PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

AI assists adversarial collaboration in debate on minority salience

Barbara Mellers[a,1], Leo Yuan[b], Yubo Zhou[b], Isabelle Mauboussin[b], Eyana C. Lao[b], Bea Corio[b], Ville Satopaa[c], Lyle Ungar[d], Sudeep Bhatia[b], Cory J. Clark[e], Rasha Kardosh[f], Ran Hassin[g], Asael Sklar[h], Surya Gayet[i], Chris Paffen[i], Stefan Van der Stigchel[i], Andre Sahakian[i], and Philip Tetlock[j]

Affiliations are included on p. 5.

Edited by Steven Pinker, Harvard University, Cambridge, MA; received December 4, 2025; accepted April 23, 2026

The advancement of science depends on rigorous tests of competing hypotheses, yet many disputes are left unresolved. Adversarial collaboration—where opposing scientists jointly design decisive tests—is one proposed solution. We examine whether large language models (LLMs) can play a role by organizing information, structuring the debate and generating candidate experimental designs. This article reports an AI-assisted adversarial collaboration designed to resolve a debate in PNAS on minority salience—an overestimation of the percentage of minority faces in a visual display. The debate focused on whether there would be further overestimation when minorities in the displays were the same minorities in participants' communities (or social environments). Using LLMs to extract and organize competing propositions, we identified central disagreements and generated initial experimental designs to test claims. Human collaborators refined the designs and created two preregistered experiments that factorially manipulated the ethnicity of minority faces and the ethnicity of participants' communities. Data showed that people exaggerated the percentage of minorities in facial displays. Furthermore, overestimation was even greater when minorities in facial displays were also minorities in participants' communities. When the two camps of researchers saw the results, their confidence in key hypotheses converged. We do not experimentally test AI-assisted adversarial collaboration relative to traditional adversarial collaboration or other forms of dispute resolution. Rather, our study illustrates how an AI tool can be used with adversarial collaboration to formalize claims, structure disagreements, lower barriers to collaboration, and serve as an impartial observer to strengthen perceptions of fairness.

adversarial collaboration | AI | minority salience | scientific disputes

Significance

Disputes are central to scientific progress, but their resolution can be difficult. Adversarial collaboration offers a powerful but underexplored method for addressing scientific debate. We offer proof of concept that an AI-assisted adversarial collaboration can help resolve a debate about minority salience. This exercise led to preregistered experiments that clarified the dispute and produced solid evidence in one direction. Furthermore, after the experiments, the beliefs of opposing research teams about the hypotheses converged. AI did not replace humans or determine final experimental designs, but it clarified the disagreements and set the stage for the design of experiments that yielded sufficiently compelling results to produce agreement between proponents and skeptics of the robustness of minority-salience effects.

Scientific debates often unfold as sequences of claims and counterclaims. Although this form of dispute may sharpen arguments, it rarely settles them. One approach to resolving such disputes is adversarial collaboration, in which disagreeing scientists work together to codesign tests that could, in principle, change minds (1–5). Adversarial collaborations have clarified disagreements in judgment and decision-making, cognition, and social psychology (6–13). When successful, adversarial collaborations can surface hidden assumptions, reduce misunderstandings, and shift attention from rhetoric to evidence. Yet adversarial collaborations can be difficult to initiate and hard to sustain. Such collaborations often require extensive negotiations over details of experimental designs and interpretations of data. The process can be slowed by interpersonal friction, closed-mindedness, and weak incentives to cooperate.

These barriers may explain why adversarial collaborations remain relatively uncommon despite their benefits. Could adversarial collaborations become more common within science if they were easier to do? This question motivated our research agenda. Can large LLMs support adversarial collaboration by structuring contested propositions and facilitating progress toward mutually acceptable tests, even when they do not replace the theoretical, empirical, and methodological contributions of domain experts?

We use AI to i) extract and formalize key claims from each side as testable propositions, ii) identify points of agreement and disagreement, iii) propose empirical tests that can serve as starting points for discussion and iv) be a neutral third party that favors neither side. AI suggestions become starting points, and final experimental designs are improved and agreed upon by collaborating researchers.

