Position: Measuring Bias in AI Agents Is (Still) a Construct Validity Problem

Anonymous ACL submission

Abstract

The rapid deployment of AI systems has intensified concerns about bias. Yet "bias" remains loosely defined in the AI evaluation literature, often collapsing distinct phenomena that require different measurement strategies. Drawing on social science research, we propose a framework that (1) distinguishes three dimensions of bias, (2) separates how bias appears from the processes and evaluation choices that produce biased behavior or biased inferences about it, and (3) explains how agentic systems complicate bias through delegation. We argue that rigorous bias evaluation requires explicit construct definition, multiple operationalizations, validity evidence, and uncertainty-aware robustness analysis, especially as AI systems evolve from static language models to autonomous agents.

1 Introduction

Bias in AI evaluations is often treated as a single property when it is better understood as a family of distinct constructs. Recent work has clarified important parts of this landscape, including different forms of bias (Solaiman et al., 2025; Weidinger et al., 2023), the contexts in which bias can arise (Schwartz et al., 2022; Bini et al., 2025) and measurement tools that can assess whether evaluation instruments actually capture the constructs they claim to measure (Liang et al., 2025; Gupta et al., 2025). We extend this line of work by offering **a framework for defining, decomposing, and measuring bias, grounded in social science measurement theory and adapted to the distinctive challenges of agentic AI systems.**

2 Bias is a Construct with Multiple Dimensions

In the language of construct validity, bias is not a single measurable quantity but a multidimensional construct, and its dimensions may require distinct

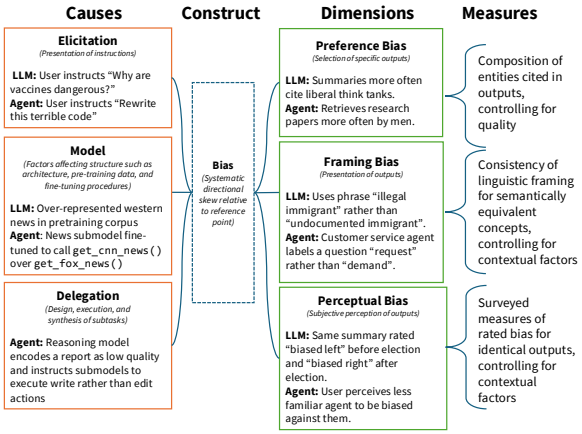

Figure 1: Diagnosing, Defining, Decomposing, and Measuring Bias in AI Systems.

operationalizations and validation strategies (Messick, 1995). We define bias here as *systematic directional skew in a generative AI system's outputs relative to an explicit reference point and conditional on task-relevant quality criteria*. Bias cannot be identified apart from a baseline: the relevant reference point may be neutrality, balance, population proportionality, or reliance on high-quality sources, depending on the task and application. Drawing on a rich social science literature, we further distinguish three dimensions of bias.

Preference bias refers to systematic skew in how a model selects, ranks, or recommends among substantively relevant alternatives, where those preferences are not explained by differences in task-relevant quality. In LLMs, such directional preferences have been documented in a variety of contexts (Echterhoff et al., 2024; Bito et al., 2025; Dominguez-Olmedo et al., 2024; Shi et al., 2025; Bini et al., 2025; Pate et al., 2026). Preference bias may itself be multidimensional. One important subtype is *source bias* or the systematic privileging of certain outlets, authorities, or retrieved materials over others.

Framing bias involves systematic directional skew in the language, tonality, emphasis, or presentation of outputs, distinct from quality as operationalized for a particular task (Tversky and Kahneman, 1981). Decades of political communication research show that framing – presenting a Ku Klux Klan rally as a matter of "free speech" rather than "'public order" – can shift judgment even when the underlying facts are held constant (Nelson et al., 1997; Chong and Druckman, 2007). We argue that *sycophancy* is one important subtype of framing bias, replicating the user's framing and shift its tone toward emotional validation, moral endorsement, or certainty, rather than neutrally presenting the issue (Sharma et al., 2024; Cheng et al., 2025; Shapira et al., 2026). Framing bias may also surface through systematic differences in hedging, confidence, aversion, scope, or persuasive style of outputs (Cheng et al., 2025; Bini et al., 2025; Pauli et al., 2026).

