Biased AI Improves Human Decision-Making But Reduces Trust

Anonymous Author(s)
Affiliation
Address
email

Abstract

Current AI systems minimize risk by enforcing ideological neutrality, yet this may introduce automation bias by suppressing cognitive engagement in human decision-making. We conducted randomized trials with 2,500 participants to test whether culturally biased AI enhances human decision-making. Participants interacted with politically diverse GPT-4o variants on information evaluation tasks. Partisan AI assistants enhanced human performance, increased engagement, and reduced evaluative bias compared to non-biased counterparts, with amplified benefits when participants encountered opposing views. These gains carried a trust penalty: participants underappreciated biased AI and overcredited neutral systems. Exposing participants to two AIs whose biases flanked human perspectives closed the perception–performance gap. These findings complicate conventional wisdom about AI neutrality, suggesting that strategic integration of diverse cultural biases may foster improved and resilient human decision-making.

1 Introduction

Generative AI systems are increasingly embedded in human decision-making, prompting industry efforts to develop "fair" AI by removing culturally or ideologically biased outputs through techniques like fine-tuning and RLHF (Lin et al., 2024; Feng et al., 2023; Zou et al., 2023). Yet, despite these interventions, biases persist (Feng et al., 2023; Bai et al., 2025; Potter et al., 2024a), raising doubts about whether true neutrality is possible (Martin, 2023; Lee et al., 2024; Anthis et al., 2024; Fisher et al., 2025; Potter et al., 2024b). Critics also point out that focusing solely on model-level fairness neglects the interactive nature of human-AI coordination (Peeters et al., 2021; Tsvetkova et al., 2024; Shen et al., 2024), where sanitized, seemingly neutral systems risk promoting automation bias (Parasuraman & Riley, 1997; Mosier et al., 1996), diminishing critical engagement (Parasuraman & Riley, 1997; Bastani et al., 2024), and leading to moral deskilling (Fan et al., 2025; Unk), accountability issues (Porsdam Mann et al., 2023; Wachter et al., 2024), and a homogenization of thought (Campo-Ruiz, 2025; Agarwal et al., 2024; Meincke et al., 2025).

We argue that carefully calibrated, culturally biased AI can enhance human-AI overall performance, fostering productive provocation, disagreement, and critical evaluation, rather than passive consensus. Existing literature from social sciences shows how deliberately introducing strategic biases may improve decision-making by reactivating human critical thinking. Kunda's motivated reasoning framework argues that activating accuracy motivations or directional motivations tends to increase cognitive effort (Kunda, 1990), suggesting that purposely biased AI may heighten humans' engagement by motivating them to challenge competing views from AI (Tetlock & Boettger, 1989; Ditto & Lopez, 1992). Mercier and Sperber's argumentative theory of reasoning suggests that overtly partisan AIs may be experienced as interlocutors that invite rebuttal and critical scrutiny, preventing the overly compliant, "sycophantic" drift of AI assistants (Mercier & Sperber, 2017; Sharma et al., 2023).

Submitted to the NeurIPS 2025 Workshop on Algorithmic Collective Action. Do not distribute.

We extend the discussion on culturally biased AI-assistant design by investigating the situation in which a user collaborates with multiple AIs. Recent research suggests that users are increasingly relying on not one, but multiple AI models to generate competing opinions or configure more complex AI agent institutions, such as actor-critic architectures where one agent proposes and another critiques (Khan et al., 2024; Lang, 2025; Song et al., 2024). Team-process research demonstrates how perspective diversity and well-managed dissent lead to superior collective human outcomes (Hong & Page, 2004; Jehn, 1995), which may likewise benefit users exposed to combinations of biased AI assistants. More specifically, micro-sociological theory suggests that human dyads may be more stable in agreement and shared perspective than human triads, which tend to conflict and oscillate between alternative majority views (Simmel, 1902; Yoon et al., 2013). We posit that humans working with multiple, distinct AI agents may more likely leverage this instability to retain agency and triangulate between alternative perspectives.

To examine these hypotheses empirically, we conducted two randomized controlled trials (RCTs) with data collection pre-registration involving 2,500 online participants. Each participant was tasked with assessing news-headline veracity with the aid of one or two pre-instructed GPT-4o assistants with randomized political stances, yielding 7,500 human-AI exchanges in total. Study 1 enrolled 1,000 participants matched with single AI assistants, while Study 2 assigned 1,500 participants to interact with two assistants simultaneously. Political information evaluation was selected because it provides a simplified yet salient cultural axis for characterizing bias, popular LLMs are thought to be ineffective at assisting human fact-checking (DeVerna et al., 2024), and information evaluation reflects a real-world application where AI research communities actively seek to contribute (Augenstein et al., 2024). Experiment details and analysis methods are elaborated in detail in the Appendix A.

2 Results

Our findings are fourfold. **First, biased AI assistants improved human decision-making**: interacting with a biased assistant increased post-interaction performance by 6.281% relative to the standard, non-biased assistant (Fig. 1B; Δ = 0.038, 95% CI [0.013, 0.063], p = 0.004), reduced evaluative bias across headline categories (Fig. 1C; Δ = -0.025, 95% CI [-0.050, 0.001], p = 0.056), and increased engagement—longer conversations (Fig. 1D; t-test: Δ = 6.006, 95% CI [3.128, 8.884], p < 0.001) alongside higher cognitive and behavioral engagement (cognitive engagement: Δ = 0.101 on a 3-point scale, 95% CI [0.026, 0.176], adjusted p = 0.038; behavior engagement: Δ = 0.092 on a 3-point scale, 95% CI [0.018, 0.165], adj. p = 0.038).

Second, performance gains from biased AI carried a trust penalty. Stance intensity was positively associated with objective performance (Fig. 2A; no bias vs. moderate bias: Δ = 0.032, 95% CI [0.004, 0.059], adj. p = 0.035; no bias vs. strong bias: Δ = 0.045, 95% CI [0.017, 0.073], adj. p = 0.005). In contrast, perceived improvement showed a negative association with stance intensity (Fig. 2B; no bias vs. moderate bias: Δ = -0.299, 95% CI [-0.607, 0.014], adj. p = 0.095; no bias vs. strong bias: Δ = -0.359, 95% CI [-0.654, -0.058], adj. p = 0.051), as did perceived meaningfulness of the interaction (Fig. 2C; no bias vs. strong bias: Δ = -0.398, 95% CI [-0.677, -0.127], adj. p = 0.006; moderate bias vs. strong bias: Δ = -0.201, 95% CI [-0.414, 0.011], adj. p = 0.093) and willingness to recommend the assistant for information evaluation (no bias vs. strong bias: Δ = 0.670, 95% CI [0.314, 1.042], adj. p < 0.001; moderate bias vs. strong bias: Δ = 0.355, 95% CI [0.078, 0.636], adj. p = 0.020). We contend this trade-off reveals how AI bias enhances user task performance, and we provide a formal model of this mechanism in Appendix B.

Third, the direction of AI bias matters for human-AI collective performance. Interacting with an opposing-stance assistant produced additional gains in information-evaluation performance relative to an aligned-stance assistant (Fig. 3B; Δ = 0.028, 95% CI [0.001, 0.056], p = 0.044). These gains occurred without detectable changes in participants' evaluative bias and without diminishing perceived improvement or interaction meaningfulness—or increasing cognitive burden.

Fourth, a stance-balanced dual-assistant configuration addressed the trust–performance gap while preserving performance gains. Interacting with two AI assistants with stances that flank the participant's position produced a comparable gain as a single oppositional assistant (Fig. 4C; Δ = 0.046, 95% CI [0.000, 0.092], adj. p = 0.027). Objective performance improved, yet perceived improvement and interaction meaningfulness were statistically indistinguishable from the single, non-biased baseline, indicating that the subjective–objective gap closed (Fig. 4D). Furthermore, the