We apply our tool to a debate in PNAS about minority salience—the tendency to overestimate the proportion of minorities in visual displays of faces. Kardosh et al. reported that observers overestimated minorities in facial arrays because of the tendency to pay

Author contributions: B.M., L.Y., Y.Z., I.M., E.C.L., B.C., V.S., L.U., S.B., C.J.C., R.K., R.H., A. Sklar, S.G., C.P., S.V.d.S., A. Sahakian, and P.T. designed research; B.M., L.Y., Y.Z., I.M., E.C.L., B.C., R.K., R.H., A. Sklar, S.G., C.P., S.V.d.S., and A. Sahakian performed research; B.M., L.Y., and Y.Z. analyzed data; and B.M., L.Y., Y.Z., and P.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mellers@wharton.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2535311123/-/DCSupplemental.

Published May 26, 2026.

attention to people and things that are less common and unexpected. When minorities in the displays are also the minorities in observers' communities, that minority group becomes even more salient, and participants pay even more attention to group members (14). Gayet et al. challenged their interpretation of the data, arguing that greater attention to minorities could be due to other processes, such as differences in visual characteristics of facial displays when minorities in the display were matched (vs mismatched) with minorities in participants' communities (15). Kardosh et al. rebutted the claims (16).

AI-Assisted Adversarial Collaboration

Both sides agreed that minority overestimation occurs. The central disagreement was whether the estimation of minorities further increased when minorities in the display were also the minorities in participants' communities. The two sides agreed to an AI-assisted adversarial collaboration. AI would take the role of an unbiased facilitator with the task of structuring claims, proposing studies and identifying paths toward resolution. We prompted a large language model (LLM) to read papers on both sides and extract key arguments as clear, concise, and testable propositions.

We then prompted the LLM to organize propositions into areas of agreement and disagreement and to highlight the most important points of conflict. Next, we asked the LLM to design an experiment to adjudicate critical conflicting propositions. This LLM pipeline (GPT-4o and Claude; 2024–2025) could, in principle, be used to assist scientific dispute resolution in a wide range of social science domains.

We did not assume the AI-generated experiment would necessarily be the best, but it served as a reasonable starting point for discussion. Both teams thought the LLM's experiments were theoretically sensible but insufficient to resolve the debate, underscoring the need for human expertise.

Eventually all parties agreed to an experimental design that addressed the core disagreement. We conducted two preregistered experiments that orthogonally manipulated i) the ethnicity of minority faces in laboratory displays and ii) the ethnicity of minorities in participants' communities. This design permitted the same set of facial stimuli to serve as minorities and majorities across participants, addressing the possibility that minority overestimation depended on differences in visual characteristics of facial displays rather than the match of minorities in the display to minorities in participants' communities.

Results

Both experiments were preregistered. Participants from China and the United States were recruited for the first experiment, and participants from South Africa and the United States were recruited for the second. All participants viewed 20 displays of 100 faces containing ~25% minorities and estimated the percentage of minority faces in the displays. In both experiments, we tested the hypotheses that participants overestimated minority proportions in visual displays and that minority overestimation became even greater when minorities in the facial displays matched minorities in participants' communities (Fig. 1).

Minority Overestimation. In both experiments, participants overestimated the actual proportion of minorities which was 25%. See Figs. 2 and 3. Mean estimates exceeded 25% by approximately 5.5% in each experiment ($t(457) = 5.62$, $P < 0.001$ in the Asian–Caucasian experiment and $t(431) = 10.81$, $P < 0.001$ in the South African–Caucasian experiment). Results provided convergent support for minority overestimation across two sets of populations and facial stimuli.

Fig. 1. Example Display of Asian minorities in the Asian–Caucasian Experiment. Each matrix display was randomly generated for each subject in each experiment. Similar matrices were used with Black and White faces in the South African–Caucasian experiment. The matrix was displayed for 2 s. Percentages of Asian faces averaged 25% and were randomly distributed.

