experiment. The research methods were approved by the New York University Ethics Committee (IRB-FY2024-8271, IRB-FY2025-9303). Qualtrics survey files, anonymized data and analysis code is available on our Open Science Framework (OSF) page.

<u>Study 1.</u>
Participants.
One-thousand and fifty-four participants residing in the United States were recruited using Prolific Academic. The sample was quota-matched to be nationally representative of the U.S. population by age, gender, and ethnicity. No participants were excluded from analysis. The final sample included 532 women, 504 men, and 18 individuals who self-identified as another gender. Participants ranged in age from 18 to 85 years (M = 45.773, SD = 15.613). Based on self-reported political orientation, 601 identified as liberal, 286 as conservative, and 167 as politically neutral. For all experiments presented in this article, ethical approval was provided by the New York University Research Ethics Committee and all participants gave informed consent. All experiments were performed in accordance with the principles expressed in the Declaration of Helsinki. All samples were politically balanced for Democrats and Republicans.

Materials.
Participants completed a 20-minute online survey. Here, we report the measure relevant to the current study. Additional measures were included as part of related projects (Globig et al., 2024); full survey materials and code are available on OSF. To assess trust in AI for politically

sensitive content, participants were presented with a scenario in which either a person or AI would explain a political debate on their behalf. They were asked to indicate who they would prefer to perform this task: a human or an AI system. This scenario was designed to probe perceptions of competence and neutrality in identity-relevant information-seeking contexts.

Analysis.
To assess whether participants exhibited a systematic preference for AI over a human in the political debate scenario, we conducted an exact binomial test comparing the observed proportion of AI selections to a null hypothesis of equal choice probability ($p = 0.5$). This test was selected due to its suitability for binary categorical data and its robustness for analyzing deviations from chance-level preference.

Study 2.
Participants.
Sample size was computed based on prior work (Marks et al., 2019) Power calculations were performed using g*Power (Faul et al., 2009) to achieve power of 0.8 ($\beta = 0.2$, $\alpha = 0.05$). Two-hundred and seventy-four participants who resided in the USA, voted in the previous general election and identified as either Republican or Democrat completed the task on Prolific Academic. As pre-registered, we excluded one participant, who did not identify as Democrat or Republican. Thus, data of 283 participants were analyzed (149 Democrats, 134 Republicans, Mage = 43.465, SDage ± 12.963, range = 21–80; female = 137, male = 142, other = 3). Participants received $12 per hour for their participation, in addition to a performance-related bonus.

Materials.
Fact-Checking Task.
Participants completed an incentivized political fact-check task designed to examine how individuals sought information from politically aligned (ingroup), misaligned (outgroup) human advisors (=sources), and artificial intelligence (AI) advisors. The task was implemented using JavaScript and JsPsych (v7.3.4), hosted via Firebase. It was self-paced and took approximately 30 minutes to complete. Participants first completed a set of measures assessing political orientation, political sectarianism, trust in AI, and usage frequency of AI tools. They were then introduced to six sources, defined by a 3 (source identity: ingroup, outgroup, AI) × 2 (accuracy: accurate =0.8, random=0.5) within-subject design. Each source was visually represented by a distinct avatar and clearly labeled during a brief categorization task to ensure participants could recognize their identity and role (see Figure 3).

Learning Stage
The learning stage was designed to help participants form judgments about each source's competence in a binary shape categorization task ("Is this a blip?"). It consisted of six blocks of 10 trials, with each block associated with a different source. The order of blocks and trials was randomized for each participant.

On each trial, participants viewed a novel abstract shape, followed by the source's response ("yes" or "no") and visual feedback indicating whether the source's response was correct. Source accuracy was experimentally manipulated: three sources responded correctly on 80% of trials (accurate), while the other three responded correctly on 50% of trials (chance level). After each block, participants rated the perceived competence of that source using a slider ranging from 0 (completely inaccurate) to 100 (completely accurate). Once all six sources had been observed, participants sorted them from best to worst.

Decision-Making Stage
Confidence ratings were collected both before and after the source input. Participants were informed that only their final response would count toward a bonus payment of up to $2, incentivizing strategic use of source advice.

During the decision-making stage, participants completed four blocks of 30 shape categorization trials each. Each trial began with a novel shape and an initial “yes” or “no” judgment and a confidence rating from 0 (not confident at all) to 100 (extremely confident). Participants were then presented with two randomly selected source avatars and asked to choose one for advice. After viewing the selected source’s response, they had the opportunity to revise their judgment and confidence rating. Source pairs varied in identity and accuracy, allowing for within-subject comparisons across all combinations of source identity and accuracy. Participants were informed that only their final decision would count toward bonus earnings, which incentivized selective and strategic use of source input.