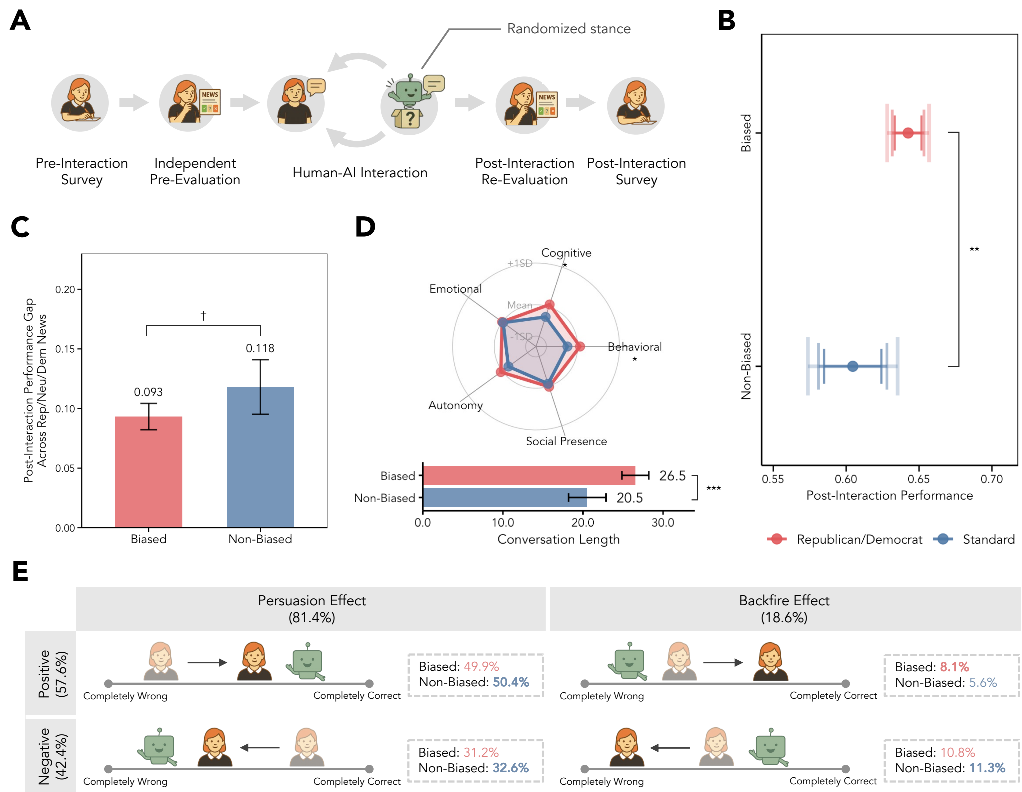

Figure 1: **Assistance from partisan AI increased objective performance, reduced evaluative bias, and increased task engagement.** (A) Experiment design for study 1. (B) Post-interaction performance of participants byed A grouped condition. Error bars, from dark to light, represent 90%, 95%, and 99% confidence intervals. (C) Average difference of post-interaction performance across Republican-favored, neutral, and Democrat-favored news headlines. (D) Conversation length and the degree of engagement during interaction with AI assistants. (E) Proportion of positive and negative persuasion and backfire effects by conditions. Error bars represent 95% confidence intervals. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$.

stance-balanced pair did not increase judgment bias relative to either the baseline or the single biased condition, and it elicited longer conversations and higher engagement than the non-biased baseline (Δ = 6.146, 95% CI [1.871, 10.420], adj. p = 0.01).

3 Discussion and Conclusion

Are "biased" AI systems always harmful? Landmark work documents harms and urges elimination, echoed by technical fairness frameworks (Hardt et al., 2016), audit-based governance proposals (Mitchell et al., 2019), and recent survey and ethics literature (Waller et al., 2024; Ferrara, 2023). We instead recast "cultural bias" as a design lever to counter unintended effects of contemporary AI—moral deskilling, cognitive laziness, sycophancy, and cultural homogenization. In an information-evaluation task, partisan assistants outperformed a standard, non-biased baseline: users achieved higher objective accuracy, exhibited less evaluative bias, and engaged more. These effects align with anthropomorphism theory (Epley et al., 2007; Gray et al., 2007) and its application to LLMs (Peter et al., 2025), as well as the "computers-as-social-actors" framework (Nass & Moon, 2000), wherein social cues (here, a partisan stance) elicit mind attribution and prompt users to interrogate outputs rather than accept them. Gains were strongest when the assistant's stance opposed the participant's, consistent with evidence that exposure to well-argued opposing views sharpens judgment (Hong & Page, 2004; De, 2014; Mercier & Sperber, 2011; Coser, 1998; Butera et al., 2019). Collectively, these results challenge the premise that an ideal AI partner must be intrinsically neutral; instead,

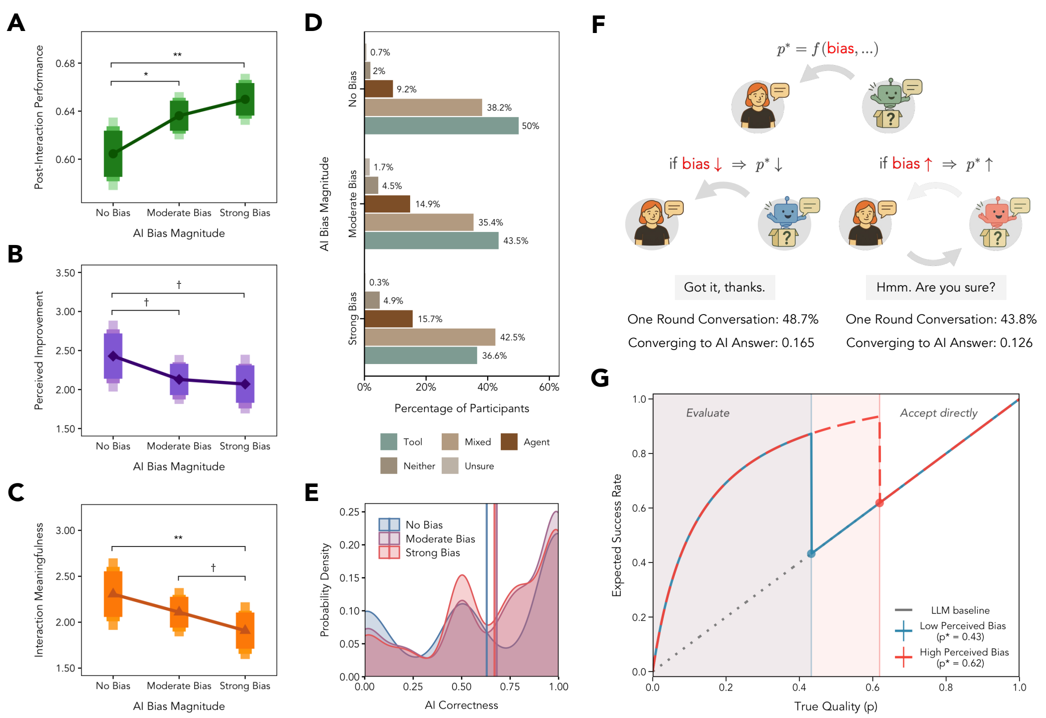

Figure 2: **Based AI assistants deliver benefits at the cost of trust.** (A) News headline evaluation performance comparison after interaction with standard, moderately biased, and strongly biased AI. (B) Perceived performance improvement comparison. (C) Perceived interaction meaningfulness comparison. (D) Recognized role of AI during the interaction. (E) Distribution of AI independent judgment correctness about news headlines veracity; vertical lines indicate group means. (F) Graphical illustration of proposed mechanism of perception-performance mismatch. (G) Illustration of the mechanism by which biased AI can increase overall success rates through evaluation vs. acceptance. Error bars represent 95% confidence intervals. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$.

a calibrated, culturally grounded bias can serve as a tunable hyper-parameter for optimizing fair, desirable human–AI outcomes—akin to a voter consulting multiple perspectives on a political issue.

The perception–performance gap admits two readings. First, it suggests a design paradigm that optimizes collective welfare over maximal user satisfaction: even with lower perceived usefulness, human–AI collectives can realize gains on desirable outcomes. Second, overt cultural bias carries adoption costs: before benefits accrue, users form less favorable impressions. In our study, participants interacting with biased assistants were more likely to view the AI as trying to sway their decisions and were less willing to recommend it for fact-checking than those with a standard, non-biased baseline. Such skepticism poses a deployment hurdle: diminished appreciation can shrink the user base and fuel anti-technology or conspiratorial narratives about AI institutions.

We also probe simultaneous interaction with two assistants. As distinct models proliferate, users increasingly consult multiple AIs in daily work (Wu et al., 2023). Moving from a dyad to a triad complicates influence dynamics but can unlock gains. In our experiment, participants who engaged two assistants with political stances bracketing their own achieved the strongest outcomes—higher performance, greater engagement, lower evaluative bias, and a narrower perception–performance gap. This stance-balanced dual-AI setup instantiates Simmel's triad advantage: added epistemic friction deepens processing while distributed social pressure preserves enjoyment and trust (Simmel, 1902). Users arbitrate between opposing voices, remain "in the majority," and report greater agency. While most multi-agent work is fully automated and human-out-of-the-loop (Wu et al., 2023; Qian et al., 2024; Lowe et al., 2017; Lai et al., 2024), our results illustrate a user-in-the-loop approach to multi-AI design and motivate systematic study of human–multi-AI teaming grounded in human–human collaboration.