Matching Minorities in Displays and Communities. Next we tested the hypothesis that minority overestimation increased when minorities in the displays were also minorities in participants' communities. Individual contrasts showed some violations of the hypothesis. Caucasians from the United States did not overestimate Asian minorities in facial displays relative to Asians from China. Similarly, Caucasians from the United States did not overestimate Black minorities in facial displays compared to Black participants from South Africa. However, a preregistered test comparing matched minorities versus mismatched minorities revealed a consistent pattern

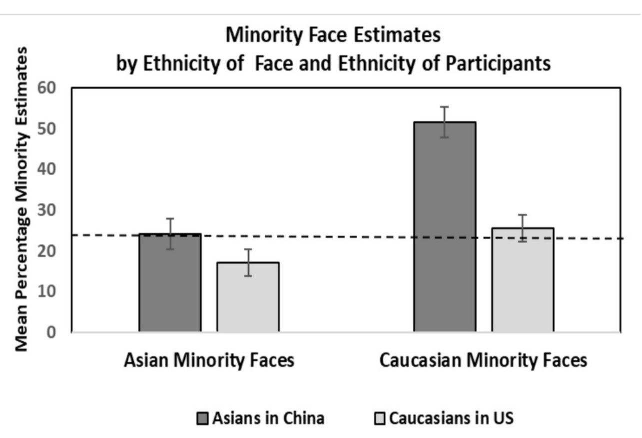

Fig. 2. Estimates of Minority Faces in the Asian–Caucasian Study. Mean estimates of minority faces in the Asian–Caucasian experiment when actual percentages were 25% (dashed line). On the *Right*, Asians in China overestimate Caucasian minority faces more than Caucasians in the United States. However, on the *Left*, Caucasians in the United States do not overestimate Asian minority faces more than Asians in China. Neither group of participants overestimates Asian minority faces. Error bars represent plus or minus 2 SE.

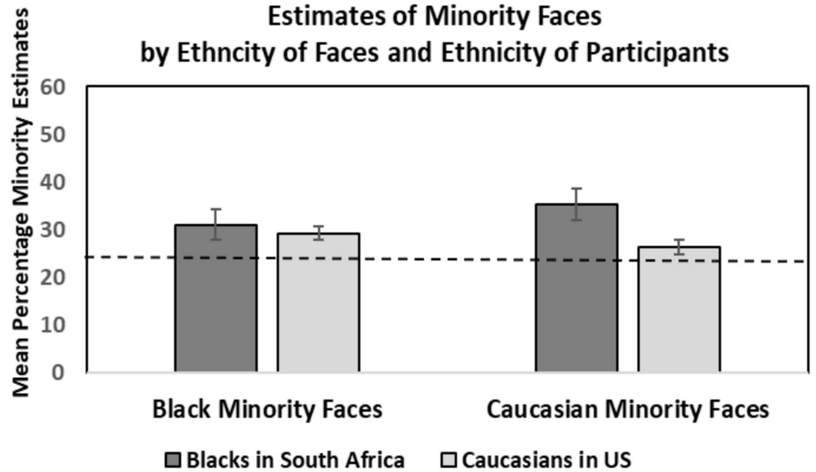

Fig. 3. Estimates of Minority Faces in the South African–Caucasian Study. Mean estimates of minority faces in the South African–Caucasian experiment when actual percentages were 25% (dashed line). On the *Right*, Blacks in South Africa overestimate Caucasian minority faces more than Caucasians in the United States. However, on the *Right*, Caucasians in the United States do not overestimate Black minority faces more than Blacks in the United States. Error bars represent plus or minus 2 SE.

in both experiments (Fig. 4). In the Asian–Caucasian experiment, minority overestimation was 9.5% higher if minorities in facial displays matched (versus mismatched) the minorities in participants' communities (t(454) = 5.01, $P < 0.001$). In the South African–Caucasian experiment, minority overestimation was 3.6% higher if minorities in displays matched (versus mismatched) minorities in participants' communities ($t(430) = 2.53$, $P < 0.01$). Thus, in two independent experiments, minority overestimation was reliably greater when the minority group in the display corresponded to the minority group in participants' communities.