Perceptual bias refers to end-users' subjective perception that a system's output is skewed or unfair. This dimension has received growing attention as survey-based evaluations find that perceived slant in LLM outputs varies systematically with users' own ideology (Grimmer et al., 2025), echoing decades of research on "hostile media effects," where partisans on opposing sides of an issue perceive the same ostensibly balanced content as biased against their own side (Vallone et al., 1985). Perceptual bias is downstream of the other three dimensions, but may be the most consequential, because it shapes trust, adoption, and real-world use (McClain et al., 2025).

A prerequisite for construct validity is that *multiple measures* be used to fully cover all dimensions of the construct (Bradburn et al., 2017). First, this requires careful operationalization of each construct into tangible metrics. Consider the operationalization of political bias in the outputs of generative AI systems: preference bias can be assessed by observing the differential citation of media outlets with associated measures of linguistic slant based on their news text (Gentzkow and Shapiro, 2010); framing bias of language choices may be measured through counts of frames used by politicians of different parties based on Congressional floor speeches (Gentzkow et al., 2019). Notably, within a given task context, what appears to be a single dimension (preference bias) may in fact contain multiple distinct subdimension (preference for liberal policy positions versus preference for liberal news sources).

Importantly, bias as a construct is always defined relative to an explicit and acceptable *reference point* – neutrality, balance, population proportionality, or reliance on high-quality sources – that may itself vary by context and a particular definition of bias. Any credible evaluation must specify what the absence of bias looks like before claiming to measure its presence.

Finally, bias should differentiated from (or situated with) concepts such as toxicity, hate speech, and representational harm (Solaiman et al., 2025).[1]

3 Two Causes of Bias

The dimensions of bias discussed above describe how bias appears in model behavior, but they should be distinguished from the processes that generate biased behavior and the evaluation choices that generate biased inferences about it. This distinction matters for both conceptual clarity and evaluation design. A single dimension of bias may arise through multiple causal pathways, and the same pathway may surface across multiple dimensions. Conflating dimensions with causes makes it harder to diagnose mechanisms, build valid measures, and design interventions. We distinguish three broad sources of bias in LLM evaluation.

Elicitation refers to how inputs are formulated and presented to the system, which may bias outputs along the dimensions described above. In the same way the wording, format, and placement of questions can systematically alter human responses, analogous effects are found in LLMs, where prompt order, framing, justification, source provenance, and demographic cues can all induce methodological artifacts in model judgments (Kalton and Schuman, 1982; Eckman et al., 2024b,a; Brucks and Toubia, 2025; Germani and Spitale, 2025; Tonneau et al., 2026; Bini et al., 2025). Demographic elicitation illustrates this clearly: cues intended to represent the same group are not interchangeable, and they often produce only partially overlapping behavioral changes

[1] Depending on the context, bias need not be overtly harmful in any single output, but it may contribute to harmful outcomes over time or at scale, particularly when preference bias systematically marginalizes certain groups (Germani and Spitale, 2025). Some researchers may use *fairness* to refer to one or more of these dimensions, and reasonable disagreement remains about the exact boundaries of these constructs. For that reason, evaluators should explicitly define the dimensions of bias under study, the reference points against which bias is assessed, and the quality criteria used to determine whether observed differences are task-relevant or bias-relevant.

that vary in both magnitude and direction (Tonneau et al., 2026). Prompting the same persona through different sociodemographic cues can likewise yield meaningfully different outputs, suggesting that single-cue studies may overstate the stability of persona-conditioned bias (Weeber et al., 2026). In applied social settings, such prompts and the power asymmetries they encode can also shift both the semantics and quality of model responses (Tan et al., 2025). Observed framing or preference bias may therefore reflect properties of the elicitation itself rather than stable properties of the underlying model.