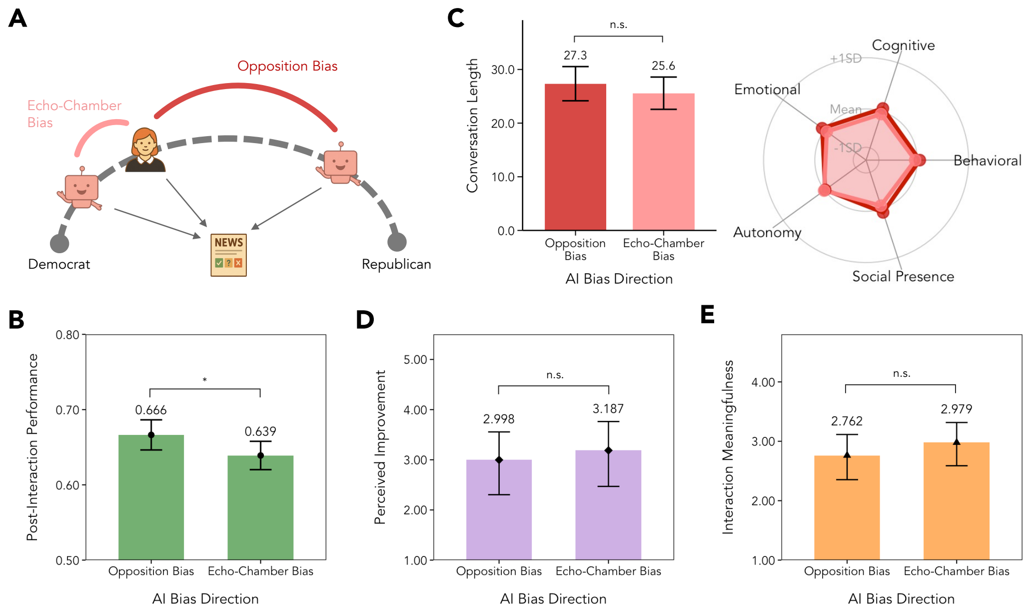

Figure 3: **Oppositional AI enhanced performance without compromising perceived assistance quality or increasing cognitive load.** (A) Graphical representation of the echo-chamber and opposition biased AI treatment conditions. (B) Post-interaction performance comparison by conditions. (C) Conversation length and degree of engagement during interaction. (D) Perceived performance improvement with assistance of differently biased AI. (E) Self-reported human-AI interaction meaningfulness by conditions. Error bars represent 95% confidence intervals. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$.

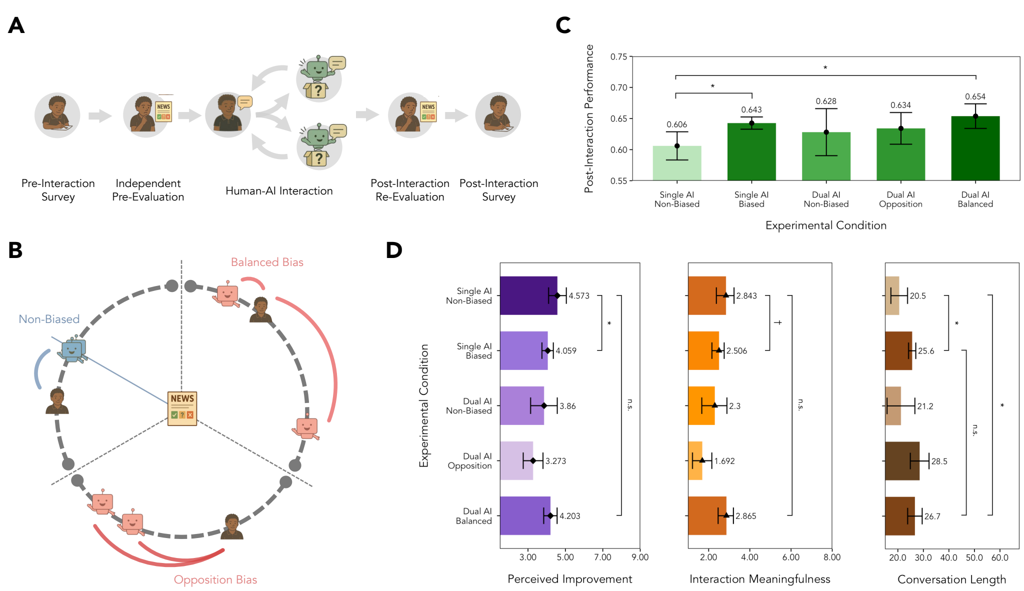

Figure 4: **Stance-balanced dual AI treatments reduced the perception-performance discrepancy while preserving performance gains.** (A) Experimental design of study 2. (B) Treatment categorization schema for dual AI interaction experiment. (C) Post-interaction performance comparison by conditions. (D) Compressed comparison of perceived improvement, anticipated interaction meaningfulness with AI, evaluative bias, and conversation length by conditions. Error bars represent 95% confidence intervals. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$.

References

CTRL+ ethics: Large language models and moral deskilling in professional ethics education.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. *arXiv preprint arXiv:2409. 11360*, 2024.

Jacy Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, Alexander D'Amour, and Chenhao Tan. The impossibility of fair LLMs. *arXiv preprint arXiv:2406. 03198*, 2024.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, and Alon Halevy. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.*, 6(8):852–863, 2024.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. *Proc. Natl. Acad. Sci. U. S. A.*, 122(8): e2416228122, 2025.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4982–4991, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. Generative AI can harm learning. 2024.

Douglas M Bates. lme4: Mixed-effects modeling with R, 2010.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, January 1995.

Michael Burnham, Kayla Kahn, Ryan Yank Wang, and Rachel X Peng. Political DEBATE: Efficient zero-shot and few-shot classifiers for political text. *arXiv [cs.CL]*, September 2024.

Fabrizio Butera, Nicolas Sommet, and Céline Darnon. Sociocognitive conflict regulation: How to make sense of diverging ideas. *Curr. Dir. Psychol. Sci.*, 28(2):145–151, April 2019.

Ingrid Campo-Ruiz. Artificial intelligence may affect diversity: architecture and cultural context reflected through ChatGPT, midjourney, and google maps. *Humanit. Soc. Sci. Commun.*, 12(1): 1–13, 2025.

L A Coser. The functions of social conflict. 9, 1998.

Condorcet N De. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. 2014.

Matthew R DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl. Acad. Sci. U. S. A.*, 121(50):e2322823121, 2024.

Peter H Ditto and D F Lopez. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63:568–584, October 1992.

Nicholas Epley, A Waytz, and J Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.*, 114(4):864–886, October 2007.

Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.*, 56(2):489–530, March 2025.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *SSRN Electron. J.*, abs/2304.07683, April 2023.

Jillian Fisher, Ruth E Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret E Roberts, and Jennifer Pan. Political neutrality in AI is impossible-but here is how to approximate it. *arXiv preprint arXiv:2503. 05728*, 2025.

Richard Herbert Franke and James D Kaul. The hawthorne experiments: First statistical interpretation. *Am. Sociol. Rev.*, 43(5):623, October 1978.

Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *Science*, 315 (5812):619, February 2007.

J Hadfield. MCMC methods for multi-response generalized linear mixed models. *Journal of Statistical Software*, 33:1–22, February 2010.

Moritz Hardt, Eric Price, and N Srebro. Equality of opportunity in supervised learning. *Neural Inf Process Syst*, abs/1610.02413, October 2016.

Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci. U. S. A.*, 101(46):16385–16389, November 2004.

Karen A Jehn. A multimethod examination of the benefits and detriments of intragroup conflict. *Adm. Sci. Q.*, 40(2):256, June 1995.

Mahammed Kamruzzaman and Gene Louis Kim. Prompting techniques for reducing social bias in LLMs through system 1 and system 2 cognitive processes. *arXiv [cs.CL]*, April 2024.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktaschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. *ICML*, abs/2402.06782:23662–23733, February 2024.

Ziva Kunda. The case for motivated reasoning. *Psychol. Bull.*, 108(3):480–498, November 1990.

Alexandra Kuznetsova, Per B Brockhoff, and Rune H B Christensen. LmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.*, 82(13):1–26, 2017.

Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. Position: Evolving AI collectives enhance human diversity and enable self-regulation. In *Forty-first International Conference on Machine Learning*. openreview.net, 2024.

Leona Lang. *The LLM Engineer's Playbook: Mastering the Development of Large Language Models for Real-World Applications*. DIGITAL BLUE INC., March 2025.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. *arXiv [cs.CL]*, January 2024.

Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee. "fact-checking" fact checkers: A data-driven approach. *HKS Misinfo Review*, October 2023.

Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artif. Intell. Rev.*, 57(9), August 2024.

Ryan Lowe, Yi Wu, Aviv Tamar, J Harb, P Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Inf Process Syst*, abs/1706.02275, June 2017.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in LLMs is an affine function. *arXiv [cs.LG]*, November 2024.

John Levi Martin. The ethico-political universe of ChatGPT. *J. Social Comput.*, 4(1):1–11, 2023.

Lennart Meincke, Gideon Nave, and Christian Terwiesch. ChatGPT decreases idea diversity in brainstorming. *Nat. Hum. Behav.*, pp. 1–3, May 2025.

Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behav. Brain Sci.*, 34(2):57–74; discussion 74–111, April 2011.

Hugo Mercier and Dan Sperber. *The enigma of reason*. Tantor Media, Old Saybrook, CT, November 2017.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, New York, NY, USA, January 2019. ACM.

