# Toward Responsible AI Scientists for Social Science

Meysam Alizadeh[1,2*], Helen Margetts[1], Jacob N Shapiro[3], Mohsen Mosleh[1], Cosmina Dorobantu[4], Kevin Munger[5], Atoosa Kasirzadeh[6], Yian Yin[7], Homa Hosseinmardi[8], Rebekka Burkholz[9], Veronika Batzdorfer[10], Fabrizio Gilardi[2], Ingmar Weber[11], Joshua Tucker[12], Michael Macy[7]

[1*]University of Oxford.
[2*]University of Zurich.
[3]Princeton University.
[4]LSE.
[5]European University Institute.
[6]Carnegie Mellon University.
[7]Cornell University.
[8]UCLA.
[9]CISPA Helmholtz Center for Information Security.
[10]Karlsruhe Institute of Technology.
[11]Saarland University.
[12]New York University.

*Corresponding author(s). E-mail(s): alizadeh@ipz.uzh.ch;

**Abstract**

Large language models (LLMs) are increasingly used to support parts of the scientific workflow, fueling interest in "AI scientist" frameworks. Their relevance to social science, however, remains largely untested. Unlike the natural sciences, social research is characterized by epistemic pluralism, limited ground truth, complex validity requirements, and heightened ethical and governance constraints. Here we examine these domain-specific challenges, propose targeted mitigation strategies, and advocate a collaborative "AI co-scientist" model. We outline pragmatic development pathways, benchmarking strategies, institutional reforms to support cumulative learning, and clear methodological boundaries. Rather than racing toward autonomous social science researchers, we argue that progress should focus on strengthening the conceptual and institutional foundations required to support them.

# 1 Introduction

Recent advances in large language models (LLMs), particularly in reasoning and multi-step task execution, are beginning to automate substantial components of the scientific workflow. These systems can now solve complex mathematical problems [1], assist with formal proofs [2], retrieve and synthesize prior work [3, 4], propose new hypotheses [5–8], and generate analytical or computational code [9–12]. Building on these advances, emerging end-to-end or human-in-the-loop frameworks can autonomously formulate hypotheses, run experiments, analyze results, and draft manuscripts. Together, these developments suggest the emergence of increasingly autonomous "AI scientist" frameworks that may reshape how scientific knowledge is produced [13, 14].

Existing AI scientist frameworks have been developed and evaluated almost exclusively in computer science (e.g. *AI Scientist* [15] and *Agent Laboratory* [14]), chemistry (e.g. *ChemCrow* [16] and *Coscientist* [17]), and the life sciences (e.g. *Biomni* [18]) (see supplementary materials). Some systems aim to automate end-to-end research pipelines, while others incorporate human feedback or domain-specific tools to support hypothesis generation, experimentation, and analysis. Demonstrations span applications such as diffusion modeling, chemical experimentation, and biomedical discovery [14–18]. *Kosmos* [19] is positioned as more domain-general, though its validation still centers on applications in metabolomics, materials science, neuroscience, and statistical genetics. While LLMs already show potential for transforming social science methods [20, 21], extending AI scientist frameworks to this domain requires more than domain transfer and introduces distinct methodological and epistemic risks. The epistemic pluralism that characterizes social science, in which multiple theoretical frameworks can be simultaneously defensible and empirically underdetermined, has deep roots in the philosophy of science [22], and poses distinctive challenges for AI systems trained to optimize toward singular solutions. Even in machine learning research, recent studies document systematic failures, including metric misuse and biased dataset selection [23, 24].

Here we identify domain-specific challenges and risks in developing autonomous AI scientist frameworks for social science, propose targeted mitigation strategies, and delineate key methodological boundaries. We caution against premature deployment in the absence of a clear understanding of system reasoning and reliability. Instead, we argue that a collaborative AI co-scientist—one designed to complement rather than replace human researchers and grounded in multi-agent architectures and systematic interpretability studies—offers a more responsible and feasible path forward.

# 2 Challenges and Risks of Social Science AI Scientists

## 2.1 Epistemic Disagreement and Lack of Ground Truth

Many existing AI scientist frameworks work on AI or natural science tasks, which are characterized by abundant, well-structured public datasets and codified knowledge [25]. For example, *AI Scientist* and *Agent Laboratory* rely heavily on HuggingFace datasets, while DeepMind's AlphaFold [26] would not have been possible without the Protein Data Bank [27]. Social science, by contrast, lacks comparable public

repositories that establish ground truth for many core constructs. The lack of such repositories can be traced to both the plurality of theoretical and operational definitions of key constructs and legal, ethical, and privacy constraints on data sharing about individuals.

Many core constructs in the social sciences, such as socioeconomic status, teaching effectiveness, and partisanship, are inherently unobservable and admit multiple, often competing, operationalizations. For example, socioeconomic status may be measured as individual income, family income, or as a latent construct combining some combination of income, wealth, education, and occupation [28]. These choices are not merely technical: they reflect underlying theoretical commitments [29] and normative assumptions about what the construct ought to capture [30, 31]. In fairness research, for instance, conflicts between individual and group fairness, or between predictive parity and error-rate balance in recidivism risk assessment, reflect differing value judgments about whose harms matter and how they should be weighed, rather than alternative measurements [28]. More broadly, the validity of many operationalizations remains contested [32, 33], underscoring that measurement in social science is inseparable from theory and interpretation. This view is consistent with philosophical accounts of validity that treat it not as a psychometric property of an instrument but as a theoretical claim about the relationship between a latent attribute and its operationalization [34], a framing that renders automated construct selection without theoretical grounding particularly problematic [28].

## 2.2 Evaluation, Robustness, and Non-Human Reasoning

In domains with fixed benchmarks, AI scientist frameworks can be evaluated against known solutions using standard performance metrics such as accuracy. Social science lacks a direct analogue to an "accuracy" score [35], and even closely related metrics (e.g., reliability) can lead to qualitatively different conclusions [36]. Evaluating social science research instead requires assessing multiple dimensions, including statistical inference and internal, construct, and external validity [30], making robust, human-like reasoning central to AI scientist development in this domain [21].

Social science AI scientists face two additional challenges. First, the growing emphasis on reasoning-centered evaluation of AI systems—where models are assessed on multi-step inference, abstraction, and analogical reasoning rather than factual recall—raises concerns about the reliability of LLM reasoning in theory-driven domains [37]. Social science research frequently requires transferring theoretical constructs across contexts, interpreting ambiguous social categories, and drawing analogies between historical or institutional settings. Yet LLMs often fail to generalize reliably in analogy-making and abstract reasoning [38, 39], producing brittle inferences that appear plausible but rest on flawed conceptual mappings [40, 41]. In social science workflows, such failures could silently propagate into theory selection, construct definition, or causal interpretation, creating risks that are difficult to detect through standard benchmark-style evaluation.