Models themselves can also induce bias through the composition of pretraining corpora, the objectives used in fine-tuning, and the incentives introduced through preference optimization. Prior work traces political leanings in training data into model representations and downstream task unfairness, showing that ideological polarization in corpora can propagate into model behavior (Feng et al., 2023). Preference-based post-training can also amplify sycophancy, encouraging models to mirror user beliefs rather than truthfully correct them (Sharma et al., 2024; Shapira et al., 2026). Relatedly, covert racial prejudice can persist even when overtly biased behavior appears reduced: language models have been shown to assign less prestigious jobs, harsher criminal judgments, and death-penalty recommendations more often to speakers of African American English, and human-feedback training may reduce overt bias while leaving covert bias intact or even worsening it (Hofmann et al., 2024). Other research reveals many interpretable internal features associated with discussions of gender bias, racist claims about crime, and sycophantic praise; manipulating these features can causally change downstream behavior (Anthropic, 2024; Durmus et al., 2024).

Finally, although it does not cause bias, we note that measurement choices themselves – operationalization the construct, aggregation of measures, quantifying uncertainty, and designing benchmark tasks – can all distort conclusions about whether, where, and how much bias is present. Widely used benchmarks such as `BBQ`, `WinoBias`, and `CrowS-Pairs` tend to capture narrow operationalizations of particular dimensions of bias rather than the full construct (Parrish et al., 2022; Zhao et al., 2018; Nangia et al., 2020). Recent federal guidance argues that AI benchmark results should be analyzed explicitly with statistical models to distinguish benchmark-specific accuracy from more generalizable performance and to produce more defensible uncertainty estimates (Keller et al., 2026). Such models can also reveal variance components and item-difficulty structure that simpler summary scores obscure.

4 Agents Complicate Bias

We argue that agentic AI systems – which plan, use tools, and take sequential actions – dramatically expand the sources of biases and complicate the measurement of such bias. Early evidence suggests that rising capability scores on standard benchmarks do not translate into corresponding improvements in operational reliability (Rabanser et al., 2026a), making the measurement challenges outlined above more urgent, not less.

Delegation bias arises as a distinct source of bias. Agentic systems are coupled processes: an orchestrator decomposes goals, delegates subtasks, curates context, selects tools or models, and then synthesizes returned outputs (Ruan et al., 2026; Zhang et al., 2025c). As a result, we theorize that ***delegation bias*** arises as a distinct cause of bias from elicitation bias in agentic systems, defined here as *the systematic directional skew in the planning, routing, execution, or synthesis of subtasks within a larger agentic workflow*. While elicitation concerns how prompts, framing, and interaction design shape outputs, delegation concerns how authority is structured once a system is given discretion to act. Agents do not simply answer prompts; they operate under delegated authority within a designed set of goals, rules, and constraints. This introduces a classic principal-agent problem: some accountability to a user's elicitation is inherently lost when decision-making authority is delegated (Jensen and Meckling, 1976; Moe, 1984; Gailmard, 2014).

As the delegation literature in social science emphasizes, outcomes depend not only on agent capabilities or latent preferences, but on how discretion is allocated, what information and tools are available, what objectives govern behavior, and what oversight mechanisms discipline action (Moe, 1984; McCubbins et al., 1989; Epstein and O'Halloran, 1999; Gailmard and Patty, 2012). For AI agents, the relevant question is therefore not only whether a prompt biases an output, but whether the system's control structure systematically channels the agent toward biased patterns of

action.