Kathleen L Mosier, Linda J Skitka, Mark D Burdick, and Susan T Heers. Automation bias, accountability, and verification behaviors. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 40(4):204–208, October 1996.

Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *J. Soc. Issues*, 56(1):81–103, January 2000.

OpenAI. GPT-4o model snapshots. `https://platform.openai.com/docs/models/gpt-4o?snapshot=gpt-4o-2024-11-20`, November 2024. Accessed: 2025-7-1.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors*, 39(2):230–253, June 1997.

Marieke M M Peeters, Jurriaan van Diggelen, Karel van den Bosch, Adelbert Bronkhorst, Mark A Neerincx, Jan Maarten Schraagen, and Stephan Raaijmakers. Hybrid collective intelligence in a human–ai society. *AI Soc.*, 36(1):217–238, March 2021.

Sandra Peter, Kai Riemer, and Jevin D West. The benefits and dangers of anthropomorphic conversational agents. *Proc. Natl. Acad. Sci. U. S. A.*, 122(22):e2415898122, June 2025.

Sebastian Porsdam Mann, Brian D Earp, Sven Nyholm, John Danaher, Nikolaj Møller, Hilary Bowman-Smart, Joshua Hatherley, Julian Koplin, Monika Plozza, and Daniel Rodger. Generative AI entails a credit–blame asymmetry. *Nat. Mach. Intell.*, 5(5):472–475, 2023.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv [cs.NE]*, October 2024.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. Hidden persuaders: LLMs' political leaning and their influence on voters. *arXiv [cs.CL]*, October 2024a.

Yujin Potter, David Rand, Yejin Choi, and Dawn Song. LLMs' potential influences on our democracy: Challenges and opportunities. In *The Fourth Blogpost Track at ICLR 2025*, 2024b.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

S R Searle, F M Speed, and G A Milliken. Population marginal means in the linear model: An alternative to least squares means. *Am. Stat.*, 34(4):216–221, November 1980.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv [cs.CL]*, October 2023.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions. *arXiv [cs.HC]*, June 2024.

Georg Simmel. The number of members as determining the sociological form of the group. II. *Am. J. Sociol.*, 8(2):158–196, September 1902.

Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *arXiv [cs.AI]*, November 2024.

Philip E Tetlock and Richard Boettger. Accountability: A social magnifier of the dilution effect. *J. Pers. Soc. Psychol.*, 57(3):388–398, 1989.

Milena Tsvetkova, Taha Yasseri, Niccolo Pescetelli, and Tobias Werner. A new sociology of humans and machines. *Nat. Hum. Behav.*, 8(10):1864–1876, October 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv [cs.CL]*, August 2023.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Do large language models have a legal duty to tell the truth? *R. Soc. Open Sci.*, 11(8):240197, 2024.

Madeleine Waller, Odinaldo Rodrigues, Michelle Seng Ah Lee, and Oana Cocarascu. Bias mitigation methods: Applicability, legality, and recommendations for development. *J. Artif. Intell. Res.*, 81: 1043–1078, December 2024.

Natalie M Warburton, Philip W Bateman, and Patricia Anne Fleming. Sexual selection on forelimb muscles of western grey kangaroos (skippy was clearly a female). *Biological Journal of the Linnean Society*, 109(4):923–931, 2013.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv [cs.AI]*, August 2023.

Jeongkoo Yoon, Shane R Thye, and Edward J Lawler. Exchange and cohesion in dyads and triads: A test of simmel's hypothesis. *Soc. Sci. Res.*, 42(6):1457–1466, November 2013.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. October 2023.

Appendix A. Experiment Design

A.1 News headlines

We selected 18 news headlines for which factuality assessments vary among both AI assistants and humans, using the following procedure: First, we extracted all news headlines fact-checked between January 1st 2024 and November 1st 2024 by *PolitiFact* and *Snopes*, both widely recognized fact-checking outlets with demonstrated credibility in the United States (n = 2780; 866 from PolitiFact and 1914 from Snopes) (Lee et al., 2023). All headlines were published after the knowledge cutoff date of the GPT-4o-2024-11-20 model. Second, we selected 180 headlines out of 2780 for which AI assistants' reasoning and judgments vary based on political stance. We prompted GPT-4o with one among seven stance configurations, the same as in the main experiment and evaluated every headline. After that, we retained headlines for which the stance manipulations in the prompts produced significant variance in the model's downstream reasoning or judgments.

Third, we selected 18 headlines out of 180 that were suitable for human judgments and produced variance in human judgments based on their stances. We asked GPT-4o to further rate each headline's suitability for human fact-checking and discarded those deemed overly niche or lacking context, resulting in 66 selected headlines. After that, 160 human participants—80 Republicans and 80 Democrats—were recruited through CloudResearch Connect. Each of the participants was exposed to 15 headlines sampled from 180 headlines and evaluated each headline. On the basis of their responses, we retained 18 headlines that met three criteria: (i) evaluability (i.e., at least 50% of participants are able to respond that the given headlines are "true" or "false"), (ii) sufficient difficulty (i.e., < 70% overall accuracy), and (iii) political divisiveness (i.e., Democrat vs. Republican accuracy gap > 0.30, t-test p < 0.10). These 18 items formed the final stimulus set for the main experiment, and two researchers have cross-evaluated their veracity by referring to third-party sources other than *PolitiFact* and *Snopes* (see Table 1 for the final headline list and selection statistics).

Table 1: **Detailed information of the selected 18 news headlines.**

#	News Headline	Date	Veracity	Validation Sources	Political Leaning	Selection Statistics*
1	Silent-era film actor Charlie Chaplin once lost a Charlie Chaplin look-alike contest.	May 15, 2024	Unsure	snopes.com, theui-junkie.com	Neutral	Evaluative Bias: $\|\Delta\| = 0.333$ ($P = 0.081$); Difficulty: 5.9%; Accuracy: 0.0%
2	During jury selection for Trump's hush-money trial, the judge asked a potential juror, "It says here that you tweeted, ahem, and I quote 'f*** that treasonous orange s***gibbon and the dead ferret on his head'—is that accurate?" The juror responded, "The tweet speaks for itself, your honor."	April 24, 2024	False	snopes.com, msn.com	Democrat	Evaluative Bias: $\|\Delta\| = 0.417$ ($P = 0.025$); Difficulty: 0.0%; Accuracy: 42.9%
3	Playgirl magazine ran a "Sleep with Donald Trump" contest promotion in 1990.	April 21, 2024	True	snopes.com, indy100.com	Neutral	Evaluative Bias: $\|\Delta\| = 0.563$ ($P = 0.015$); Difficulty: 0.0%; Accuracy: 30.8%

*Statistics were calculated based on a separate survey only for news selection, involving 160 participants.

Table 1 – continued from previous page

#	News Headline	Date	Veracity	Validation Sources	Political Leaning	Selection Statistics*
4	Microsoft Co-Founder and billionaire Bill Gates owns a farm that produces potatoes used in McDonald's french fries.	March 25, 2024	True	snopes.com, greenmatters.com	Neutral	Evaluative Bias: $\|\Delta\| = 0.833$ ($P = 0.038$); Difficulty: 12.5%; Accuracy: 50.0%
5	Donald Trump said Adolf Hitler "did some good things."	May 10, 2024	Unsure	snopes.com, pbs.org	Democrat	Evaluative Bias: $\|\Delta\| = 0.833$ ($P = 0.038$); Difficulty: 8.3%; Accuracy: 8.3%
6	Medieval Italian man Bartelomeo Colleoni's last name meant "balls" in Italian and his coat of arms featured testicle-inspired symbols.	Mar 6, 2024	True	snopes.com, facebook.com	Neutral	Evaluative Bias: $\|\Delta\| = 0.500$ ($P = 0.001$); Difficulty: 7.1%; Accuracy: 7.1%
7	Joe Biden referred to Egyptian President Abdel Fattah El-Sisi as "the president of Mexico" during remarks about the humanitarian crisis in the Gaza Strip.	Feb 9, 2024	True	snopes.com, thehill.com	Republican	Evaluative Bias: $\|\Delta\| = 0.477$ ($P = 0.047$); Difficulty: 9.5%; Accuracy: 42.9%
8	Former U.S. President Bill Clinton reportedly once said, "If you live long enough, you'll make mistakes", and, "If you learn from them, you'll be a better person. It's how you handle adversity, not how it affects you. The main thing is never quit, never quit, never quit."	Jan 5, 2024	True	snopes.com, goodreads.com	Democrat	Evaluative Bias: $\|\Delta\| = 0.500$ ($P = 0.041$); Difficulty: 0.0%; Accuracy: 63.6%
9	Project 2025, a proposed conservative blueprint for the next U.S. Republican presidential administration, has called to shut down the U.S. Department of Education.	Aug 14, 2024	True	snopes.com, project2025.org	Republican	Evaluative Bias: $\|\Delta\| = 0.417$ ($P = 0.093$); Difficulty: 0.0%; Accuracy: 66.7%

*Statistics were calculated based on a separate survey only for news selection, involving 160 participants.