For example, an AI system tasked with analyzing online protest movements might draw analogies to historical revolutions, such as comparing a contemporary hashtag campaign to the early stages of the Arab Spring. While digital platforms played

an important role in mobilization during events such as the Egyptian and Tunisian uprisings [42–44], such comparisons can rest on superficial similarities, such as rapid information spread, while ignoring deeper differences in institutional context, political repression, or communication infrastructure.

Second, modern AI systems exhibit well-documented behavioral failure modes [45], including hallucinations in large language models [46–48] and sycophantic responses [49], which pose particular risks in data-sparse scientific domains, such as fabricating nonexistent datasets [50]. One documented driver of hallucination is data scarcity [51], and given that many core social-scientific constructs remain underrepresented in training corpora, hallucination risks are likely elevated in social science applications [51]. Synthetic data does not fully resolve this limitation and may instead induce model collapse under repeated reuse [52].

## 2.3 Ethical, Contextual, and Governance Constraints

Social science research is shaped by ethical, contextual, and governance constraints that AI systems must carefully navigate, especially when applications draw on digital trace and platform data [53, 54]. In addition to standard human-subject and privacy safeguards, the field relies on evolving norms around consent, data protection, and platform governance, and it often uses social data that raise concerns about downstream harms [55]. Automated agents that propose studies, collect data, or design interventions may unintentionally circumvent IRB safeguards and restrictions, or engage vulnerable populations in ways that human researchers would recognize as inappropriate. LLMs have been shown to miss ethical concerns identified by human reviewers and produce inconsistent assessments across runs [56]. At the same time, as LLMs are increasingly used in social science, researchers still lack clear guidance on tool reliability, responsible use, and appropriate evaluation standards [57]. Recent work further highlights emerging threats from adversarial content designed to manipulate, deceive, or exploit visiting agents (see [58] for a recent taxonomy of AI agent traps). Together, these challenges position governance and ethical reasoning as core requirements rather than peripheral safeguards for AI scientists operating in social science settings.

## 2.4 Undisclosed Search and the Risk of Fishing

Recent work suggests that AI agents may inadvertently enable forms of exploratory or selective data searching, raising concerns analogous to "fishing" in social science when analytical search processes are not fully disclosed. Iterative prompt refinement can function as a form of selective reporting, analogous to p-hacking, allowing users to steer models toward preferred outputs without transparent documentation. At the same time, evidence on sycophancy indicates that LLMs often produce plausible or user-aligned responses under uncertainty rather than abstaining. For example, reframing specification search as uncertainty reporting has been shown to be associated with p-hacking-like behavior in AI coding agents [59, 60], while the non-determinism of LLM outputs suggests that repeated interactions with the same model may constitute an implicit and undisclosed search over outputs [61]. These behaviors—captured under the concept of "LLM hacking" [62] (see Supplementary Information)—suggest

that poorly governed AI scientists could unintentionally amplify long-standing risks of false discovery and biased estimation associated with exploratory flexibility in social research. A constructive path forward is to systematically characterize how LLMs perform core social science tasks before granting them greater autonomy in the research workflow [63].

### 2.5 Production-Progress Paradox and the Risk of Slow Science

AI systems primarily accelerate the production of conventional research outputs, such as text, code, and summaries, without necessarily reshaping how knowledge is represented, evaluated, or integrated across studies [64]. Similar concerns were raised in earlier work on electronic publication, which showed that increased efficiency in accessing literature was associated with narrowing the range of ideas scientists engage with [65]. As a result, AI scientist frameworks could increase research output without accelerating, and potentially even slowing, scientific progress when the primary bottlenecks lie not in producing artifacts but in resolving conceptual disagreements, integrating conflicting evidence, or designing informative tests [66, 67].

Recent commentary by Sayash Kapoor and Arvind Narayanan [68] highlights this "production–progress" mismatch, in which AI systems optimize for throughput despite key epistemic constraints lying elsewhere (see supplementary information). Supporting this concern, a large-scale analysis of 41.3 million papers finds that scientists using AI tools publish more and receive more citations, yet AI adoption is associated with a contraction in collective scientific focus and reduced follow-on engagement, as AI-augmented work concentrates in data-rich domains and automates established research programs rather than opening new ones [69–71]. This dynamic may be particularly consequential in social science, where foundational disagreements persist and progress depends heavily on resolving contested interpretations rather than scaling routine production. For example, literature reviews are intended to synthesize competing perspectives and demonstrate scholarly judgment, but risk becoming machine-generated artifacts produced and consumed by other machines, thereby weakening their role as vehicles for conceptual integration.

## 3 Recommendations

### 3.1 Multi-agent AI Co-Scientists for Social Science

Recent work has proposed the concept of an "AI co-scientist": a multi-agent system designed to collaborate with, rather than replace, human researchers by generating, critiquing, and iteratively refining hypotheses and study plans [17, 72, 73]. These systems typically rely on specialized agents that debate and rank competing proposals across multiple rounds, improving outputs through additional test-time computation and access to external tools [74–76]. In social science, such systems could scaffold core elements of the research workflow, including literature synthesis, theory comparison, preregistration drafting, and measurement design, while leaving substantive judgment and final decisions to human researchers. Beyond accelerating exploration,

these systems may also help detect errors and improve transparency by strengthening expectations for documenting research choices and analytical pathways.

This emerging literature suggests a preliminary design space for social science applications. One potential component is a meta-review agent that explicitly surfaces competing theoretical framings of a construct rather than collapsing distinctions between them [28]. A second component is a knowledge-graph agent that builds and maintains links among theories, methods, datasets, and prior findings, supporting cumulative reasoning across studies. A third is a governance-oriented agent that records prompt iterations, tool calls [77], and specification changes, thereby increasing visibility into otherwise opaque decision processes and reducing opportunities for undisclosed analytical search. To date, however, none of these components have been systematically implemented or evaluated in social science contexts.