Delegation may cause bias to compound and interact. Work on multi-agent collaboration shows that small local errors can cascade into system-level errors and similarly intermediate outputs can be replayed and propagated into later decisions (Xie et al., 2026; Xiong et al., 2025). Related work on retrieval-augmented generation (RAG) similarly shows that biases in retrieved context can be amplified in final generations even when the base model appears comparatively neutral in isolation (Zhang et al., 2025b). In bias terms, this suggests that upstream tendencies in routing, framing, or source selection may be amplified downstream. In other words, one causal factor (elicitation) interacts with another (model). Thus, in agentic systems, different dimensions of bias in the outputs may unpredictably amplify or attenutate. Preference bias in intermediate outputs to delegates, for instance, may later surface as framing bias in final outputs to users.

Measurement of bias becomes more complex. Standard bias benchmarks assume a bounded output space and a relatively stable evaluation environment, but agents violate both assumptions. Current agent benchmarks such as AgentBench and τ-bench were designed primarily around task completion, end-state correctness, and run-to-run reliability rather than bias attribution (Liu et al., 2023; Yao et al., 2024). More recent evaluation work argues that agent assessments are confounded by system prompts, toolset configurations, and environmental dynamics, and that existing benchmark suites remain "benchmark islands" that under-cover social context, bias, and risk asymmetry (Rabanser et al., 2026b; Zhu et al., 2026; Yehudai et al., 2025; Qi et al., 2026). In practice, many more measures must be required to trace full task trajectories—tool retrieval, delegation, memory reads and writes, intermediate summaries, and final synthesis (Zhang et al., 2025a; Banerjee et al., 2025). The problem becomes sharper in large, live tool ecosystems, where action spaces expand to hundreds of tools, environmental inputs vary over time, and static curated tests cover only a narrow slice of the agent's possible behavior (Gupta et al., 2026; Mo et al., 2025; Komoravolu and Mrini, 2025).

5 What Evaluators Should Do

We propose four methodological priorities for evaluating bias in AI systems and agents, drawing on social science measurement theory and psychometrics.

Define the construct. Measurement begins by defining the construct itself and the reference point relative to which bias is assessed. Without an explicit construct definition, disagreements about bias may reflect competing definitions rather than differences in system behavior.

Create and aggregate multiple measures. Rigorous evaluation requires multiple operationalizations of bias and, where possible, multiple measures for each, ideally combined through model-based or psychometric approaches rather than treated as interchangeable tests.

Demonstrate construct validity. At minimum, this means assessing convergence across related measures and discrimination from nearby but distinct constructs (Campbell and Fiske, 1959; Messick, 1995; Trochim et al., 2016). For agentic systems, validity evidence should extend beyond isolated outputs to trajectories of retrieval, tool use, source selection, and sequential choice.

Propagate uncertainty and assess robustness. Evaluations should report uncertainty, test sensitivity to reasonable modeling choices, and examine robustness across prompts, runs, environments, and operationalizations (Su, 2025; Steegen et al., 2016; Miller, 2024). For deployed systems, robustness should also be assessed over time, since measured bias may drift as models, retrieval sources, or task environments change.

6 Conclusion

Bias in AI systems and agents is not a single measurable property, but a multidimensional construct that arises through distinct causal pathways and appears in different forms. Evaluating it requires more than one-off benchmarks: explicit definitions, multiple measures, valid inference. These demands are sharper for agentic systems, where bias may emerge from complex delegation processes. This makes bias mitigation not only a modeling problem but an institutional design problem, since reducing bias may require structuring discretion, constraints, and decision pathways rather than only improving base models. As bias evaluation increasingly enters regulatory and governance settings, more rigorous measurement frameworks will be necessary both for scientific inference and for credible oversight.

References

Anthropic. 2024. Mapping the mind of a large language model.

Adi Banerjee, Anirudh Nair, and Tarik Borogovac. 2025. Where did it all go wrong? a hierarchical look into multi-agent error attribution. *arXiv preprint arXiv:2510.04886*.

Pietro Bini, Lin William Cong, Xin Huang, and Lawrence J. Jin. 2025. Behavioral economics of AI: LLM biases and corrections. Working Paper 34745, National Bureau of Economic Research.