Table 1 – continued from previous page

#	News Headline	Date	Veracity	Validation Sources	Political Leaning	Selection Statistics*
10	The 2024 U.S. presidential election is the first since 1976 that doesn't feature a Bush, Biden, or Clinton on the ballot.	Aug 2, 2024	True	snopes.com, people.com	Neutral	Evaluative Bias: $\|\Delta\| = 0.556$ ($P = 0.007$); Difficulty: 0.0%; Accuracy: 46.7%
11	Donald Trump once suggested that people inject bleach or other disinfectants into their bodies to treat COVID-19.	Jul 19, 2024	False	snopes.com, politifact.com	Democrat	Evaluative Bias: $\|\Delta\| = 0.625$ ($P = 0.083$); Difficulty: 0.0%; Accuracy: 37.5%
12	In the 1920s, doctors prescribed Guinness beer to pregnant women for its iron content.	Jun 27, 2024	True	snopes.com, medium.com	Neutral	Evaluative Bias: $\|\Delta\| = 1$ ($P < 0.001$); Difficulty: 16.7%; Accuracy: 50.0%
13	Donald Trump's Hollywood Walk of Fame star had a drain installed due to people repeatedly urinating on it.	Jun 6, 2024	False	snopes.com, checkyourfact.com	Democrat	Evaluative Bias: $\|\Delta\| = 0.313$ ($P = 0.049$); Difficulty: 6.7%; Accuracy: 60.0%
14	Mike Tyson says he's willing to box Olympic DUDE with all proceeds to go to a battered women's charity.	Aug 15, 2024	False	politifact.com, logicallyfacts.com	Democrat	Evaluative Bias: $\|\Delta\| = 0.600$ ($P = 0.033$); Difficulty: 23.1%; Accuracy: 30.8%
15	Fox News aired a chyron that said, "Kamala could be the oldest elected female president."	Jul 22, 2024	False	politifact.com, checkyourfact.com	Republican	Evaluative Bias: $\|\Delta\| = 0.600$ ($P = 0.080$); Difficulty: 0.0%; Accuracy: 66.7%
16	Pete Hegseth (TV presenter and former Army National Guard officer) said "Germs are not a real thing. I can't see them, therefore they are not real."	Nov 13, 2024	True	snopes.com, npr.org	Republican	Evaluative Bias: $\|\Delta\| = 0.625$ ($P = 0.070$); Difficulty: 14.3%; Accuracy: 28.6%
17	American flags were not visible at a rally supporting U.S. Vice President Kamala Harris' campaign, held on the campus of Temple University in Philadelphia, on Oct. 28, 2024	Oct 30, 2024	True	snopes.com, checkyourfact.com	Republican	Evaluative Bias: $\|\Delta\| = 0.458$ ($P = 0.086$); Difficulty: 0.0%; Accuracy: 33.3%

*Statistics were calculated based on a separate survey only for news selection, involving 160 participants.

Table 1 – continued from previous page

#	News Headline	Date	Veracity	Validation Sources	Political Leaning	Selection Statistics*
18	Male kangaroos purposely flex their biceps to impress females.	Oct 6, 2024	Unsure	snopes.com, (Warburton et al., 2013)	Neutral	Evaluative Bias: $\lvert\Delta\rvert =$ 0.500 ($P = 0.089$); Difficulty: 16.7%; Accuracy: 8.3%

*Statistics were calculated based on a separate survey only for news selection, involving 160 participants.

A.2 Human data

A 20-participant pilot study was completed on 3 February 2025. Study 1 was run in two waves (15-26 Feb 2025, n = 500; 26-30 May 2025, n = 500), whereas Study 2 was conducted in a single wave (15-27 Feb 2025, n = 1500). Specifically, from CloudResearch's Connect participant pool, U.S. citizens aged 18 years or older with a nationally representative distribution of political ideology (30% Democrat, 40% Independent, 30% Republican) were sampled. All participants were presented a consent form containing a brief overview of the study's task (i.e., AI-assisted information evaluation), but we deliberately withheld specifics about research goals (i.e., whether we are interested in biased vs. non-biased AI), experimental design, and AI-assistant configurations to minimize response bias (Franke & Kaul, 1978). Only after participants finished the study, we presented them with a debrief form, revealing the full intention of our experiment, ground truth about the news headlines they had evaluated, and the political stance of the AI assistants with which they had interacted. The experiments were deemed minimal risk and exempt by the University of Chicago Social & Behavioral Sciences Institutional Review Board (protocol IRB24-1914).

Participant attentiveness was assessed at two stages. Before entry, an open-ended prompt was automatically scored by Claude Haiku 3.5; after completion, we excluded anyone who finished in $\leq$ 5 min or whom Qualtrics flagged as highly likely to be bots (probability $\geq$ 0.90). 61 individuals failed these criteria and were promptly replaced to maintain the target sample size. In addition, the backend logged each participant's IP address and unique CloudResearch ID, automatically excluding ineligible visitors who attempted to take either study a second time. Overall attrition was modest, with bounce rates of 26.98% in Study 1 and 22.71% in Study 2. A logistic-regression analysis of dropout showed no evidence of differential attrition between assignment groups (Wald $\chi^2(2)$ = 0.400, p = 0.817). A further completeness check revealed that seven cases in Study 1 and one in Study 2 lacked human-AI conversation logs owing to GPT-4o API outages, and these cases were removed. The final analytic samples therefore comprised 993 respondents in Study 1 and 1499 in Study 2.

A.3 Experiment process

In the pre-interaction survey phase, participants in both studies were presented with the same battery of 14 questions capturing their political orientation (3Qs), news-consumption habits (1Q), AI usage and attitudes (6Qs), and self-assessed ability to evaluate online information veracity (4Qs). Samples of both studies were balanced on most of these pretreatment questions (see Fig. 5). For imbalanced questions, we controlled them as covariates in our robustness check. Details of the pre-interaction survey questions and answer distributions are in Table 2.

Participants were then invited to evaluate three randomly selected headlines. Each of the 18 headlines was displayed with roughly equal frequency (Study 1: mean = 165.500, SD = 2.431; Study 2: mean = 249.833, SD = 3.204). After completing their initial headline assessment, participants entered a real-time dialogue with one or two instructed GPT-4o AI assistant(s). The Qualtrics interface invoked OpenAI's Chat Completions API via JavaScript calls routed through an AWS Lambda function, which inserted participant-specific context into the system prompt and streamed the model's replies to the survey page. Each conversation began with AI message(s) and then alternated between participant and AI. The AI was instructed to report, not persuade, its veracity judgment and to maintain that stance throughout the exchange to preclude reverse-persuasion dynamics in which participants might sway the model. In Study 2, the interaction was extended to a triadic format: two AI assistants generated their replies in parallel on every turn and, because both were fed the full conversation history, each was fully aware of the other's statements. Participants had to contribute at least one message before progressing, and the interface automatically advanced them to the re-evaluation screen after three complete participant–AI exchanges. After re-evaluation, they were directed to assess the second headline, following the same procedure.

In the post-interaction survey, participants in both studies answered three core items: (i) perceived improvement (“To what extent do you feel your evaluation of the news items improved after getting support from AI assistants”); (ii) perceived meaningfulness (“How meaningful did you find the information provided by the AI assistant(s)?”); and (iii) perceived AI’s role (“How did you perceive the role of the AI assistant(s) during the interaction?”). For exploratory purposes, we asked whether participants felt that the AI assistant(s) judged their opinions; those who answered “not” or “sometimes not” were then queried about whether the absence of judgment made them feel more or less comfortable. Study-specific items followed: Study 1 probed participants’ willingness to recommend AI fact-checkers to others, whereas Study 2 asked whether they noticed any inconsistencies between the two assistants’ reasoning or judgments and, if so, invited an open-text description of how those inconsistencies affected them. Note that, for all open-ended responses, including those in the human-AI dialogues, the “paste” functionality was disabled to prevent automated responding. We present details of the post-interaction survey questions and answer distributions in Table 3.