These proposed components should be understood as design hypotheses rather than validated prescriptions. Their motivation draws from capabilities demonstrated in adjacent domains: meta-review and critique mechanisms have been evaluated in peer-review simulations [78]; knowledge-graph systems have supported cumulative reasoning in biomedical discovery [17]; and tournament-style self-critique has improved output quality in general multi-agent frameworks [74]. Whether these capabilities transfer to social science remains an open empirical question. Social science research is characterized by contested constructs, limited ground truth, and multidimensional validity requirements, raising uncertainties about whether meta-review agents can meaningfully preserve theoretical disagreements, whether knowledge graphs can encode context-sensitive generalizability constraints [79], and whether governance logging alone can detect or deter undisclosed analytical search in practice.

Accordingly, we frame these components as candidate mechanisms to be evaluated through narrowly scoped prototypes, rather than as elements of a finalized architecture. This position is consistent with the staged, interpretability-first development strategy outlined in Section 3.5, which prioritizes empirical validation of individual components before large-scale system integration.

## 3.2 Pragmatic Starting Points

A pragmatic strategy is to begin with tasks and domains that offer clear benefits alongside well-specified data and evaluation criteria. On the task side, reproducibility [80], replication [81], preregistration compliance checks [81], early-stage manuscript review [78], and experiments that use structured interactions with AI systems as treatments provide natural entry points [82]. Recent evidence further shows that AI coding agents can reproduce published social science findings with near-perfect accuracy, highlighting the feasibility of structured evaluation pipelines [60]. An independent audit demonstrates that Claude Code exactly reproduced the main results of a published political science paper, accurately reconstructed treatment variables, collect new data, and re-estimate core specifications with minimal human intervention [83].

On the domain side, early development should prioritize data-rich areas with established measurement practices and partial ground truth. These include elections and voting (for example, MIT Election Data and Science Lab datasets [84] and ParlGov [85]); political polarization (for example, American National Election Studies

feeling thermometer measures); legislative behavior (for example, Voteview roll-call voting records [86]); credit and risk modeling (for example, German Credit data [87], credit card fraud detection datasets [88], and explainable credit scoring benchmarks); psychometrics (for example, the myPersonality dataset [89]); public opinion (for example, OpinionQA [90] and SubPOP [91]); cognition (for example, Psych-101 [92]); social experiments (for example, SOCSCI210 [93]); and human–AI interaction studies [94]. Prioritizing task structure and data availability over disciplinary labels enables systematic validation before deployment in more ambiguous, data-poor domains.

### 3.3 Non-Static and Fluid Benchmarking

AI co-scientist benchmark design involves defining tasks, curating data, and specifying evaluation protocols. Beyond creating gold-standard annotations, rigorous assessment requires benchmarks that move beyond static knowledge retrieval [25]. Rather than only ImageNet-style classification benchmarks [95], researchers can design "miniature research episodes" that mirror core stages of the research workflow [96]. For example, consider political ideology: an ImageNet-style benchmark treats ideology as a fixed label and evaluates whether a model can classify individuals as "liberal" versus "conservative". A "miniature research episode" benchmark would supplement ML ideological labeling by situating political ideology within a broader research workflow, linking it to research questions (e.g., the role of ideology in voting behavior or opinion formation), alternative operationalizations (e.g., self-reported surveys or inference from social networks), methodological choices (e.g., survey experiments versus observational designs), tool or software use (e.g., Prolific or R packages), interpretation, and robustness checks.

For data curation, benchmarks can draw on existing research data and code repositories (e.g. OSF and Dataverse), public opinion surveys (e.g., the American Trends Panel [97]), election studies (e.g., the Comparative Study of Electoral Surveys (CSES) [98]), and open experimental archives (e.g., NSF's Timesharing Experiments [99]) to assess LLM reasoning. For example, AI co-scientists can be tasked with reconstructing or proposing research questions and theoretical constructs based on preregistered designs. For evaluation, model outputs should be assessed by panels of domain experts using explicit rubrics (e.g., conceptual clarity, internal validity, and ethical acceptability), rather than against single ground-truth labels. Recently proposed approaches such as Fluid Benchmarking [100, 101], which support evaluation across multiple relevant dimensions, are well suited to this context. While such benchmarks cannot resolve the ground-truth problem, they enable principled, pluralistic evaluation under real epistemic and ethical constraints.

### 3.4 Institutional Reform

To enable productive uses of AI in social science, research must move beyond the peer-reviewed manuscript as the dominant unit of knowledge production. Rather than using AI to accelerate existing outputs, we should leverage its capacity for structured, cumulative knowledge representation [79]. LLM-based systems could maintain machine-readable links among theories, methods, data, and findings, integrating prior

work directly into ongoing research rather than relegating it to narrative literature reviews (see supplementary materials). In domains with repeated studies using comparable methods, AI systems could also support continuously updated evidence bases, with individual contributions feeding into shared, living databases rather than standalone papers. Realizing these designs will require institutional reforms, including new norms and credit-allocation mechanisms to recognize individual contributions within collaborative research infrastructures.

### 3.5 Mechanistic Interpretability and Incremental Extension of Prior Work

Advancing AI systems without first understanding how existing models perceive and reason about social data risks producing tools that generate plausible but conceptually unsound research. A more prudent path is to replicate and dissect prior demonstrations of LLM capabilities in social science using mechanistic interpretability techniques [102, 103]. By uncovering the internal representations and circuit-level reasoning behind tasks [104, 105], we can evaluate whether LLMs are genuinely capturing social mechanisms or merely exploiting superficial correlations.

Aligned with this view, Melanie Mitchell's NeurIPS 2025 keynote emphasizes the importance of incremental, well-grounded extensions rather than sweeping leaps toward human-level scientific reasoning. Building an AI co-scientist for social science should follow this philosophy: extend existing systems only as far as we can explain and validate them. This means prioritizing replication studies, mechanistic audits, controlled evaluations of failure modes, and narrowly scoped prototypes that tackle specific components of the research pipeline. Such a staged approach not only mitigates risks but also ensures that progress is cumulative, interpretable, and scientifically meaningful. In short, rather than racing to build an autonomous social science researcher, the field should focus on understanding and strengthening the foundations upon which such a system would eventually rest.

## 4 Methodological Boundaries

### 4.1 Methodological Boundaries Across Research Designs

AI co-scientists should not be expected to perform uniformly across social science methods. Their reliability depends on both the epistemic task – whether data is obtained through observational studies or controlled experiments – and whether the mode of interaction with the social world is online or in-person (Figure 1). In observational online research, such as social media analysis, AI co-scientists are well suited to core tasks including data collection, analytic planning, and large-scale measurement. Observational offline studies, by contrast, often rely on situated interpretation, embodied presence, and contextual knowledge (e.g., schools, organizations, communities) that current systems cannot access, sharply limiting their role beyond preparatory or analytic support.

