Position: Evaluating Bias in AI Agents Is (Still) a Social Science Measurement Challenge

Soubhik Barari
NORC at the University of Chicago
barari-soubhik@norc.org

Diwas Bhattarai
NORC at the University of Chicago
bhattarai-diwas@norc.org

Joshua Y. Lerner
NORC at the University of Chicago
lerner-joshua@norc.org

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 distinguishes three dimensions of bias (*preference*, *framing*, and *perception*) grounded in known typologies, separates the diagnosis of bias (*causes*) from how it appears (*dimensions*), and explains how *delegation* arises as a novel diagnosis for bias in AI agents. We illustrate the framework with a stylized example of an agentic meta-analysis, showing how discretion over search strategy, study inclusion, outcome prioritization, evidence weighting, and reporting language may shift conclusions toward different policy orientations. Rigorous AI bias evaluation in the agentic era remains a social science measurement challenge, requiring attention to construct validity, external validity, and an understanding of delegation in complex systems.

1 Introduction

AI bias is often treated as a single property when it is better understood as a context-specific, multidimensional construct. Prior work demonstrates that AI bias evaluations routinely deploy operationalizations poorly matched to the harms they target (Blodgett et al., 2020), that fairness harms in computational systems frequently trace to mismatches between theoretical constructs and their operationalizations (Jacobs and Wallach, 2021), and that AI evaluation practice tends to skip systematizing a concept before building instruments to measure it (Wallach et al., 2025). Related work has mapped broader social forms of bias (Solaiman et al., 2025; Weidinger et al., 2023), the contextual factors that produce them (Schwartz et al., 2022; Bini et al., 2025), and the psychometric validity of AI evaluation instruments (Wallach et al., 2025; Jung et al., 2026).

We extend Wallach et al. (2025)'s original provocation on AI evaluation as a social science measurement problem in two ways. First, we propose a three-dimensional decomposition of bias grounded in established social science typologies and adapted to agentic AI outputs. Second, whereas prior accounts focus on large language models (LLMs), we argue that *delegation* — the transfer of decision-making from a human principal to an autonomous agent — constitutes a structurally novel source of bias in AI agents. We ground these claims in a stylized agentic meta-analysis example showing how delegated methodological choices can produce distinct shifts in preference and framing bias, requiring evaluators to measure both rather than treating one as a sufficient proxy for the other.

In sum, we argue that **bias evaluation in agentic AI systems requires social science tools at every stage: conceptualizing and operationalizing the construct (construct validity), measuring across real-world contexts (external validity), and diagnosing causes (delegation).**

2 Bias is a Construct with Multiple Dimensions, Measures, and Contexts

As Figure 1 illustrates, bias is not a single measurable quantity but a multidimensional construct (Messick, 1995). Following Blodgett et al. (2020) and Wallach et al. (2025), we call the preliminary step *conceptualization*, or specifying the relevant form of bias under investigation, which may be context-dependent (e.g., having to do with political, gender, or racial entities) and scoped narrowly or broadly. We offer a general definition applicable across settings: *systematic skew in an agentic AI system's outputs relative to an explicit reference point, conditional on task-relevant quality criteria*. By *quality criteria*, we refer to legitimate task-