Belief Updating and Convergence. We investigated the value of AI-assisted adversarial collaboration by assessing the confidence that both teams placed in the hypotheses before and after sharing the results. Both sides had relatively high prior confidence in minority overestimation (90% vs. 72%). That confidence increased modestly after exposure to the data (95% vs. 75%). Prior confidence in the hypothesis about additional overestimation when minorities in displays matched minorities in participants' communities differed substantially across sides (65% vs. 41%) but posterior confidence converged following exposure to the results (80% vs. 74%). Beliefs in key hypotheses on both sides increased, suggesting stronger scientific consensus after the collaboration than before.

Discussion

This study pursued two goals. The first was substantive: to investigate competing claims in an ongoing debate about minority salience and determine whether overestimation of minority groups in visual displays is amplified when the minority group in the display is also a minority in participants' social environment. The second was procedural: to document a case in which LLMs were incorporated into an adversarial collaboration to structure disagreements and generate candidate research designs to resolve conflicting claims.

The experiments provided evidence consistent with the hypothesis that minority overestimation increased when minorities in displays corresponded to minorities in participants' communities. The experiments addressed Gayet et al.'s concerns by independently manipulating the ethnicity of minority faces and ethnicity of minorities in participants' communities. We could use the same

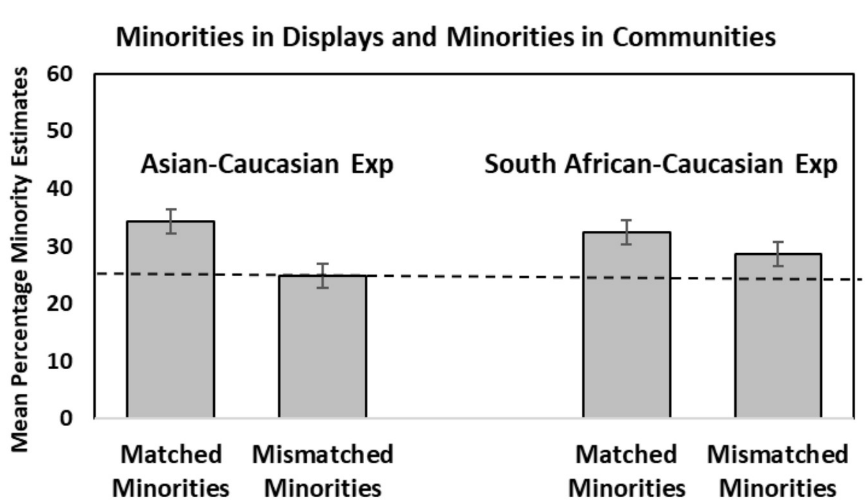

Fig. 4. More Overestimation If Minorities Match in Displays and Environments. Mean estimates of percentage of minority faces when actual percentages were 25% (dashed line). The *Left* side shows results from the Asian–Caucasian experiment, and the *Right* side shows results from the South African–Caucasian experiment. In both experiments, overestimation is greater when minorities in the display correspond to the minorities in participants' communities. Error bars represent plus or minus 2 SE.

faces as minorities and majorities within each experiment. Results were consistent in both experiments.

We also devised a method to evaluate our efforts by measuring belief change, a feature that was not present in previous adversarial collaborations but should be added. Beliefs converged after exposure to the results of codesigned experiments, indicating a greater degree of scientific consensus. These measures reveal the effectiveness of adversarial collaboration.

AI-assisted adversarial collaboration may be especially useful when issues are complex, data are ambiguous and working relationships are strained. Even without these challenges, AI can organize information into essential claims and present disagreements in a simple, shared language. It can analyze data and act as an unbiased third party, strengthening perceptions of fairness. Finally, it can provide access to researchers with fewer resources or social connections who wish to explore scientific disputes.