Experimental research also faces important constraints. In online settings, AI systems can support experimental design, power analysis, preregistration, analysis, and

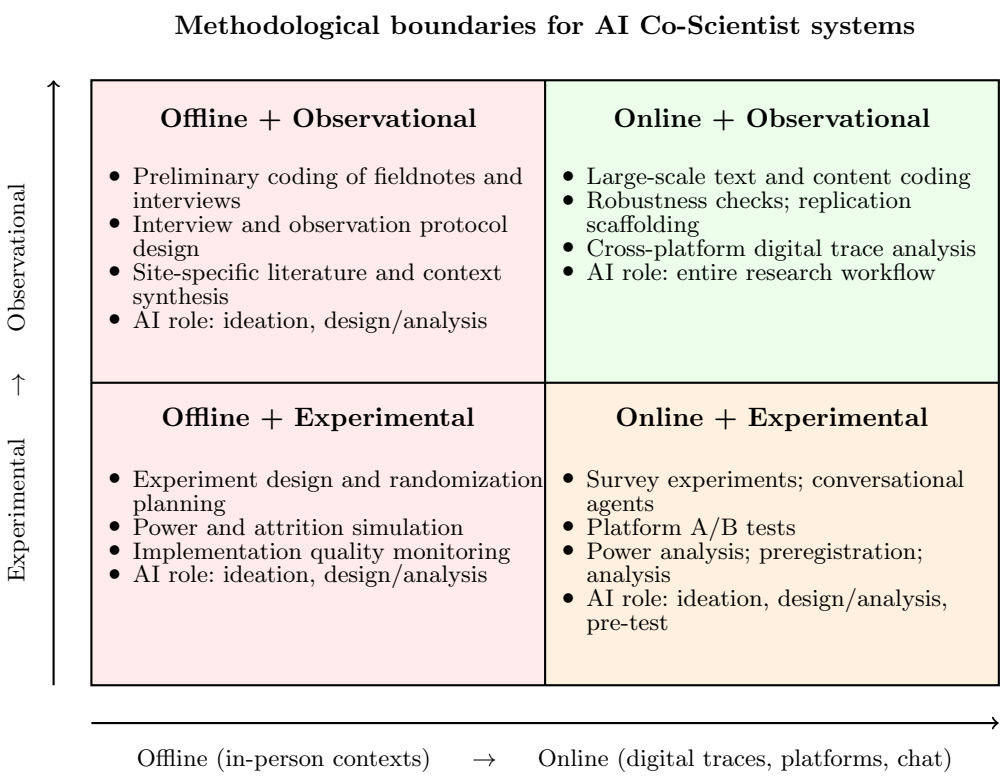


**Fig. 1 A framework for AI co-scientist support across social science research designs.** The matrix organises social science research along two axes: epistemic approach (experimental versus observational) and setting (offline in-person versus online digital). Color coding indicates the expected degree of task automation feasible for AI co-scientists in each quadrant (green, high; orange, moderate; red, low). In online observational settings, AI co-scientists can support the entire research workflow, including large-scale text and content coding, robustness checks, replication scaffolding, and cross-platform digital trace analysis. In online experimental settings, their role spans ideation, study design, power analysis, preregistration, and implementation of platform-level A/B tests and survey experiments, though reliability declines when causal identification depends on behavioral responses or spillover effects. In offline contexts, AI co-scientists are limited to preparatory and analytic support, such as interview and observation protocol design, site-specific literature synthesis, and preliminary coding of fieldnotes, because situated interpretation, embodied presence, and real-world institutional access remain beyond current system capabilities.

in some cases implement interventions via platform-level A/B tests or conversational agents [82]. Their reliability, however, declines when causal identification hinges on behavioral responses, incentives, compliance, or spillover effects, where AI-generated behavior is a poor proxy for human decision-making. Offline experiments present even more difficult challenges. Without the capacity to intervene, observe, and adapt within real-world institutional contexts, AI co-scientists cannot engage mechanisms central to

causal identification. Even for image- or video-based experiments, where multimodal models can assist with stimulus characterization and manipulation checks, it remains unclear whether they capture human perceptual, affective, and culturally mediated responses.

## 4.2 "Silicon Sampling" as Survey Takers for Public Opinion Measurement

The original "silicon sampling" approach proposes using LLMs as conditional simulators of human responses by prompting them with detailed sociodemographic backstories drawn from real survey data [106]. Recent research suggests that silicon sampling can be valuable for pretesting surveys, refining question wording, and exploring initial messaging concepts [107] (see Supplementary Materials). However, it is essential to avoid an AI "ouroboros" drifting away from the social world social scientists seek to explain. While formal and computational theory has an important role, the empirical core of social science must remain outward-facing, and replacing survey or social data with silicon samples for reasons of economic efficiency risks eroding that core. Until its underlying mechanisms are better understood and its limitations can be reliably mitigated, we do not recommend its use for estimating real-world attitudes, preferences, or voting intentions. The central concern is not only bias or insufficient variance, but unverifiable reliability, as model outputs vary with prompting choices, sampling parameters, and model updates [108–110]. Further, there are also much larger questions at stake here in terms of what we actually mean by public opinion – e.g., is it simply a set of answers to survey questions about public issues (which we know is filled with a myriad of challenges for making inferences to opinions of larger populations) or are there other aspirations here tied to norms of democratic governance? – that go beyond the scope of this current manuscript.

## 4.3 High-Stakes Data Collection in Sensitive Settings

Recent work suggests that carefully designed LLM-based interviewers can complement traditional methods in low-risk settings, for example, by scaling short qualitative interviews, eliciting richer responses than open-text surveys, and enabling rapid hypothesis generation when deployed under clear ethical guardrails and human oversight [111] (see Supplementary Information).

Prior research shows that computer-administered surveys and virtual interviewers can reduce social desirability pressures and increase willingness to disclose sensitive information [112–115]. LLM-based interviews may therefore offer similar advantages for certain sensitive topics, although this remains an open empirical question requiring systematic validation.

However, recruitment, interviewing, or data collection involving vulnerable populations or legally sensitive behaviors should not be automated end-to-end [116, 117]. Such contexts require harm-aware protocols, informed consent, and the capacity to respond to distress, coercion, or safety risks. Automation increases the likelihood of inappropriate questioning, unintended disclosure, and unverifiable compliance with ethical standards.