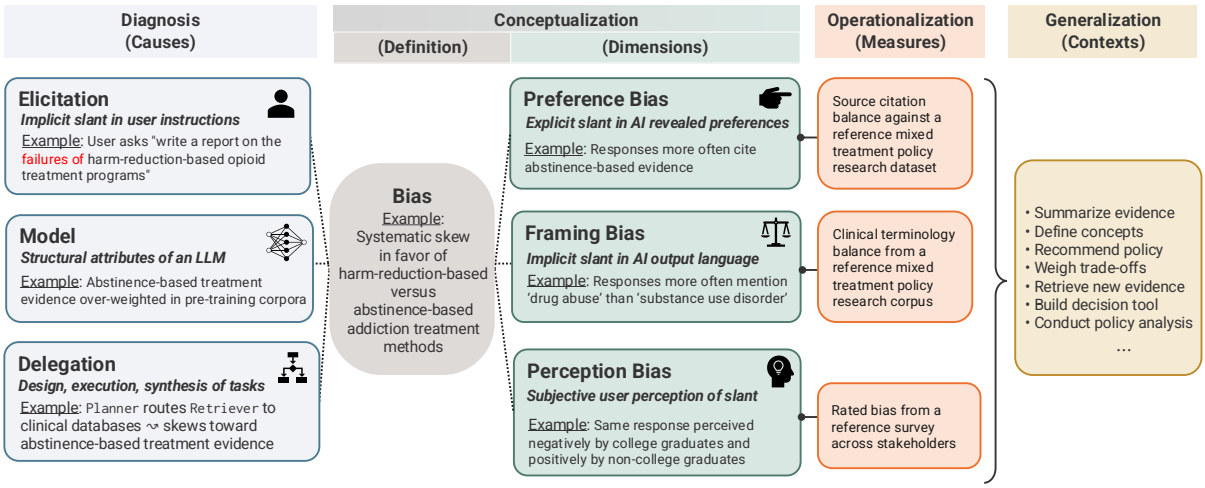


Figure 1: **Components of Valid Measurement of Bias in Agentic AI Systems.** Framework is illustrated with the example of querying an agentic AI system for evidence synthesis on opioid use disorder interventions

relevant standards, for example, the credibility of evidence in a literature review task. Observed differences in outputs should not be attributed to bias when they are explained by such criteria; bias is the residual skew in outputs favoring certain groups, ideas, viewpoints, or other concepts that persists after accounting for these quality differences. Crucially, any such definition requires an explicit *reference point* — the relevant baseline may be neutrality, balance across groups or viewpoints, population proportionality, or exclusion of low-quality sources — and no claim of bias is interpretable without one.[1] We depart from prior frameworks in expressly stipulating three dimensions of bias as a shared structure for conceptualization, which practitioners may further specify for their application context.

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 systematic 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).

Framing bias involves systematic 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 different framing choices — for instance, presenting food stamp benefits as "welfare" versus "poverty aid" — can shift judgment even when the underlying factual content is held constant (Nelson et al., 1997; Chong and Druckman, 2007). 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).

Perception 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 beliefs (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). Perception bias is downstream of the other dimensions, but may be the most consequential, because it shapes trust, adoption, and real-world use (McClain et al., 2025).

These three dimensions are not mutually exclusive. Preference and framing bias can co-occur and reinforce one another, for instance with a model that systematically selects ideologically skewed sources (preference bias) which propagates the framing conventions of those sources in its synthesis (framing bias). Although perception bias is

[1]Depending on the context, bias may be defined to encompass concepts such as toxicity, hate speech, fairness, or representational harm (Solaiman et al., 2025). Evaluators should explicitly define the dimensions of bias under study, the reference points against which it is assessed, and the quality criteria used to determine whether observed differences are task-relevant or bias-relevant.

downstream of the other two and may vary across users who observe the same output, it may still be shaped by whatever preference and framing biases are present in the underlying output. Evaluators are free to treat these dimensions as conceptually distinct, but empirically correlated targets.[2]

A prerequisite for construct validity is *operationalization* wherein multiple measures are developed to fully cover all dimensions of the construct (Bradburn et al., 2017). 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) while framing bias may be measured through differential counts of frames used by partisans 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 subdimensions (preference for liberal policy positions versus preference for liberal news sources).

A final challenge arises with *generalization*, or ensuring that estimated bias appropriately represents a broad range of contexts in which a system is actually deployed. This requires that evaluators specify two things. First, they must identify the contexts resembling realistic use cases in which bias may arise. A single evaluation setting captures one task type, topic, or user population, but bias may manifest differently across use cases; evaluators should therefore sample from the actual distribution of deployment contexts rather than from benchmark tasks constructed for convenience, a canonical concern in external validity (Findley et al., 2021). This connects to recent work arguing that frontier AI capabilities — and their failure modes — are only fully visible when systems are tested in realistic, open-ended contexts rather than closed benchmarks (Kapoor et al., 2026). Second, the evaluator must select the appropriate aggregation method, which may be a non-parametric average or some kind of model-based aggregation.