A solid test of AI-assisted adversarial collaboration vs traditional adversarial collaboration would require a much larger and systematic endeavor, such as that used in the many-labs study (17). Researchers would need to develop metrics of success beyond belief change in hypotheses. We recommend this undertaking as a valuable future direction. In the meantime, AI-assisted adversarial collaboration can help scientists by lowering barriers to entry, promoting productive debates and providing a deeper understanding of opposing theoretical and methodological perspectives (18).

Materials and Methods

All studies were approved by the University of Pennsylvania's Institutional Review Board, and all participants gave their consent to participate.

Preregistration Documents. Procedures, sample size, exclusion criteria, and data analysis plans were preregistered prior to data collection and can be obtained at https://aspredicted.org/7pq9bk.pdf. Study materials, datasets, and analysis code are publicly available on the Open Science Framework at https://osf.io/95g7z/overview?view_only=9511541179bd4af5844de728daa4b9d0.

Asian–Caucasian Experiment. We conducted a pilot study to identify unambiguous Asian and Caucasian faces. Fifty Chinese participants (Huixiang Data) and 50 Caucasian participants (Prolific) rated the degree to which individual male and female faces looked White or Asian on a 101-point scale (White to Asian). We needed a minimum of 40 validated faces for each ethnicity × gender category to avoid repeating faces within a matrix. We selected those faces for which the average rating of the correct ethnicity was 80 or above on a 101-point scale. This left us with 52 White female faces, 48 White male faces, 47 Asian female faces, and 41 Asian male faces, and we randomly selected 40 faces of each ethnicity x gender.

We preregistered a 2 × 2 factorial design crossing minority-face ethnicity (Asian vs. White) with participant population (Asians in China vs. Caucasians in the United States), with a target of $N = 400$ (200 per minority-face and participant ethnicity condition). Participants completed an online task and were excluded if they 1) failed to answer all major questions (progress <90), 2) failed an attention check, 3) were not members of the designated local majority group (living in areas with >50% majority and <50% minority residents), or 4) produced minority estimates ±2 SD from the group mean.

After initial exclusion, we retained 165 Asian participants and 190 Caucasian participants. To reach preregistered sample goals, we recruited additional participants, anticipating similar exclusion rates. The final sample included 252 Asians (China) and 206 Caucasians (United States). Participants from each country were randomly assigned to Asian-minority face or White-minority face conditions.

All participants viewed 20 matrices of 100 grayscale faces. Each matrix was displayed for 2 s. Minority faces were either Asian or White. The proportion of minority faces varied normally from 20% to 30% ($M = 25\%$). Participants were told, "You will quickly view sets of 100 faces for 2 s. After each trial, estimate whether there were more males or females in the set." Following all trials, participants estimated the percentage of minority faces across the trials (wording changed for each minority condition: Asian or Caucasian).

To verify local demographic context, participants reported a) whether their current neighborhood had relatively more Asian vs. White residents (5-point scale) and b) the perceived percentage of Asian and Caucasian residents in their neighborhood.

South African–Caucasian Experiment. Design, procedure, and exclusion criteria matched the Asian–Caucasian experiment unless otherwise noted. Fifty Black participants in South Africa and 50 Caucasian participants (Prolific, United States) rated the extent to which faces appeared Black or White. Faces were retained if average ratings were ≥90 that the face belonged to the correct ethnicity, yielding 59 Black female, 48 Black male, 58 White female, and 50 White male faces.

We preregistered $N = 400$ (200 per minority-face condition). After exclusions identical to the Asian–Caucasian experiment, we initially retained 117 Black participants (South Africa) and 143 Caucasian participants (United States). We then recruited additional participants; the final sample had 235 Blacks (South Africa) and 197 Caucasians (United States).

AI Prompts and Protocol. We began with the goal of identifying potential debates that could be used in an AI-assisted adversarial collaboration. We examined "Reply" sections of journals that included disagreements about published articles. For PNAS, we searched based on Reply, then Letter and Reply, then Psychological and Cognitive Sciences. For Nature Human Behavior, we searched based on Reply, then Psychology. For JPSP, we searched based on Reply, then the year 2000 or later.