# 5 Conclusion

Developing "AI scientist" systems for social science remains constrained by epistemic pluralism, ethical obligations, and reliance on situated human judgment. Poorly governed systems risk reinforcing fragile inference and misaligned incentives — and the case for caution is not that human researchers are reliable where AI is not, but that ungoverned automation may amplify the very incentive failures, including publication pressure, selective reporting, and motivated reasoning, that already compromise social science. Rather than treating these constraints as obstacles, we argue that they clarify where AI can most productively complement human expertise and where oversight must remain central. Our recommendations emphasize pragmatic development and domain-specific validation to strengthen the foundations of social inquiry. By prioritizing online observational studies and recommending multi-agent AI co-scientists with specialized agents to address core methodological and epistemic challenges, the field can harness computational advances to support more transparent and theoretically ambitious social science. In short, the limits of AI scientists are not only computational but interactional. This perspective has sought to clarify where automation is feasible, where it raises governance concerns, and where it remains speculative.

# Declarations


- Funding: Not applicable.
- Conflict of interest/Competing interests: There are no competing interests to declare.
- Ethics approval and consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Data availability: Not applicable.
- Materials availability: Not applicable.
- Code availability: Not applicable.
- Author contribution: MA conceived the study and wrote the first draft. HM, JS, MM, CD, KM, IW, and JT contributed to the development of the manuscript by proposing new content and refining existing sections. All authors reviewed and edited the manuscript.

# References


[1] Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482 (2024).

[2] Collins, K. M. *et al.* Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences* **121**, e2318124121 (2024).

[3] Press, O. *et al.* Citeme: Can language models accurately cite scientific claims? *Advances in Neural Information Processing Systems* **37**, 7847–7877 (2024).

[4] Asai, A. *et al.* Synthesizing scientific literature with retrieval-augmented language models. *Nature* 1–7 (2026).

[5] Girotra, K., Meincke, L., Terwiesch, C. & Ulrich, K. T. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN* **4526071** (2023).

[6] Lu, C., Hu, S. & Clune, J. *Automated capability discovery via model self-exploration* (2025).

[7] Faldor, M., Zhang, J., Cully, A. & Clune, J. *Omni-epic: Open-endedness via models of human notions of interestingness with environments programmed in code* (2025).

[8] Hu, S., Lu, C. & Clune, J. *Automated design of agentic systems* (2025).

[9] Tian, M. *et al.* Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems* **37**, 30624–30650 (2024).

[10] Huang, Q., Vora, J., Liang, P. & Leskovec, J. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302* (2023).

[11] Lu, C. *et al.* Discovering preference optimization algorithms with and for large language models. *Advances in Neural Information Processing Systems* **37**, 86528–86573 (2024).

[12] Ma, Y. J. *et al. Eureka: Human-level reward design via coding large language models* (2024).

[13] Yamada, Y. *et al.* The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066* (2025).

[14] Schmidgall, S. *et al.* Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227* (2025).

[15] Lu, C. *et al.* Towards end-to-end automation of ai research. *Nature* **651**, 914–919 (2026).

[16] M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535 (2024).

[17] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

[18] Huang, K. *et al.* Biomni: A general-purpose biomedical ai agent. *biorxiv* (2025).

[19] Mitchener, L. *et al.* Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824* (2025).

[20] Grossmann, I. *et al.* Ai and the transformation of social science research. *Science* **380**, 1108–1109 (2023).

[21] Bail, C. A. Can generative ai improve social science? *Proceedings of the National Academy of Sciences* **121**, e2314021121 (2024).

[22] Longino, H. E. Theoretical pluralism and the scientific study of behavior. *Scientific pluralism* **19**, 102–131 (2006).

[23] Luo, Z., Kasirzadeh, A. & Shah, N. B. *The more you automate, the less you see: The hidden pitfalls of ai scientist systems.*

[24] Tang, X. *et al.* Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications* **16**, 8317 (2025).

[25] Song, Z. *et al.* Evaluating large language models in scientific discovery. *arXiv preprint arXiv:2512.15567* (2025).

[26] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *nature* **596**, 583–589 (2021).

[27] Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research* **47**, D520–D528 (2019).

[28] Jacobs, A. Z. & Wallach, H. *Measurement and fairness*, 375–385 (2021).

[29] Feyerabend, P. *Against method: Outline of an anarchistic theory of knowledge* (Verso Books, 2020).

[30] Shadish, W. R., Cook, T. D. & Campbell, D. T. *Experimental and quasi-experimental designs for generalized causal inference.* (Houghton, Mifflin and Company, 2002).

[31] Kusmaryono, I., Wijayanti, D. & Maharani, H. R. Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. *International Journal of Educational Methodology* **8**, 625–637 (2022).

[32] Benoit, K., Munger, K. & Spirling, A. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science* **63**, 491–508 (2019).

[33] Bramson, A. *et al.* Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science* **84**, 115–159 (2017).

[34] Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. The concept of validity. *Psychological review* **111**, 1061 (2004).

[35] Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* **16**, 199–231 (2001).

[36] Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).

[37] Mitchell, M. The turing test and our shifting conceptions of intelligence (2024).

[38] Mitchell, M. Artificial intelligence learns to reason. *Science* **387**, eadw5211 (2025).

[39] Mitchell, M. The metaphors of artificial intelligence (2024).

[40] Stevenson, C. E., Pafford, A., van der Maas, H. L. & Mitchell, M. Can large language models generalize analogy solving like children can? *arXiv preprint arXiv:2411.02348* (2024).

[41] Lewis, M. & Mitchell, M. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215* (2024).

[42] Tufekci, Z. & Wilson, C. Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of Communication* **62**, 363–379 (2012).

[43] Breuer, A. Social media and protest mobilization: Evidence from the tunisian revolution. *Democratization* **22**, 764–792 (2015).

[44] Howard, P. N. & Hussain, M. M. *Democracy's Fourth Wave? Digital Media and the Arab Spring* (Oxford University Press, 2013).

[45] Nguyen, A., Yosinski, J. & Clune, J. *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*, 427–436 (2015).

[46] Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. *On faithfulness and factuality in abstractive summarization*, 1906–1919 (2020).

[47] Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM computing surveys* **55**, 1–38 (2023).

[48] Sun, Y., Sheng, D., Zhou, Z. & Wu, Y. Ai hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications* **11**, 1–14 (2024).

[49] Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J. & Tenenbaum, J. B. Sycophantic chatbots cause delusional spiraling, even in ideal bayesians. *arXiv preprint arXiv:2602.19141* (2026).

[50] Mitchell, M. Why ai chatbots lie to us (2025).

[51] Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. Why language models hallucinate. *arXiv preprint arXiv:2509.04664* (2025).

[52] Shumailov, I. *et al.* Ai models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).