[2]This is a property of *reflective constructs*, where dimensions and their observable measures are downstream reflections rather than upstream causes of the underlying construct, which is true for *formative constructs* (MacKenzie et al., 2005).

3 Delegation and Bias in Agentic AI Systems

The dimensions described above characterize how bias appears, but they must be distinguished from their causes. We refer to the identification of these causal pathways as *diagnosis*. Although evaluators may want a single aggregate bias estimate for a particular stable implementation of an AI system, the task of diagnosis is essential for understanding how that estimate responds to changes in elicitation choices, model structure, or delegation configuration. Notably a single bias dimension may arise through multiple causes, and the same cause may surface across multiple dimensions. Identifying and further varying causal factors via randomized experiments can help disentangle these phenomena.

We argue that there are three broad types of causes for bias. While the first two are well known to appear in monolithic large language models, this paper introduces a third which arises in agentic AI systems comprising many models.

Elicitation refers to how inputs are formulated and presented to the system. Similar to how question wording, framing, and ordering can systematically alter survey responses, prompt structure, demographic cues, and source provenance produce methodological artifacts in model judgments (Kalton and Schuman, 1982; Eckman et al., 2024; Beck et al., 2025; Brucks and Toubia, 2025; Germani and Spitale, 2025; Tonneau et al., 2026; Bini et al., 2025). Observed preference or framing bias may therefore reflect properties of the elicitation rather than stable properties of the underlying model.

Models themselves induce bias through different structural attributes such as their network architecture, weights, built-in safety features or guardrails, and default generation parameters. These, in turn, may arise from aspects of the pretraining data, fine-tuning objectives, and preference optimization, as well as from the defaults chosen by application developers. Systematically imbalanced corpora may propagate systematically imbalanced citations into outputs (Feng et al., 2023); preference-based post-training may favor the representation of certain groups or "sycophantic" linguistic frames (Sharma et al., 2025; Shapira et al., 2026); and covert biases can persist or worsen even when overt bias appears reduced through human-feedback training (Hofmann et al., 2024).

Delegation refers to the assignment of decision-making authority to an AI system or subsystem. Agentic AI systems are coupled processes: an orchestration model decomposes goals and delegates subtasks, while a planning model may select tools for an execution model to use; the outputs may then be delivered to a synthesis model and propagated back to a user-facing response model (Ruan et al., 2026; Zhang et al., 2025b). While elicitation concerns what a user explicitly instructs, delegation concerns how decision-making authority is allocated in order to interpret and execute those instructions. This allocation is shaped by system-level prompts, instructions embedded in external skills and tool documentation, available tools and sources, task-routing rules, and synthesis procedures relayed through agent protocols such as the Model Context Protocol (MCP). These conditions may be opaque to the end user and may cascade across model calls: a biased routing or source-selection decision at one stage may amplify framing decisions downstream. This introduces the classic principal-agent problem, where accountability to the user's elicitation may be attenuated once decision-making authority passes to the agent (Jensen and Meckling, 1976; Moe, 1984; Gailmard, 2014).

A vast literature in political economy emphasizes that delegation is structured by *constraints*: which choices are left to agents, which are governed by rules or procedures, what information and tools agents can access, what objectives guide behavior, and what oversight mechanisms discipline action (McCubbins and Schwartz, 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 allocation of decision authority and constraints channels the agent toward biased patterns of outputs.