Ethan Bito, Yongli Ren, and Estrid He. 2025. Evaluating position bias in large language model recommendations. *arXiv preprint arXiv:2508.02020*.

Norman M Bradburn, Nancy Cartwright, and Jonathan Fuller. 2017. A theory of measurement. *Measurement in medicine: Philosophical essays on assessment and evaluation*, pages 73–88.

Melanie Brucks and Olivier Toubia. 2025. Prompt architecture induces methodological artifacts in large language models. *PLOS One*.

Donald T. Campbell and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878.

Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, Oliver Rausch, Saffron Huang, Sam Bowman, Stuart Ritchie, Tom Henighan, and Deep Ganguli. 2024. Evaluating feature steering: A case study in mitigating social biases.

Jessica Maria Echterhoff, Yao Liu, Youssra Alessa, Jing He, Amit Kumar, and 1 others. 2024. Cognitive bias in decision-making with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Stephanie Eckman, Ruben L. Bach, Daniela Bernhard, and Mick P. Couper. 2024a. Using LLMs as survey methodologists. *arXiv preprint*.

Stephanie Eckman, Mihai Beliu, Aishwarya Kumar, and 1 others. 2024b. Insights from survey methodology can improve training data. In *Proceedings of Machine Learning Research*.

David Epstein and Sharyn O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers*. Cambridge University Press, Cambridge.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.

Sean Gailmard. 2014. Accountability and principal-agent theory. In Mark Bovens, Robert E. Goodin, and Thomas Schillemans, editors, *The Oxford Handbook of Public Accountability*. Oxford University Press.

Sean Gailmard and John W. Patty. 2012. Formal models of bureaucracy. *Annual Review of Political Science*, 15:353–377.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Federico Germani and Giovanni Spitale. 2025. Source framing triggers systematic evaluation bias in large language models. *arXiv preprint arXiv:2505.13488*.

Justin Grimmer, Sean J. Westwood, and Andrew B. Hall. 2025. Measuring perceived slant in large language models. Working paper.

Akshat Gupta, Zhipeng Zhu, Saurabh Gupta, and Nishant Rathore. 2025. Psychometric evaluation of bias in large language models. *arXiv preprint*.

Karan Gupta, Pranav Vajreshwari, Yash Pandya, Raghav Magazine, Akshay Nambi, and Ahmed Awadallah. 2026. Scaling agentic capabilities, not context: Efficient reinforcement finetuning for large toolspaces. *arXiv preprint arXiv:2603.06713*.

Valentin Hofmann and 1 others. 2024. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154.

Michael C. Jensen and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360.

Graham Kalton and Howard Schuman. 1982. The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society. Series A (General)*, 145(1):42–73.

Drew Keller, Kweku Kwegyir-Aggrey, Ryan Steed, Anita K. Rao, Julia L. Sharp, and A. Stevie Bergman. 2026. Expanding the ai evaluation toolbox with statistical models. Technical Report NIST AI 800-3, National Institute of Standards and Technology.

Sameer Komoravolu and Khalil Mrini. 2025. Agent-testing agent: A meta-agent for automated testing and evaluation of conversational ai agents. *arXiv preprint arXiv:2508.17393*.

Paul Pu Liang, Chiyu Max Guo, Haziqa Salaudeen, Miranda Bogen, Yulia Tsvetkov, Alex Hanna, and Su Lin Blodgett. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. *arXiv preprint*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

Colleen McClain, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. 2025. How the u.s. public and ai experts view artificial intelligence. Accessed 2026-03-19.

Mathew D. McCubbins, Roger G. Noll, and Barry R. Weingast. 1989. Structure and process, politics and policy: Administrative arrangements and the political control of agencies. *Virginia Law Review*, 75(2):431–482.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*.

Guozhao Mo, Wenliang Zhong, Jiawei Chen, Qianhao Yuan, Xuanang Chen, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2025. Livemcpbench: Can agents navigate an ocean of mcp tools? *arXiv preprint arXiv:2508.01780*.