[53] Lazer, D. M. J. *et al.* Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).

[54] Breuer, J., Bishop, L. & Kinder-Kurlanda, K. The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society* **22**, 2058–2080 (2020).

[55] Dunleavy, P. & Margetts, H. Data science, artificial intelligence and the third wave of digital era governance. *Public Policy and Administration* **40**, 185–214 (2025).

[56] Li, J.-J. *et al.* Pluriharms: Benchmarking the full spectrum of human judgments on ai harm. *arXiv preprint arXiv:2601.08951* (2026).

[57] Chakravorti, T. *et al.* Social scientists on the role of ai in research. *arXiv preprint arXiv:2506.11255* (2025).

[58] Franklin, M., Tomašev, N., Jacobs, J., Leibo, J. Z. & Osindero, S. Ai agent traps (2026).

[59] Asher, S. G. *et al.* Do claude code and codex p-hack? sycophancy and statistical analysis in large language models (2026).

[60] Alizadeh, M., Mosleh, M., Gilardi, F. & Tucker, J. A. Evaluating ai coding agents in social science reproducibility (2026).

[61] Atıl, B. *et al. Non-determinism of "deterministic" LLM system settings in hosted environments*, 135–148 (2025).

[62] Baumann, J. *et al.* Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825* (2025).

[63] Balluff, P. *et al.* Newer, larger, better? a critique of the unreflective llm adoption in communication research. *Political Communication* 1–10 (2026).

[64] Sutton, R. S. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html (2019). Online essay.

[65] Evans, J. A. Electronic publication and the narrowing of science and scholarship. *science* **321**, 395–399 (2008).

[66] Jones, B. F. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies* **76**, 283–317 (2009).

[67] Bloom, N., Jones, C. I., Van Reenen, J. & Webb, M. Are ideas getting harder to find? *American Economic Review* **110**, 1104–1144 (2020).

[68] Kapoor, S. & Narayanan, A. Could ai slow science? (2025). URL https://www.normaltech.ai/p/could-ai-slow-science.

[69] Wu, L., Wang, D. & Evans, J. A. Large teams develop and small teams disrupt science and technology. *Nature* **566**, 378–382 (2019).

[70] Chu, J. S. & Evans, J. A. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* **118**, e2021636118 (2021).

[71] Hao, Q., Xu, F., Li, Y. & Evans, J. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature* 1–7 (2026).

[72] Gottweis, J. *et al.* Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

[73] Chowa, S. S. *et al.* From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review* (2026).

[74] Wu, Q. *et al. Autogen: Enabling next-gen llm applications via multi-agent conversations.*

[75] Haase, J. & Pokutta, S. Beyond static responses: Multi-agent llm systems as a new paradigm for social science research. *arXiv preprint arXiv:2506.01839* (2025). URL https://arxiv.org/abs/2506.01839.

[76] Madden, E. R. Evaluating the use of large language models as synthetic social agents in social science research. *Journal of Social Computing* **6**, 334–341 (2025).

[77] Schick, T. *et al.* Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems* **36**, 68539–68551 (2023).

[78] Jin, Y. *et al. Agentreview: Exploring peer review dynamics with llm agents*, 1208–1226 (2024).

[79] Goroff, D., Lewis Jr, N., Scheel, A. M., Scherer, L. & Tucker, J. A. The inference engine: A grand challenge to address the context sensitivity problem in social science research (2018).

[80] Hu, C. *et al. Repro-bench: Can agentic ai systems assess the reproducibility of social science research?*, 23616–23626 (2025).

[81] Nguyen, B. *et al.* Replicatorbench: Benchmarking llm agents for replicability in social and behavioral sciences. *arXiv preprint arXiv:2602.11354* (2026). URL https://arxiv.org/abs/2602.11354.

[82] Hackenburg, K. *et al.* The levers of political persuasion with conversational artificial intelligence. *Science* **390**, eaea3884 (2025).

[83] Straus, G. & Hall, A. How accurately did claude code replicate and extend a published political science paper? (2026). URL https://www.andrewbenjaminhall.com/Straus_Hall_Claude_Audit.pdf.

[84] MIT Election Data and Science Lab. Election data and science lab datasets. https://electionlab.mit.edu/data (2024). Accessed: 2026-04-02.

[85] Döring, H. & Manow, P. Parliaments and governments database (parlgov): Release 2024 (2024).

[86] Lewis, J. B. *et al.* Voteview: Congressional roll-call votes database. https://voteview.com (2019). Accessed: 27 July 2018.

[87] Dua, D. & Graff, C. German credit data set. http://archive.ics.uci.edu/ml (2019). UCI Machine Learning Repository.

[88] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S. & Bontempi, G. *Credit card fraud detection: A realistic modeling and a novel learning strategy*, 1597–1604 (2015).

[89] Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* **110**, 5802–5805 (2013).

[90] Santurkar, S. *et al. Whose opinions do language models reflect?*, 29971–30004 (PMLR, 2023).

[91] Suh, J., Jahanparast, E., Moon, S., Kang, M. & Chang, S. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761* (2025).

[92] Binz, M. *et al.* A foundation model to predict and capture human cognition. *Nature* 1–8 (2025).

[93] Kolluri, A., Wu, S., Park, J. S. & Bernstein, M. S. *Finetuning llms for human behavior prediction in social science experiments*, 30084–30099 (2025).

[94] Lin, H. *et al.* Persuading voters using human–artificial intelligence dialogues. *Nature* 1–8 (2025).

[95] Deng, J. *et al.* *ImageNet: A large-scale hierarchical image database*, 248–255 (IEEE, 2009).

[96] Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).

[97] Pew Research Center. The american trends panel. https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/ (2024). Accessed: 2026-04-02.

[98] Comparative Study of Electoral Systems (CSES). Comparative study of electoral systems (cses). https://cses.org/ (2024). Cross-national post-election survey project; accessed 2026-04-02.

[99] Freese, J. Collaborative research: Time-sharing experiments for the social sciences (tess). National Science Foundation Award #1627769 (2016). Time-sharing Experiments for the Social Sciences (TESS) infrastructure.

[100] Hofmann, V. *et al.* *Fluid language model benchmarking* (2025).

[101] Li, P. *et al.* Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689* (2025). URL https://arxiv.org/abs/2511.04689.

[102] Meng, K., Bau, D., Andonian, A. & Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems* **35**, 17359–17372 (2022).

[103] Bereska, L. & Gavves, E. Mechanistic interpretability for ai safety: A review. *Transactions on Machine Learning Research* (2024). URL https://openreview.net/forum?id=8X3IMKz3jH.