Delegation may moderate model-based and elicitation-based biases. Work on multi-model collaboration in agentic systems shows that small local errors can cascade into system-level errors, and intermediate outputs can be 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., 2025a). In agentic systems, delegation may not merely add a third independent cause of bias; it may also moderate how elicitation and model-based biases propagate through the pipeline. Elicitation-caused skew in intermediate task outputs can be amplified, attenuated, or rerouted depending on how the pipeline is structured: which subtasks are delegated, what constraints govern those subtasks, in what order they occur, and which models execute them. Because each nested model in an agentic system carries its own fixed attributes, the effects of upstream elicitation choices may interact with the properties of whichever model executes each stage. Evaluators diagnosing delegation-caused bias may therefore need to trace effects across subtask outputs rather than attributing observed skew to a single causal pathway implied by Figure 1.[3]

4 Stylized Illustration: Delegation in an Agentic Meta-Analysis

We illustrate the framework with a stylized example involving an AI agent asked to conduct a meta-analysis of interventions for opioid use disorder (OUD). The example is not intended as an empirical benchmark or a clinical recommendation. Rather, it shows how preference, framing, and perception bias can arise in a consequential evidence-synthesis task where bias concerns skew toward one intervention orientation rather than another.

We ground the example in the Department of Health and Human Services' overdose-prevention framework (U.S. Department of Health and Human Services, 2026a,b). Suppose a researcher asks an AI agent: "Conduct a meta-analysis comparing the effects of harm-reduction-oriented and abstinence-oriented interventions for OUD. Estimate effects on overdose, treatment initiation, treatment retention, opioid use, infectious-disease outcomes, and quality of life. Summarize the strength of the evidence and write a policy-facing conclusion." The task requires many choices before any final answer is produced: which search terms to use, which studies to include, how to classify interventions, which out-

[3] Model attributes and user elicitation may also moderate each other through what is called multi-turn context drift (context, here, referring to a language model's context window), where model behavior degrades or shift as conversation history accumulates and earlier instructions become diluted (Sharma et al., 2025; Rabanser et al., 2026). The distinction with delegation is in the locus of control: in multi-turn settings, context drifts because the user's own inputs accumulate in the prompt, but the user still remains the principal. Bias arises from a structural transfer of decision authority that the user did not author, typically cannot observe, and cannot modify.

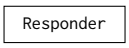
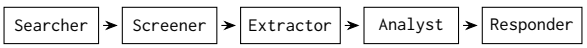
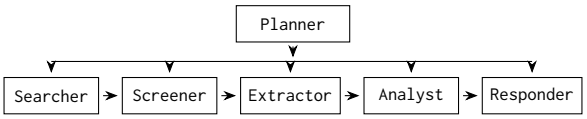
Condition	Workflow
T1: Single Responder	
T2: Constrained Delegation	
T3: Model-Designed Delegation	

Table 1: **Stylized Delegation Conditions for the Agentic Meta-Analysis.** Each agentic implementation (condition) varies in how much discretion is delegated to models. Each box is a separate LLM call; arrows show direction of inputs between models.

comes to privilege, how to handle heterogeneous populations, how to assess study quality, and how to translate statistical results into policy language.

In this setting, we define bias as *systematic skew toward either a harm-reduction-oriented or abstinence-oriented interpretation of the evidence, relative to a pre-specified meta-analytic protocol, and conditional on task-relevant quality criteria*. Here, the reference point may be considered the balance of evidence found in a high-quality research archive or compiled by an expert reviewer. Relevant criteria include study quality, outcome relevance, causal identification, external validity, clinical relevance, publication bias, and transparent handling of heterogeneity. Hence, a finding is not biased merely because it favors one intervention family but because it does so even after accounting for the protocol and the quality of the underlying evidence and above some baseline.

Preference bias appears in the agent's evidence-synthesis choices. An agent may disproportionately retrieve harm-reduction studies, classify ambiguous interventions as harm reduction, privilege overdose outcomes over abstinence or retention outcomes, or emphasize study designs that favor low-barrier services. Conversely, it may disproportionately retrieve abstinence-oriented treatment studies, privilege abstinence and program completion outcomes, exclude studies focused on overdose reversal or syringe services, or treat relapse as the dominant endpoint. These choices affect the apparent direction of the evidence before the final report is written.