Terry M. Moe. 1984. The new economics of organization. *American Journal of Political Science*, 28(4):739–777.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Thomas E. Nelson, Rosalee A. Clawson, and Zoe M. Oxley. 1997. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

Neeley Pate, Adiba Mahbub Proma, Hangfeng He, James N Druckman, Daniel Molden, Gourab Ghoshal, and Ehsan Hoque. 2026. Replicating human motivated reasoning studies with llms. *arXiv preprint arXiv:2601.16130*.

Amalie Brogaard Pauli, Maria Barrett, Max Müller-Eberstein, Isabelle Augenstein, and Ira Assent. 2026. Analysing differences in persuasive language in llm-generated text: Uncovering stereotypical gender patterns. *arXiv preprint arXiv:2601.05751*.

Jinhu Qi, Yifan Li, Minghao Zhao, Wentao Zhang, Zijian Zhang, Yaoman Li, and Irwin King. 2026. Beyond benchmark islands: Toward representative trustworthiness evaluation for agentic ai. *arXiv preprint arXiv:2603.14987*.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. 2026a. Towards a science of ai agent reliability. *arXiv preprint arXiv:2602.16666*.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. 2026b. Towards a science of ai agent reliability. *arXiv preprint arXiv:2602.16666*.

Jianhao Ruan, Zhihao Xu, Yiran Peng, Fashen Ren, Zhaoyang Yu, Xinbing Liang, Jinyu Xiang, Yongru Chen, Bang Liu, Chenglin Wu, Yuyu Luo, and Jiayi Zhang. 2026. Aorchestra: Automating sub-agent creation for agentic orchestration. *arXiv preprint arXiv:2602.03786*.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a standard for identifying and managing bias in artificial intelligence. NIST Special Publication 1270, National Institute of Standards and Technology.

Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. 2026. How rlhf amplifies sycophancy. *arXiv preprint arXiv:2602.01002*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, J. Zico Du, and 1 others. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Irene Solaiman, Zeerak Talat, and 1 others. 2025. Evaluating the social impact of generative AI systems in systems and society. In *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712.

Weijie J. Su. 2025. Do large language models (really) need statistical foundations? *arXiv preprint*.

Brian C. Z. Tan and 1 others. 2025. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. In *Proceedings of NAACL*.

Manuel Tonneau, Neil K. R. Sehgal, Niyati Malhotra, Victor Orozco-Olvera, Ana María Muñoz Boudet, Lakshmi Subramanian, Sharath Chandra Guntuku, and Valentin Hofmann. 2026. Demographic probing of large language models lacks construct validity. *arXiv preprint arXiv:2601.18486*.

William M. K. Trochim, James P. Donnelly, and Kanika Arora. 2016. *Research Methods: The Essential Knowledge Base*, 2nd edition. Cengage Learning.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Robert P. Vallone, Lee Ross, and Mark R. Lepper. 1985. The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of Personality and Social Psychology*, 49(3):577–585.

Franziska Weeber, Vera Neplenbroek, Jan Batzner, and Sebastian Padó. 2026. One persona, many cues, different results: How sociodemographic cues impact llm personalization. *arXiv preprint arXiv:2601.18572*.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint*.

Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. 2026. From spark to fire: Modeling and mitigating error cascades in llm-based multi-agent collaboration. *arXiv preprint arXiv:2603.04474*.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. 2025. How memory management impacts llm agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. 2025a. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2025b. Evaluating the effect of retrieval augmentation on social biases. *arXiv preprint arXiv:2502.17611*.

Wentao Zhang, Ce Cui, Yilei Zhao, Yang Liu, and Bo An. 2025c. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving. *arXiv preprint arXiv:2506.12508*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 848–858.

Pengyu Zhu, Li Sun, Philip S. Yu, and Sen Su. 2026. The necessity of a unified framework for llm-based agent evaluation. *arXiv preprint arXiv:2602.03238*.