[104] Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S. & Garriga-Alonso, A. *Towards automated circuit discovery for mechanistic interpretability*, Vol. 36, 16318–16352 (2023).

[105] Rai, D., Zhou, Y., Feng, S., Saparov, A. & Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646* (2024).

[106] Argyle, L. P. *et al.* Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**, 337–351 (2023).

[107] Wihbey, J. & D'Alonzo, S. Ai simulations of audience attitudes and policy preferences:"silicon sampling" guidance for communications practitioners (2025).

[108] Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B. & Larson, J. M. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis* **32**, 401–416 (2024).

[109] Boelaert, J., Coavoux, S., Ollion, E., Petev, I. & Präg, P. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research* 00491241251330582 (2025).

[110] Lyman, A. *et al.* Balancing large language model alignment and algorithmic fidelity in social science research. *Sociological Methods & Research* **54**, 1110–1155 (2025).

[111] Geiecke, F. & Jaravel, X. Conversations at scale: Robust ai-led interviews with a simple open-source platform. *Available at SSRN 4974382* (2024).

[112] Tourangeau, R., Couper, M. P. & Steiger, D. M. Humanizing self-administered surveys: Experiments on social presence in web and ivr surveys. *Computers in Human Behavior* **19**, 1–24 (2003).

[113] Kreuter, F., Presser, S. & Tourangeau, R. Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly* **72**, 847–865 (2008).

[114] Lucas, G. M., Gratch, J., King, A. & Morency, L.-P. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* **37**, 94–100 (2014).

[115] Papneja, H. & Yadav, N. Self-disclosure to conversational ai: A literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing* **29**, 119–151 (2025).

[116] Quinn, C. R. General considerations for research with vulnerable populations: Ten lessons for success. *Health & Justice* **3**, 1–7 (2015).

[117] Uuk, R. *et al.* A taxonomy of systemic risks from general-purpose ai. *arXiv preprint arXiv:2412.07780* (2024). URL https://arxiv.org/abs/2412.07780.

# Appendix A Background on Existing AI Scientists

The rapid advancement of large language models (LLMs), particularly in their knowledge and reasoning abilities, has enabled a wide range of scientific applications, from solving complex mathematical problems [1] and assisting with formal proofs [2] to retrieving related work [3], and generating analytical or computational code [9]. Building on these capabilities, emerging end-to-end frameworks can now autonomously formulate hypotheses, run experiments, analyze results, and draft manuscripts. These systems aim to replicate substantial portions of the scientific workflow and therefore have the potential to reshape how scientific research is conducted [13, 14].

Although existing studies present AI scientist frameworks as domain-agnostic, most have been trained or evaluated primarily on problems in computer science, chemistry, and the life sciences. The *AI Scientist* provides an end-to-end framework for automating the research pipeline, but its validation is limited to tasks in diffusion modeling, language modeling, and grokking. *Agent Laboratory* [14], in contrast, is a human-in-the-loop system that incorporates user feedback at each stage and is evaluated on five LLM-related research questions. In chemistry, systems such as *ChemCrow* [16] and *Coscientist* [17] demonstrate autonomous ideation and experimentation. *Biomni* [18] extends this approach to biomedicine, supporting tasks ranging from gene prioritization and drug repurposing to rare-disease diagnosis and protocol design.

# Appendix B Background on Cross-Disciplinary Knowledge Transfer

Recent work highlights a distinct pathway by which LLMs enable cross-disciplinary knowledge transfer: they act as "conceptual translators" that reframe a problem in the language of another field and then surface theorems, techniques, and verification routines that would not typically be in a domain expert's default toolkit. For example, in the Gemini mathematics case studies, researchers report that frontier models can identify analogies across areas (e.g., mapping discrete combinatorial questions onto continuous formulations), retrieve relatively obscure but relevant results from adjacent literatures, and help assemble proof strategies by decomposing tasks into verifiable sub-lemmas and then iteratively refining them with human feedback [10].

A parallel form of transfer appears in work that applies transformer methods beyond text by exploiting structural similarities between domains. For example, the life2vec study [11] treats human lives as sequences of events, encoded as a "synthetic language" spanning health, education, occupation, income, residence, and related life events, and then adapts NLP-style transformer training objectives to learn embeddings and make predictions across very different targets (e.g., early mortality risk and personality nuances). This approach illustrates how techniques developed for language representation can move into longitudinal social and administrative data when sufficiently comprehensive traces are available, enabling a shared representation space that integrates heterogeneous domains and supports cross-domain inference and interpretation.

# Appendix C Background on Existing Social Science Benchmarks

Benchmarks for "AI scientists" in social science span several complementary targets. OpinionQA evaluates whether a model's answers to subjective political and social questions align with the empirical distribution of responses observed in large public-opinion surveys across demographic groups (rather than a single "correct" label), making it a natural testbed for representational and distributional validity [90]. SubPOP similarly leverages survey structure but focuses on predicting subpopulation-level response distributions at scale—pairing thousands of survey questions with tens of thousands of subgroup response distributions to support training and evaluation of models as distributional forecasters of public opinion [91]. In a more applied direction, an emerging e-commerce benchmark evaluates whether language models can predict human click and purchase sequences in real-world online shopping sessions. By focusing on next-action and outcome prediction, it assesses behavioral fidelity in incentive-driven settings beyond static question answering [14].

Two additional resources broaden evaluation from opinions to behavioral prediction. SOCSCI210 aggregates individual-level responses from a large collection of open social science experiments (millions of responses across hundreds of thousands of participants), enabling tests of whether fine-tuning on experimental data improves models' ability to predict human behavior across domains and demographic groups [93]. Psych-101, introduced alongside the Centaur line of work, provides trial-by-trial transcripts from a large set of laboratory psychology experiments (160 studies; tens of thousands of participants; over ten million choices), offering a common natural-language format for modeling and evaluating cognition-like generalization across experimental paradigms [92].

# Appendix D Background on LLM Hacking

Baumann et al. [62] define LLM hacking as the production of incorrect scientific conclusions caused by researchers' configuration choices when using large language models for data annotation, including model selection, prompt formulation, decoding parameters, and output mapping. Because LLM outputs are highly sensitive to small changes, different plausible configurations can yield different annotations that propagate into downstream analyses and flip statistical conclusions, distort effect sizes, or reverse effect directions. This creates a "garden of forking paths" at the data-generation stage, analogous to p-hacking but occurring before formal analysis [62]. LLM hacking can be intentional, when researchers cherry-pick configurations to support desired results, or accidental, when defaults and unvalidated choices lead to errors [62]. To reduce these risks, the authors argue that LLMs should be treated as measurement instruments rather than black boxes. They recommend combining LLM outputs with human annotations, using bias-corrected estimators, pre-registering and reporting all tested configurations, validating models on ground-truth data, and adopting transparent hybrid workflows to prevent both manipulation and unintended error [62].