Framing bias appears in how the agent describes the same body of evidence. A report might describe harm-reduction interventions as "lifesaving," "pragmatic," and "evidence-based," while describing abstinence-oriented programs as "restrictive" or "difficult to sustain." Another report might describe abstinence-oriented programs as "recovery-focused" and "comprehensive," while describing harm-reduction interventions as "temporary" or "not addressing root causes." These differences in tone, emphasis, and scope can shift interpretation even when the same studies are cited.

Perception bias appears in how different users interpret the same synthesis. A public health official may view a harm-reduction emphasis as evidence-based overdose prevention. A treatment provider may view the same report as underweighting long-term recovery. A recovery advocate may perceive the report as dismissive of abstinence goals. A methodologist may focus instead on whether the meta-analysis followed its protocol, handled heterogeneity appropriately, and distinguished evidence strength from policy preference.

Delegation enters when the system allocates methodological discretion across agents, laid out across hypothetical system implementations in Table 1. In T1, a single `Responder` model writes a narrative synthesis from the prompt alone. In T2, the system delegates subtasks to specialized models under strict constraints: a `Searcher` uses pre-specified search strings, a `Screener` applies fixed inclusion criteria, an `Extractor` codes outcomes using a fixed schema, an `Analyst` applies a pre-specified meta-analytic model, and a `Responder` writes under a constrained rubric. In T3, a `Planner` chooses the search strategy, intervention taxonomy, inclusion criteria, outcome hierarchy, quality weights, and reporting frame before downstream models execute the workflow.

The contrast between T2 and T3 isolates the role of delegation constraints. Both workflows delegate evidence synthesis, but only T3 gives the agent discretion over methodological choices that can

shift the direction of the final conclusion. If the `Selector` privileges overdose mortality and low-barrier access, the synthesis may tilt toward harm reduction. If it privileges program completion and relapse prevention, the synthesis may tilt toward abstinence-oriented treatment. The resulting skew may not necessarily arise from the user's prompt, the quality of the retrieved evidence, or the final `Responder` alone. It would arise from upstream choices about how the agent defines, filters, weighs, and summarizes the evidence.

This example illustrates why bias dimensions should be measured separately. A workflow could include a balanced set of studies but frame one intervention family more favorably. Conversely, a workflow could use neutral language while selecting outcomes or inclusion criteria that favor one policy orientation. Treating study selection and report language as interchangeable measures of bias would miss these distinctions.

5 Conclusion

Measuring bias in AI agents inherits the same methodological challenges that social scientists have long grappled with and that AI researchers have already begun to apply to evaluate LLMs. Agents make those challenges harder, not different in kind. Rather than reinvent the wheel, we extend Wallach et al. (2025)'s position forward to the agentic era and assert that bias evaluations should — where appropriate — draw on well-established tools in psychometrics (e.g. construct validity and scale development), survey methodology (e.g. elicitation design), causal inference (e.g. experimental identification of bias sources), and applied statistics more broadly (e.g. generalizability).[4]

Our framework introduces two further tools whose relevance to agentic bias evaluation has been underappreciated. From political science and social psychology, ideal point measurement motivates a three-dimensional structure for bias — preference, framing, and perception — that treats these as related but empirically distinct dimensions of a single construct rather than entirely interchangeable or entirely independent indicators. From political economy, the concept of delegation provides a diagnostic for identifying a structurally distinct causal pathway through which agentic systems introduce bias that cannot be reduced to model attributes or elicitation choices.