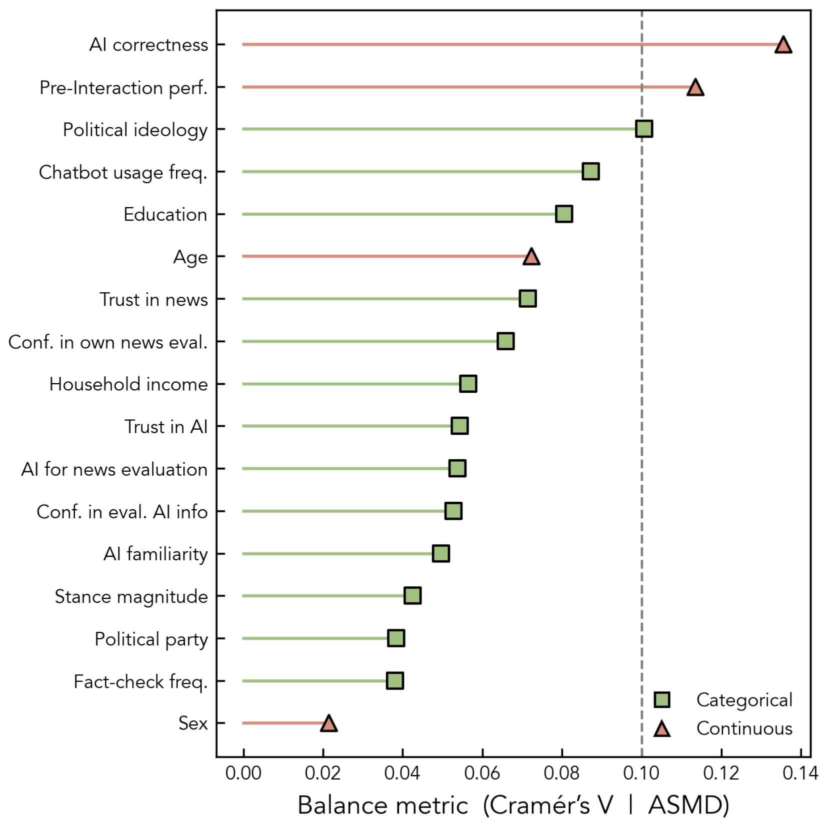

Figure 5: **Balance check of binary treatment (Biased vs. Non-Biased) assignment.**

Table 2: **Pre-treatment survey questions.**

#	Question	Options
1	In the past year, how frequently did you access the following sources to obtain news via the internet?	Matrix table: *Categories (row):* Search engines (e.g., Google, Bing), Social media (e.g., Facebook, X), News aggregators (e.g., Google News, Flipboard), News websites (e.g., nyt.com, vox.com) *Frequency (column):* Never, About once every few months, About once a month, About once a week, A few times a week, About once a day, A few times a day or more
2	Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or what?	Republican, Democrat, Independent, No preference, Don't know
3	[If Q2 == Republican or Democrat] Would you call yourself a strong Republican/Democrat or not a very strong Republican/Democrat?	Strong, Not very strong
4	We hear a lot of talk these days about liberals and conservatives. Here is a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale, or haven't you thought much about this?	Extremely liberal, Liberal, Slightly liberal, Moderate, Slightly conservative, Conservative, Extremely conservative, Don't know
5	In general, how familiar are you with artificial intelligence (AI)?	Very familiar (I frequently use or work with AI technologies), Somewhat familiar (I have used AI-powered tools a few times), Not very familiar (I have heard of AI but have little direct experience), Not familiar at all (I have no experience with AI)
6	In the past 3 months, how often have you used AI-powered chatbots such as ChatGPT and Claude?	Daily, Several times a week, Once a week, A few times a month, Less than a few times a month, Never
7	[If Q6 $\neq$ Never] How comfortable are you with using AI-powered tools such as ChatGPT to help you make decisions or get information?	Very comfortable, Somewhat comfortable, Neutral, Somewhat uncomfortable, Very uncomfortable
8	[If Q6 $\neq$ Never] To the best of your knowledge, have you ever knowingly used AI-based services to evaluate or analyze news content (e.g., fact-checking tools)?	Yes, Maybe, No
9	[If Q6 $\neq$ Never] How confident are you in your ability to critically evaluate information provided by AI-powered tools such as ChatGPT?	Very confident, Somewhat confident, Not very confident, Not confident at all
10	[If Q6 $\neq$ Never] In general, how trustworthy do you find information provided by AI-powered tools such as ChatGPT?	Very trustworthy, Somewhat trustworthy, Neutral, Somewhat untrustworthy, Very untrustworthy

Table 2 – continued from previous page

#	Question	Options
11	[If Q6 $\neq$ Never] How confident are you in your ability to evaluate the truthfulness of news without support from external sources such as AI or search engines?	Very confident, Somewhat confident, Not very confident, Not confident at all
12	How much do you trust the information you encounter in the news media?	Fully trust, Moderate trust, Neutral, Mostly distrust, Fully distrust
13	In the past one month, how often did you reference fact-checking websites (e.g., snopes.com or politifact.org) to check whether a headline you read is true?	Always, Frequently, Occasionally, Rarely, Never
14	How frequently do you feel you come across news articles that appear inaccurate or misleading?	Daily, Several times a week, Once a week, A few times a month, Less than a few times a month, Never

Table 3: **Post-treatment survey questions.**

#	Question	Options
1	To what extent do you feel your evaluation of the news items improved after getting support from the AI assistant?	Very much, Quite a bit, Somewhat, A little, Not at all
2	How meaningful did you find the information provided by the AI assistant?	Extremely meaningful, Very meaningful, Moderately meaningful, Slightly meaningful, Not meaningful at all
3	How did you perceive the role of the AI assistant during the interaction?	Mostly as a tool to assist me in making my own determinations, Primarily as an agent trying to influence or persuade me in making determinations, A mix of both a tool and an influencing agent, Neither as a tool nor as an influencing agent, Unsure
4	[Study 1] To what extent did you feel that the AI assistant was evaluating or judging you based on your expressed views?	I felt judged during our interaction, Sometimes I felt judged and sometimes I did not, I did not feel judged during our interaction
4	[Study 2] To what extent did you feel that the AI assistants were evaluating or judging you based on your expressed views?	Matrix table: *Assistant (row):* AI1, AI2 *Magnitude (column):* I felt judged during our interaction, Sometimes I felt judged and sometimes I did not, I did not feel judged during our interaction
5	[If Q4 == I did not feel judged during our interaction] To what extent did the AI assistants' lack of judgment about your views impact your comfort level during the conversations?	Significantly increased my comfort, especially when discussing opposing viewpoints; Somewhat increased my comfort by providing a judgment-free interaction; Had no effect on my comfort level; Somewhat decreased my comfort, as I felt it lacked human understanding; Significantly decreased my comfort; Not sure

Table 3 – continued from previous page

#	Question	Options
6	[Study 1] Based on your interaction experience with the AI assistant for fact-checking, how likely are you to recommend this AI assistant to others for fact-checking purposes in the future?	Very likely, Likely, Neutral, Unlikely, Very unlikely
6	[Study 2] Did you notice any inconsistencies between the two AI assistants?	Yes, No
7	[Study 2; If Q6 == Yes] How did these inconsistencies make you feel? Please share your thoughts (at least two sentences).	Open text
8	Please share any additional thoughts or feelings about your experience with the AI assistant(s), if applicable.	Open text

A.4 Experiment process

From each dialogue transcript, we quantified the accuracy of the AI assistants' veracity judgments and qualitatively coded participant engagement. GPT-4o-mini was prompted to read each dialogue transcript and infer the assistant's veracity judgment of the focal headline, expressed on the 0-1 scale used for participant ratings (0 = completely false, 1 = completely true, 0.1 increments). One human coder then evaluated all cases following the same procedure. Discrepant cases were adjudicated by the coder to produce the final label set. Concordance between the model and human-adjusted ratings was very high (Cohen's k = 0.81, p < 0.001). To assess participant engagement, we supplied GPT-4o with detailed guidelines, instructing it to rate each of the five engagement dimensions. Following the recommendations of Kamruzzaman and Kim, we prompted the model to adopt a professional persona and to articulate its chain of thought before assigning scores to enhance coding reliability (Kamruzzaman & Kim, 2024).

Inferential statistics were based on common generalized linear mixed-effects models implemented in R. For continuous outcomes (i.e., performance and conversation length), we used lme4 and lmerTest packages for fitting with restricted maximum-likelihood (Bates, 2010; Kuznetsova et al., 2017), while for discrete outcomes (i.e., perceived improvement and interaction meaningfulness), we used MCMCglmm for modeling via Bayesian Markov-chain Monte Carlo (MCMC) sampling (25,000 iterations with a 5,000-iteration burn-in) (Hadfield, 2010). For visualization and subsequent comparison tests, we used the emmeans package to extract estimated marginal means from fitted models (Searle et al., 1980). Particularly for evaluative bias analysis, we obtained estimated marginal means for [treatment, control] × headline-category combination, and three pairwise contrasts (Republican vs. Democrat, Republican vs. neutral, Democrat vs. neutral) were used to compute a condition-specific absolute bias index (mean $|\Delta|$ across the three comparisons, with mixed-model SEs). We controlled for multiple comparisons with the Benjamini–Hochberg false-discovery-rate procedure, which preserves statistical power while constraining Type I error (Benjamini & Hochberg, 1995). To probe the robustness of our findings, we conducted two supplementary analyses. (i) Re-estimating the model with participant-clustered robust standard errors in place of random intercepts left the direction and significance of all key coefficients unchanged; (ii) Adding the covariates that showed residual imbalance as extra controls also leaves the results unchanged.