We downloaded relevant articles. When the debate consisted of two articles (one on each side), we uploaded one article at a time for each run of Step 1. In the case of more than two articles, we uploaded articles on each side and asked for propositions to describe claims on both sides. In each step, we asked for specific details and thorough justifications of choices. Below are the prompts.

Step 1: "Read the paper and extract its key arguments. Summarize these as clear, concise, and testable propositions, each capturing a single idea or claim by the authors."

Step 2: "Organize the papers into two opposing sides based on their stances. List the propositions both sides agree upon, highlighting common ground for designing a fair experiment."

Step 3: "List the most critical conflicting propositions between the two sides, emphasizing those with significant implications for research design (e.g., methodology differences, variables considered, interpretation of results). Explain why these conflicts are pivotal to the dispute."

Step 4: "Develop a comprehensive experimental design to test the most significant conflicting proposition, capable of determining which side's claim is more accurate. Address the research-design concerns raised by both sides. Provide detailed examples and justifications for each of the following steps:

a. Describe the specific task or procedure for participants. Explain how it distinguishes between conflicting propositions and why it is the most effective choice.
b. Identify the variables you will manipulate, specifying conditions or levels for each. Justify how manipulating these variables tests the conflicting propositions.
c. Detail the variables you will measure, including how you will operationalize and quantify them. Explain why these variables are critical for assessing the conflicting propositions.
d. Identify potential confounding variables and describe strategies to control or eliminate their influence.
e. Discuss how your design addresses methodological concerns from both sides, such as sample size, randomization, blinding, and validity issues.
f. Explain the statistical methods you will use to analyze the data. Justify how these methods are appropriate for testing the conflicting propositions.
g. Describe possible results and explain how each outcome would support one side over the other. Illustrate how the experiment could conclusively resolve the dispute.

Prior Beliefs Survey. In this survey, we asked authors from both sides to "Consider the main hypotheses as follows: 1) Using statistical, perceptual, and social information the environment, people overestimate the relative size of minority groups and 2) When social information is consistent with statistical information, minority overestimation will increase relative to that driven by statistical information alone. For example, both White and Black participants from the United States tend to overestimate the percentage of Black minorities."

"Next, please tell us how confident you are in these hypotheses. To answer the confidence questions, use any integers you wish from 0 to 100, inclusive, (100 = I'm absolutely certain that it is true, 50 = I'm just as certain that it is true as I am certain that it is false, and 0 = I'm absolutely certain that it is false). How confident are you in Hypothesis 1? How confident are you in Hypothesis 2?" Finally, we asked, "What evidence would make you believe your hypotheses were false? (Open-ended)".

Posterior Beliefs Survey. This survey asked both sides, "Now that you have seen the data and results from our study, please answer the following questions: How confident are you in Hypothesis 1? How confident are you in Hypothesis 2? To answer the confidence questions, use any integers you wish from 0 to 100, inclusive (100 = I'm absolutely certain that it is true, 50 = I'm just as certain that it is true as I am certain that it is false and 0 = I'm absolutely certain that it is false". Finally, we said, "If you would like, please use the space below to briefly explain the reasoning behind your confidence ratings, share any additional comments about the data or results, or raise questions or concerns. This section is entirely optional but will help us better interpret your responses (Open-ended)".

Data, Materials, and Software Availability. Anonymized CSV files data have been deposited in OSF (https://osf.io/95g7z/overview?view_only=9511541179bd4af5844de728daa4b9d0 (19)). Study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We wish to thank the Wharton-INSEAD Alliance for providing financial assistance and the Wharton Behavioral Lab for conducting the experiments.