Comprehension and Engagement Checks
Comprehension checks were presented before both the learning and decision-making stages. Participants who failed were required to reread the instructions. Additional attention checks and source-identification trials ensured participants understood the roles and accuracy levels of each source.

Post-task Measures
After the main task, participants rated their own performance on a slider and completed standard demographic questions. They also completed Inclusion of Other in the Self (IOS) diagrams to indicate perceived closeness to each source and responded to open-ended questions about how they evaluated sources and made decisions. These were designed to assess if participants were aware of the study hypotheses.

Analysis.
Perceived Competence.
We examined whether participants differentially perceived the competence of ingroup, outgroup, and AI advisors depending on their accuracy. Each advisor (Democrat, Republican, AI) was recoded into ingroup, outgroup, or AI based on the participant’s self-reported political orientation. For each participant, competence ratings were averaged separately for each source identity (ingroup, outgroup, AI) and for each level of accuracy (accurate, random). These values were entered into a 3 (source identity: ingroup, outgroup, AI) × 2 (source accuracy: accurate, random) within-subject ANOVA. Post hoc pairwise comparisons were conducted to unpack significant effects. We additionally performed one-sample t-tests to assess whether each group’s perceived competence exceeded the midpoint of the scale. All statistical tests conducted in the present article are two sided. Analysis was conducted using IBM SPSS 27 and R Studio (Version 1.3.1056). All results of interest hold when controlling for political orientation and political sectarianism (**see Supplementary Materials**).

Advisor Selection Analysis.
We next examined whether participants differentially preferred ingroup, outgroup, or AI advisors. For each participant, we computed the proportion of times each source was selected (when available) as a function of agent accuracy. These proportions were entered into a 3 (source identity: ingroup, outgroup, AI) × 2 (source accuracy: accurate, random) within-subject ANOVA. Post hoc pairwise comparisons were conducted to unpack significant main and interaction effects. To further characterize the preference data, we conducted a series of planned paired t-tests comparing proportions across sources, correcting for multiple comparisons using a

Bonferroni adjustment. All results of interest hold when controlling for political orientation and political sectarianism (**see Supplementary Materials**).

DDM.
Our aim in modeling our task using the drift-diffusion framework was to assess how source identity and accuracy impacted the evidence accumulation process during source selection. In particular, we wanted to assess (1) whether identity concerns manifest as a starting-point bias, a process bias in drift rate, or a combination of both; and (2) whether accuracy considerations operate through similar or distinct computational processes.

We implemented and compared 16 different specifications of a DDM (**see Supplementary Materials**). The models included the following parameters: (1) t_0, amount of non-accumulation/non-decision time; (2) a, distance between decision thresholds; (3) z, starting point of the accumulation process; and (4) v, drift rate, the rate of evidence accumulation. Source identity was coded with two orthogonal contrasts (ingroup vs. AI; outgroup vs. AI), with AI serving as the reference category, and accuracy difference between both source options was entered as a trial-wise predictor. These variables modulated the parameters depending on the model specification.

The model space systematically varied parameter specifications. The simplest model (Model 1) assumed all parameters were fixed, while successive models allowed variation in either z or v as a function of accuracy alone (Models 2-4) or source identity alone (Models 5-7). We then estimated models in which both accuracy and identity influenced single parameters (Models 8-12), followed by models where both parameters varied simultaneously (Models 13-16). This comprehensive model space enabled us to isolate the specific computational processes through which identity and accuracy biases emerge.

We used the HSSM software toolbox (Fengler et al., *in prep*) to estimate the parameters of our drift–diffusion models. The HSSM package employs Bayesian parameter estimation, using the No-U-Turn Sampler (NUTS) to draw samples from the posterior distributions of the parameters. In fitting the models, we used HSSM's default weakly informative priors for all parameters. Models were fit jointly to choices and reaction times (RTs). We drew 10,000 posterior samples per chain after tuning and used 5,000 tuning (burn-in) iterations, running 4 independent chains in total. To assess convergence, we inspected the Gelman–Rubin R^R^ statistic from ArviZ summaries. In each case, R^ was close to one (<1.1), suggesting adequate convergence. Model fits were compared using the Watanabe–Akaike Information Criterion (WAIC).We assessed whether differences in WAIC between models were meaningful by comparing the difference in expected log predictive density (elpd) to its standard error, with differences exceeding 4 standard errors considered statistically significant. When models showed equivalent fit (differences < 4 SE), we selected the most parsimonious model with fewer effective parameters.

<u>Study 3.</u>
Participants.
Sample size was computed based on prior work (Marks et al., 2019). Power calculations were performed using g*Power (Faul et al., 2009) to achieve power of 0.8 ($\beta = 0.2$, $\alpha = 0.05$). Two-hundred and seventy-seven participants who resided in the USA, voted in the previous general election and identified as either Republican or Democrat completed the task on Prolific Academic. As pre-registered, we excluded five participants, who did not identify as Democrat or Republican. Thus, data of 272 participants were analyzed (133 Democrats, 139 Republicans, Mage = 39.246, SDage ± 12.881, range = 18–75; female = 140, male = 130, other = 2).