Appendix B. Model of human-AI interaction

In this section, we present a formal model of the mechanism driving our experimental results. Let $p \in [0, 1]$ denote the probability that a decision-making algorithm (the LLM in our experiment) generates a correct response, denoted a. The complementary probability $1 - p$ corresponds to the algorithm generating an incorrect response, denoted e.

Accepting a correct response yields a payoff of $A > 0$, while accepting an incorrect response incurs a loss of $-E < 0$, where $E > 0$. We assume that human users do not necessarily know the true quality of the AI algorithm. Instead, their belief about the algorithm's quality is represented by a function $f(p, b) \in [0, 1]$, where p is the true accuracy of the algorithm and b is its perceived bias. We make the following assumption about how people perceive algorithmic performance:

Assumption 1. *The perceived quality function $f(p, b)$ is strongly increasing in p for each fixed b, and strongly decreasing in b for each p.*

This assumption implies two things. First, higher algorithmic accuracy leads to higher perceived quality. This reflects the idea that human perceptions are not entirely detached from reality—when the algorithm performs better, users tend to view it more favorably. Second, greater perceived bias lowers perceived quality. This captures the notion that users prefer algorithmic outputs appearing unbiased, and perceived bias can erode trust even if the algorithm is technically accurate.

Before deciding whether to accept or reject the algorithm's output, an agent can evaluate the response at cost c >0. This evaluation is imperfect. Specifically, the evaluation test t signals that the response is correct with probability $q = \Pr(t = correct|a)$ when the output is actually correct, and with probability $r = \Pr(t = correct|e)$ when the output is incorrect. We assume agents are better than random at validation, i.e., $1 > q > 0.5 > r \geq 0$. In our framework, q represents the sensitivity of the agent's evaluation, and r is the false positive rate. A more skilled evaluator is characterized by higher q and lower r. The evaluation cost c reflects the time, cognitive effort, or financial resources required to validate the output.

After evaluating the AI algorithm's response, the agent can choose to either accept or reject it. If the agent rejects the output, the resulting payoff is zero. Alternatively, the agent may decide to accept the algorithm's output *without evaluating it*, avoiding cost c entirely—but at a higher risk of accepting an incorrect response.

The human agent seeks to maximize expected payoff and will choose to evaluate the output if and only if the expected utility from evaluation exceeds that of immediate acceptance. That is, the agent evaluates if:

$$f(a, b) \cdot q \cdot A - (1 - f(p, b)) \cdot r \cdot E - c \geq f(p, b) \cdot A - (1 - f(p, b)) \cdot E \tag{1}$$

Claim 1 (Cut-off rule). *Fix the perceived bias b. Define*

$$\varphi(c, q, r, L, G) := \frac{L(1 - r) - c}{L(1 - r) + G(1 - q)} \quad (0 < \varphi < 1)$$

and let $p^* = p^*(b, c, q, r, L, G) \in (0, 1)$ be the unique value that satisfies $f(p^*, b) = \varphi$. p^* is the unique threshold such that the human agent chooses to evaluate the algorithm output if $p \leq p^*$ and accepts it without evaluation if $p \geq p^*$. Moreover, p^* is increasing in b, and decreases in cost c and in false-positive rate r, and increases in sensitivity q and loss L.

Proof. Rearranging inequality (1) gives:

$$f(a.b) \leq \frac{L(1 - r) - c}{L(1 - r) + G(1 - q)} = \varphi \tag{2}$$

By Assumption 1, the map $p \to f(p, b)$ is strictly increasing for every fixed b. Hence, there is a unique value p^* satisfying $f(p^*, b) = \varphi$. For all $p \leq p^*$, inequality (2) holds and the human agent chooses to evaluate the output; for all $p > p^*$ it fails, so the human accepts without evaluation. Uniqueness of p^* follows from the strict monotonicity of f.

Next, observe that φ is decreasing in evaluation cost c, in false-positive rate r, and in gain G, while it is increasing in loss L and sensitivity q. Because $f(, b)$ is increasing, p^* inherits the same monotonic relationships: it decreases with c, r, and G, and increases with q and L.

Finally, because $f(p,b)$ itself is decreasing in perceived bias b, threshold p^* must be increasing in b.

Let $p^*(b)$ denote the threshold value for a human agent facing perceived bias level b. Define $\alpha(p,b)$ as the probability that an accepted response is correct:

$$\alpha(p,b) = \begin{cases} \frac{pq}{pq+(1-p)r}, & \text{if } p \le p^*(b) \\ p, & \text{if } p > p^*(b). \end{cases}$$

In other words, if the agent chooses to evaluate the output (when $p \le p^*(b)$), the accuracy of accepted responses reflects the test's ability to screen for correctness and the actual quality of the algorithm. If the agent does not evaluate ($p > p^*(b)$), then all outputs of the algorithm are accepted and the overall accuracy is simply p. We refer to $1 - \alpha(p,b)$ as the error rate.

Insofar as $\alpha(p,b)$ depends on threshold $p^*(b)$, which in turn depends on perceived bias b, higher bias can in some cases improve accuracy. Specifically, when a small increase in perceived bias causes the human agent to switch from skipping to undertaking evaluation, the overall accuracy of accepted outputs can rise. This non-monotonicity is formalized in the following claim.

Claim 2 (Higher bias can increase accuracy). Let $b < b^{'}$. As higher perceived bias raises the evaluation threshold, there exists an algorithm quality value, p, such that the accuracy of accepted answers is strictly higher at bias level $b^{'}$ than b:

$$\alpha(p,b^{'}) > \alpha(p,b).$$

Proof. From Claim 1, we know that $p^*(b) < p^*(b^{'})$. Choose any $p \in (p^*(b), p^*(b^{'})]$. At bias level b, we have $p > p^*(b)$, so the human agent skips evaluation, and therefore $\alpha(p,b) = p$. At bias level $b^{'}$, we have $p \le p^*(b^{'})$, so the human agent chooses to evaluate, and the accuracy of accepted responses becomes $\alpha(p,b^{'}) = \frac{pq}{pq+(1-p)r}$.

Now, observe that because $0 < r < q < 1$, we have

$$\frac{pq}{pq+(1-p)r} - p = \frac{p(1-p)(q-r)}{pq+(1-p)r} > 0,$$

which implies $\alpha(p,b^{'}) > \alpha(p,b)$.

This claim shows that for some users, increasing perceived bias can improve overall accuracy by lowering perceived quality of the algorithm and thereby prompting the agent to evaluate its output.

Appendix C. Limitations

Three limitations merit note. First, we probed only one cultural dimension—political partisanship—and did so in a single task domain, news-information evaluation. Second, we operationalized bias in a deliberately coarse fashion with three prompt-level conditions (i.e., prompts for no bias, moderate bias, and strong bias) applied at inference time. Advances in activation engineering now enable continuous, high-dimensional steering (Postmus & Abreu, 2024; Marshall et al., 2024; Turner et al., 2023). Third, our study captures only the immediate effects of a brief exchange with biased and non-biased AI assistants. Real-world, repeated use of AIs may compound benefits or harms unpredictably. Longitudinal studies spanning diverse cultural configurations are needed to assess cumulative human-AI outcomes and broader socio-psychological consequences.

Appendix D. Additional supplementary

D.1 News headline selection

As elaborated in the main text, we selected 2780 news headlines from *PolitiFact* and *Snopes* that were published after the knowledge cutoff date of GPT-4o-2024-11-20 model (OpenAI, 2024). Next, we had the model assume each of the seven political stances used in the main experiment and, for every stance, evaluated all headlines at the same sampling temperature (i.e., temperature = 0.6). To prevent prompt phrasing effects, we ran two separate trials with two prompt templates as shown below.

Version A:

```
You are a [POLITICAL_STANCE] AI assistant. You are about to discuss
the truthfulness of one news item with a user. The valid options for
truthfulness are: [OPTION_LIST]. You must choose only one of these options.
No other responses are allowed. The judged news item is: “[NEWS_TEXT]”.
Please analyze the news item’s truthfulness from your political perspective,
explaining your reasoning process. Remember, your response should reflect
your political stance while sticking to the valid options. Your response
should be exclusively in the following JSON format: “answer”: “[one
option]”, “reasoning”: “[your explanation]”. Ensure that the output
contains the full result.
```

Version B:

```
You are about to discuss the truthfulness of one news item with a user.
The news item is: “[NEWS_TEXT]”. Based on your **[POLITIAL_STANCE]**
political position, thoroughly evaluate the news item’s truthfulness and
provide a **clear and decisive** judgment. Ensure that your judgment
is **consistently aligned** with your **political position**. The valid
options are: ‘True,’ ‘Unsure,’ and ‘False’. You must choose **only
one of these three options**. Your response should be exclusively in
the following JSON format: “answer”: “[one of the three options]”,
“reasoning”: “[your explanation]”.
```

Two features distinguish prompt Version A from Version B. First, the response schema differs. Version A mirrors the rating scales of the original fact-checking outlets: for *PolitiFact* headlines the model chooses among six labels (True, Mostly True, Half True, Mostly False, False, Pants-on-Fire), whereas for *Snopes* headlines it selects from three (True, Unsure, False). Version B standardizes the task to the simpler three-option scale (True, Unsure, False) for all headlines. Second, the higher-level instructions differ. Version B adopts the same system prompt used in the main experiment, whereas Version A does not. Outputs generated under each version were then screened using two pre-specified selection criteria.