Author affiliations: [a]Psychology and Marketing Department, University of Pennsylvania, Philadelphia, PA 19104; [b]Psychology Department, University of Pennsylvania, Philadelphia, PA 19104; [c]Technology and Operations Management, Institut europeen d'administration des affaires, Fontainebleau Cedex 77305, France; [d]Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA 19104; [e]New College of Florida, Sarasota, FL 34243; [f]Psychology Department, New York University, New York, NY 10003; [g]Department of Psychology, The Hebrew University of Jerusalem, Jerusalem 9190500, Israel; [h]Arison School of Business, Reichman University, Herzliya 4610101, Israel; [i]Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht 3584 CS, The Netherlands; and [j]Management Department and Psychology Department, University of Pennsylvania, Philadelphia, PA 19104

1. D. Kahneman, G. Klein, Conditions for intuitive expertise: A failure to disagree. *Am. Psychol.* **64**, 515–526 (2009).
2. B. Mellers, R. Hertwig, D. Kahneman, Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychol. Sci.* **12**, 269–275 (2001).
3. T. Gilovich, V. Medvec, D. Kahneman, Varieties of regret: A debate and partial resolution. *Psychol. Rev.* **105**, 602–605 (1998).
4. I. Bateman, D. Kahneman, A. Munro, C. Starmer, R. Sugden, Testing competing models of loss aversion: An adversarial collaboration. *J. Public Econ.* **89**, 1561–1580 (2005).
5. M. A. Killingsworth, D. Kahneman, B. Mellers, Income and emotional well-being: A conflict resolved. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208661120 (2023).
6. S Rathje Martel *et al.*, On the efficacy of accuracy prompts across partisan lines: An adversarial collaboration. *Psychol Sci.* **35**, 437–454 (2024).
7. S. J. Ceci, S. Kahn, W. M. Williams, Exploring gender bias in six key domains of academic science: An adversarial collaboration. *Psychol. Sci. Public Interest* **24**, 15–73 (2023).
8. J. M. Doherty *et al.*, Dual-task costs in working memory: An adversarial collaboration. *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 1529–1551 (2019).
9. D. Matzke *et al.*, The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *J. Exp. Psychol. Gen.* **144**, e1–e15 (2015).
10. J. R. Corrigan, A. C. Drichoutis, J. L. Lusk, R. M. Nayga Jr., M. C. Rousu, Repeated rounds with price feedback in experimental auction valuation: An adversarial collaboration. *Am. J. Agric. Econ.* **94**, 97–115 (2012).
11. N. L. Kerr, X. Ao, M. A. Hogg, J. Zhang, Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *J. Exp. Soc. Psychol.* **78**, 66–76 (2018).
12. C. B. Cadsby, R. Croson, M. Marks, E. Maynes, Step return versus net reward in the voluntary provision of a threshold public good: An adversarial collaboration. *Public Choice* **135**, 277–289 (2008).
13. Cogitate Consortium *et al.*, Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature* **642**, 133 (2025).
14. R. Kardosh, A. Y. Sklar, A. Goldstein, Y. Pertzov, R. R. Hassin, Minority salience and the overestimation of individuals from minority groups in perception and memory. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2116884119 (2022).
15. S. Gayet, A. Sahakian, C. L. E. Paffen, S. Van der Stigchel, No evidence for social factors in the overestimation of individuals from minority groups. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2214740 (2022).
16. R. Kardosh, A. Y. Sklar, R. R. Hassin, Reply to Gayet et al.: Minority salience as a social and cognitive phenomenon. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2215179119 (2022).
17. R. Klein *et al.*, Investigating variation in replicability: A "Many Labs" replication project. *Soc. Psych.* **2**, a000178 (2014).
18. C. Isch, P. E. Tetlock, C. J. Clark, Reflections on adversarial collaboration from the adversaries: was it worth it? *Theory Soc.* **54**, 929–946 (2025).
19. B. Mellers *et al.*, AI-Assisted Adversarial Collaboration Accelerates Resolution of Debate on Minority Salience. Open Science Framework. https://osf.io/95g7z/overview?view_only=9511541179bd4af5844de728daa4b9d0. Deposited 17 November 2025.