Participants received $12 per hour for their participation, in addition to a performance-related bonus.

Materials.
Shape Categorization Task. Participants completed an incentivized shape categorization task. The task was identical to the task in Study 2, except that this time participants had to decide whether a shape was a “blip” or not.

Analysis.
Analysis was identical to that in Study 2.

References

Blunt, D. M. and K. (2025, September 4). *Google Ruling Shows Antitrust Tools Struggle to Keep Up With Tech Markets*. The Wall Street Journal. https://www.wsj.com/us-news/law/google-search-antitrust-enforcement-limits-alphabet-cb1e5b6e

Bromberg-Martin, E. S., & Sharot, T. (2020). The value of beliefs. *Neuron*, *106*(4), 561–565.

Castells, M. (1996). The information age: Economy, society and culture (3 volumes). *Blackwell, Oxford*, *1997*, 1998.

Costello, T. (2025). Large language models as disrupters of misinformation. *Nature Medicine*, *31*(7), 2092. https://doi.org/10.1038/s41591-025-03821-5

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160.

Globig, L. K., Xu, R., Rathje, S., & Van Bavel, J. J. (2024). Perceived (Mis) alignment in Generative Artificial Intelligence Varies Across Cultures. *Preprint. DOI*, *10*. https://osf.io/suqa2/download

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Kelly, C. A., & Sharot, T. (2021). Individual differences in information-seeking. *Nature Communications*, *12*(1), 7062.

Kim, J. W., & Kim, E. (2021). Temporal Selective Exposure: How Partisans Choose When to Follow Politics. *Political Behavior*, *43*(4), 1663–1683. https://doi.org/10.1007/s11109-021-09690-1

Knobloch-Westerwick, S. (2012). Selective Exposure and Reinforcement of Attitudes and Partisanship Before a Presidential Election. *Journal of Communication*, *62*(4), 628–642. https://doi.org/10.1111/j.1460-2466.2012.01651.x

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2019). Epistemic spillovers: Learning others' political views reduces the ability to assess and use their expertise in nonpolitical domains. *Cognition*, *188*(April 2018), 74–84. https://doi.org/10.1016/j.cognition.2018.10.003

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2020). Cognitive Dissonance or Credibility? A Comparison of Two Theoretical Explanations for Selective Exposure to Partisan News. *Communication Research*, *47*(1), 3–28. https://doi.org/10.1177/0093650215613136

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience & Biobehavioral Reviews*, *37*(8), 1530–1536.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, Roger., & McKoon, Gail. (2008). Drift Diffusion Decision Model:Theory and data. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, *12*(3), 148.

Rubin, D. B., & Gelman, A. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472.

Sharot, T., & Sunstein, C. R. (2020). How people decide what they want to know. *Nature Human Behaviour*, *4*(1), 14–19. https://doi.org/10.1038/s41562-019-0793-1

Stapleton, A., Lavelle, J., & McHugh, L. (2022). *Chatbot-delivered acceptance and commitment therapy with adolescents: A pilot randomized controlled trial*. OSF. https://doi.org/10.31234/osf.io/j2kpt

Stoyanovich, J., Van Bavel, J. J., & West, T. V. (2020). The imperative of interpretable machines. *Nature Machine Intelligence*, *2*(4), 197–199. https://doi.org/10.1038/s42256-020-0171-8

Stroud, N. J. (2010). Polarization and Partisan Selective Exposure. *Journal of Communication*, *60*(3), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Sunstein, C. (2018). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. https://doi.org/10.1515/9781400890521

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The Neural Substrates of In-Group Bias: A Functional Magnetic Resonance Imaging Investigation. *Psychological Science*, *19*(11), 1131–1139. https://doi.org/10.1111/j.1467-9280.2008.02214.x

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

Winter, S., Metzger, M. J., & Flanagin, A. J. (2016). Selective Use of News Cues: A Multiple-Motive Perspective on Information Selection in Social Media Environments. *Journal of Communication*, *66*(4), 669–693. https://doi.org/10.1111/jcom.12241

Xu, R., Le, N., Park, R., Murray, L., Das, V., Kumar, D., & Goldberg, B. (2025). "Information Modes": A Framework for Trust and Information Seeking. *Journal of Online Trust and Safety*, *3*(1). https://doi.org/10.54501/jots.v3i1.245

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*(11), 1448–1459. https://doi.org/10.1037/xge0000190

Zhang, Y., & Rand, D. G. (2023). Sincere or motivated? Partisan bias in advice-taking. *Judgment and Decision Making*, *18*, e29. https://doi.org/10.1017/jdm.2023.28