Our paper carries implications for how bias benchmarks for agentic systems should be designed. Current benchmarks treat task completion as the primary outcome and hold delegation structure fixed — typically delegating all subtasks to the same model in the same configuration (Liu et al., 2023; Yao et al., 2024). Our illustration suggests three design principles that such benchmarks should incorporate. First, variations in delegation structure should be baked in. Varying which decisions are delegated to the model is necessary to isolate delegation-induced bias from model and elicitation effects. Second, benchmarks should operationalize multiple bias dimensions, not just measure one and treat it as a proxy for all others. In our example, measures of preference bias would capture whether the agent selected studies, outcomes, or inclusion criteria that favored one intervention orientation, but would miss whether the final synthesis framed the same evidence more favorably or skeptically. Measures of framing bias would capture language and emphasis, but would miss upstream methodological choices that determine which evidence enters the synthesis. The two dimensions may respond differently to the same delegation manipulation, which justifies measuring them separately. Third, sampled contexts should reflect realistic deployment distributions rather than convenience tasks. A benchmark's external validity depends on whether its items represent the range of contexts in which a deployed system will actually be used (Kapoor et al., 2026).

This paper situates itself in a growing body of work arguing for deeper integration of social science theory and methods into AI research (Eckman et al., 2024; Bini et al., 2025; Wallach et al., 2025; Jung et al., 2026). Looking forward, the connection between agentic AI evaluation and the social science literature on institutional design deserves particular attention. That literature has studied how principals design delegation mechanisms — through procedural constraints, structural restrictions, and oversight requirements — to limit systematic drift in agent behavior. Future work would do well to treat delegation architecture in AI agents as a design space in which bias can be actively constrained.

[4] Although construct validity frameworks do not assume that language models reason or respond like human survey respondents, specific conceptualization or operationalization may make such assumptions. The psychometric validity of instruments designed for humans is not guaranteed for LLMs (Jung et al., 2026; Pate et al., 2026) and should be treated as an open question rather than an inheritance from the human measurement literature.

References

Jacob Beck, Stephanie Eckman, Christoph Kern, and Frauke Kreuter. 2025. Bias in the Loop: How Humans Evaluate AI-Generated Suggestions. *Preprint*, arXiv:2509.08514.

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*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

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*.

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.

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, Frauke Kreuter, and Barbara Plank. 2024. Position: 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.

Michael G. Findley, Kyosuke Kikuta, and Michael Denly. 2021. External Validity. *Annual Review of Political Science*, 24:365–393.

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.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI Generates Covertly Racist Decisions About People Based on Their Dialect. *Nature*, 633:147–154.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 375–385.

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.

Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. 2026. Do psychometric tests work for large language models? evaluation of tests on sexism, racism, and morality. *Preprint*, arXiv:2510.11254.

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.

Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, J.J. Allaire, Rishi Bommasani, Magda Dubois, Gillian Hadfield, Andy Hall, Sara Hooker, Seth Lazar, Steve Newman, Dimitris Papailiopoulos, Shoshannah Tekofsky, Helen Toner,

Cozmin Ududec, and Arvind Narayanan. 2026. Open-World Evaluations for Measuring Frontier AI Capabilities. Technical report.

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*.

Scott B MacKenzie, Philip M Podsakoff, and Cheryl Burke Jarvis. 2005. The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. *Journal of Applied Psychology*, 90(4):710.

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.

Mathew D McCubbins and Thomas Schwartz. 1984. Congressional oversight overlooked: Police patrols versus fire alarms. *American journal of political science*, pages 165–179.

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.

Terry M. Moe. 1984. The New Economics of Organization. *American Journal of Political Science*, 28(4):739–777.

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.

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*.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. 2026. 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, 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. 2025. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

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.

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*.

Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *science*, 211(4481):453–458.

U.S. Department of Health and Human Services. 2026a. Harm reduction. `https://www.hhs.gov/overdose-prevention/harm-reduction`. Accessed May 15, 2026.

U.S. Department of Health and Human Services. 2026b. Overdose prevention. `https://www.hhs.gov/programs/overdose-prevention.html`. Accessed May 15, 2026.

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.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Proceedings of the 42nd International Conference on Machine Learning*.

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*.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2025a. 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. 2025b. AgentOrchestra: A Hierarchical Multi-Agent Framework for General-Purpose Task Solving. *arXiv preprint arXiv:2506.12508*.