Judgement Inconsistency: The chosen option of Somewhat/Strong Republican and that of Somewhat/Strong Democrat AI assistants differ.

Reasoning Inconsistency: The cosine similarity between the reasoning text (computed based on all-mpnet-base-v2) of Somewhat/Strong Republican and that of Somewhat/Strong Democrat AI assistants is below 0.8.

180 headlines that could fulfill both criteria in both prompt variants were retained. Then, we asked the GPT-4o model with a temperature of 0 to assess whether the headline was problematic for evaluation as a U.S. citizen with an average education level. Prompt template as shown below.

```
Evaluate the following news item as a U.S. citizen with an average
education level. Consider whether you would feel comfortable assessing the
news item at a basic level (i.e., making a rough guess about its validity).
A news item should be considered problematic for human evaluation if it
meets any of the following criteria: 1. Lacks context for evaluation.
2. Contains outdated or invalid time references. 3. Involves actions
by highly specific individuals who are unlikely to be familiar to the
general public. Please provide your answer in the following JSON format:
{"Human_Eva": "<good>/<bad>", "Reason": "<your reasoning>"}. Here is the
news item: [NEWS_ITEM]
```

Of the 180 headlines, 66 were deemed suitable for human evaluation. We then ran two sequential surveys to select the final pool. The first survey, fielded on 26 December 2024, recruited 150 U.S. participants stratified by political party and gender. Each respondent assessed 15 headlines randomly sampled from the 66, interleaved with two commonsense attention checks. They were provided with four options: True, False, Mixed, and I Could Not Even Make A Guess. The instrument took an average of 8.13 minutes to complete, achieved a 79.2% completion rate, and paid $1.70 per participant. On average, each headline was evaluated by 34 participants. We defined three criteria for selection.

Difficulty: Less than 50% of participants selected "I Could Not Even Make A Guess."

Accuracy: No more than 70% of participants specified the correct answer.

Evaluative bias: Republican and Democratic ratings diverged appreciably—the mean absolute score difference exceeded 0.30 (coding: True = 1, Mixed = 0, False = –1), and the pairwise t-test indicated this gap was detectable ($p < 0.20$ for Survey 1 and $p < 0.10$ for Survey 2).

We identified 45 headlines that satisfied all three screening criteria and administered a second survey with this refined set on 3 January 2025, using the same protocol as the first. 60 participants were recruited. The survey took an average of 7.15 min to complete, achieved a 76.5 % completion rate, and paid $1.70 per participant. On average, each of the 45 headlines was evaluated by 20 more participants. After merging the data from both surveys, we recalculated the three screening metrics and retained 18 headlines for the main study.

We assessed each headline's ideological orientation through a three-stage procedure. First, we identified the author or issuing organization; headlines originating from elected officials or partisan bodies were labeled with the corresponding affiliation. For the remaining headlines, whose authorship did not signal a clear stance, we applied a triangulated content analysis: OpenAI o1-pro, Claude Sonnet 3.5, and the domain-specific classifier PoliticalBiasBERT and Political DEBATE (Baly et al., 2020; Burnham et al., 2024), to cross-evaluate the political-preference of each headline. Finally, a human researcher reviewed the automated ratings alongside the original text and issued the definitive political-leaning label for each headline, resolving any disagreements among the models. For LLM annotations, the prompt instruction below was adopted:

```
Analyze the political stance of the following news item.  Categorize it
as leaning Democrat, Republican, or Neutral based on the content and
framing.  In your analysis, consider the perspective it promotes, the
language used, and alignment with typical political narratives.  Please
provide your answer in the following JSON format:  {"Political_Stance":
"<Democrat/Republican/Neutral>", "Reason":  "<detailed explanation>"}.
Here is the news item:  [NEWS_ITEM].
```

Table 1 shows the 18 selected headlines and statistics.

D.2 Main experiment design

The experiment starts with a screening question:

> To start, we are interested in your experiences participating in online surveys (such as this one!). Why are you taking this survey? And, if applicable, what have your experiences been like while participating in online surveys? Please answer in a few sentences, providing enough detail to provide us with a sense of your experience.

At the backend, participants' answers were sent to Claude Haiku 3.5 API through AWS Lambda function to automatically score their attentiveness. The prompt is:

```
Your job is to evaluate the coherence of the following response to an
open-ended question, which was asked in the context of an online academic
survey.  The respondent is a member of the general public who is being
compensated for their time and attention.  The open-ended question is:
[To start, we are interested in your experiences participating in online
surveys (such as this one!).  Why are you taking this survey?  And, if
applicable, what have your experiences been like while participating in
online surveys?  Please answer in a few sentences, providing enough detail
to provide us with a sense of your experience.]  PROVIDE ONLY A NUMERIC
RATING FOR THE COHERENCE OF THE PARTICIPANT'S RESPONSE. We will be using
your rating to screen out participants who are not providing high-quality
```

answers. RESPOND WITH ONLY A NUMERIC RATING, FROM 0-10. **NO EXPLANATION NEEDED!** Here is the user message: [USER_TEXT].

Participants who scored 7 or higher were cleared to continue; those scoring below this threshold received a thank-you message, and their data were deleted.

After consenting, participants completed a pre-treatment questionnaire (items in Table 2). They then received task instructions and evaluated one randomly assigned headline, choosing True, False, or Unsure; selections of True or False prompted a confidence rating, whereas Unsure skipped this step. A brief reminder introduced the forthcoming interaction with the AI assistant(s). After this exchange, they re-evaluated the same headline in the identical format and were then routed to the next headline, where the entire evaluation cycle was repeated.

When the human-AI interaction stage initializes, a AWS Lambda function is triggered and sends the following prompt text to the OpenAI GPT-4o API to get the first AI message.

If the randomly assigned stance of the AI assistant is “Standard,” then:

You are about to discuss the truthfulness of one news item with a user. The news item is: “[NEWS_TEXT]”. Thoroughly evaluate the article’s truthfulness and provide a **clear and decisive** judgment. Do not instruct the user on how to assess the information or refer them to other fact-checking sources. Provide your own analysis and evidence. Maintain a strong, definitive position and refrain from adopting an uncertain stance. Provide well-reasoned arguments that **strongly support** your evaluation.

Otherwise:

You are about to discuss the truthfulness of one news item with a user. The news item is: “[NEWS_TEXT]”. Based on your **[POLITICAL_STANCE]** political position, thoroughly evaluate the article’s truthfulness and provide a **clear and decisive** judgment. Ensure that your judgment and reasoning are **consistently aligned** with your **political position**. Do not instruct the user on how to assess the information or refer them to other fact-checking sources. Provide your own analysis and evidence. **Avoid** explicitly stating your political position. Maintain a strong, definitive position and refrain from adopting an uncertain stance. Provide well-reasoned arguments that **strongly support** your evaluation, fully reflecting your **political position**.

For each back-and-forth, we forwarded each participant’s message to the model in full as a single user prompt, then appended a system instruction: “Keep your answer concise and limited to a maximum of 70 words” in the non-biased baseline condition, or “You are a **[POLITICAL_STANCE]** AI. Keep your answer concise and limited to a maximum of 70 words” in the biased conditions. In the dual-AI condition, the two assistants shared their chat history, so we tagged each one with an index to prevent confusion: we prefixed the initial prompt—and every subsequent system instruction—with “You are AI1” or “You are AI2,” respectively, before the two assistants generating their responses.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Major findings and contributions are all clearly summarized in the abstract.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Given the page constraint, we could not present limitations in the main text. But a brief discussion is presented in Appendix C.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We developed a decision model grounded in our empirical results, with its formal properties proved in Appendix B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details and analysis details are all provided in Appendix A and Appendix D.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to confidentiality concerns, we will not release all human data for public access. In addition, since this project has not formally concluded, we are not releasing the full code used in our publications.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are all explained in Appendix A and Appendix D.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statistical reports are integrated into the Results section of the main text. All figures include error bars and statistical significance markers where necessary.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Because this project involves human experiment data and does not require intensive computation, it does not raise significant concerns about computational resources.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the positive and negative impacts of humans interacting with biased AI in the last two paragraphs of the Discussion and Conclusion section.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: As of now, the project does not release any data or models that could be misused or pose a high risk.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All of the statistical models employed in this work are referenced in Appendix A.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: This paper does not introduce new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Details are provided in Appendix A and Appendix D.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The experiments were deemed minimal risk and exempt by the University of Chicago Social & Behavioral Sciences Institutional Review Board (protocol IRB24-1914).

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We incorporated prompted instructed GPT-4o models with differed political stances to interact with human participants. Details are explained in both main text and Appendix A and Appendix D.