# Appendix E Background On Production-Progress Paradox

The production–progress paradox is the idea that scientific production (papers, researchers, funding, datasets, compute) can rise rapidly while progress (genuine conceptual, theoretical, or practical advances) grows much more slowly—or even appears to stagnate. Kapoor and Narayanan [68] argue that many current "AI-for-science" visions focus on accelerating production (more papers, faster pipelines) even though the binding constraints on progress often sit elsewhere—an analogy they frame as "adding lanes to a highway when the slowdown is caused by a toll booth." They also emphasize that "progress" is hard to define and measure, and that what looks like a slowdown may depend on which notion of progress we adopt and how we operationalize it.

Several mechanisms can plausibly generate this mismatch. One is misaligned incentives: career rewards favor publishable, lower-risk work, which can push researchers toward incremental contributions and away from high-variance exploration that yields fewer papers but occasional breakthroughs [70]. Another is quality control and synthesis bottlenecks: as output volume grows, the capacity to verify results, share code and data, and integrate findings into coherent knowledge does not scale proportionally, weakening cumulative learning [20]. A further mechanism is the prediction–explanation gap: progress in science often requires better theories and understanding, not only improved prediction. Kapoor and Narayanan [68] illustrate this with the "epicycles" analogy—systems can keep improving predictive performance within a flawed framework, potentially delaying theoretical replacement.

AI could amplify the paradox by sharply lowering the marginal cost of producing analyses and manuscripts, increasing volume without addressing upstream bottlenecks in theory-building, validation, and synthesis [68]. It could also deepen the prediction–explanation gap if black-box modeling improves forecasts while offering limited insight into underlying mechanisms, prolonging reliance on weak theories [68]. At the same time, AI could mitigate the paradox if deployed to target the true bottlenecks: improving error detection and reproducibility workflows, supporting synthesis and interpretation rather than paper generation, and shifting incentives away from production metrics toward practices that build durable human understanding [68].

# Appendix F Background on the Cumulative Knowledge Representation

One concrete way to "integrate prior work directly into ongoing research" is to treat the literature as a structured input to an inference engine rather than a narrative backdrop. Goroff and colleagues [79] frame the central obstacle as context sensitivity: even when an effect is real in one setting, we often cannot predict whether (or how much) it will change when populations, places, implementation details, or time periods change. In that spirit, an AI co-scientist should not merely summarize prior studies, but continuously maintain machine-readable representations of each study's design, measures, sampling frame, implementation conditions, and estimated effects, along

with explicit metadata about plausible moderators (e.g., participant demographics, institutional setting, mode of delivery).

This kind of representation makes prior work actionable: when a researcher proposes a new study ("run X in Y context"), the system can retrieve the closest prior evidence, surface which contextual dimensions are well-covered versus underspecified, and generate principled transportability-style predictions or sensitivity analyses about where results are likely to hold and where they are likely to break. Crucially, this reframes "literature review" from a retrospective narrative into a living, computable knowledge base that supports design decisions, identifies the next most informative replication or extension, and helps cumulate evidence specifically around the contextual factors that determine generalizability.

# Appendix G Background On Fluid Benchmarking

Current LM benchmarks are increasingly strained by a "benchmarking crisis": evaluation has become expensive, noisy, and sometimes misleading, as static item sets saturate for strong models while remaining uninformative for weaker ones, and as ad hoc fixes (e.g., sampling fewer items for efficiency) can unintentionally increase variance and reduce practical utility [22]. Hofmann et al. [100] argue that many efforts to improve benchmarking target single problems in isolation—efficiency, variance, label noise, or difficulty—rather than optimizing evaluation quality jointly. They propose *Fluid Benchmarking*, which treats benchmarking as "benchmark refinement" and uses item response theory (IRT) to estimate each item's difficulty and discrimination from prior evaluation results, then applies adaptive testing to dynamically select the most informative items for a given model's capability level and to report performance in a latent ability space rather than raw accuracy. By tailoring items to the model, *Fluid Benchmarking* can reduce step-to-step evaluation variance and improve external validity while using far fewer items than random sampling, thereby addressing multiple dimensions of the benchmarking crisis simultaneously.

# Appendix H Background on Mechanistic Interpretability

Mechanistic interpretability is a line of work that aims to reverse-engineer neural networks into human-understandable mechanisms, explaining model behavior in terms of internal representations ("features"), components (neurons/heads), and higher-level "circuits", rather than only input–output correlations [105]. In transformer LMs, a widely used family of methods combines observational tools (e.g., analyzing attention patterns; using linear probes to test whether particular concepts are linearly decodable from activations; and feature-discovery methods such as sparse autoencoders to extract interpretable features) with causal interventions that test necessity/sufficiency [105]. The most common intervention techniques include ablation (removing or zeroing specific heads/neurons/paths and measuring the effect on a target behavior) and activation patching/causal tracing (swapping activations from a "clean" run into a

"corrupted" run to localize which layers/MLP blocks/heads causally mediate a behavior), exemplified by work that localizes factual recall to specific mid-layer computations and then edits them [102].

# Appendix I Background on Silicon Sampling

The original "silicon sampling" approach proposes using large language models as conditional simulators of human responses by prompting them with detailed sociodemographic backstories drawn from real survey data [106]. Rather than treating models as generic respondents, the method aims to capture how attitudes vary across social groups, introducing the notion of algorithmic fidelity to assess whether models reproduce the structure of human opinions rather than individual answers. Across several political opinion tasks, the authors show that conditioned models can approximate aggregate response patterns and some demographic correlations, particularly for exploratory analysis and instrument pretesting. At the same time, they stress that silicon samples are not substitutes for human respondents, given sensitivity to prompt design and persistent non-human reasoning patterns, and should therefore be used to support, rather than replace, human-centered social research.

While promising, subsequent studies have revealed limitations, including insufficient variance, unreliability on subjective questions, and persistent social biases [108–110]. Recently, drawing on a systematic review of roughly thirty empirical studies, Wihbey and D'Alonzo [107] evaluate the use of large language models (LLMs) across the survey research pipeline, distinguishing complementary from substitutive uses of "silicon sampling." They show that LLMs are most effective when used upstream as design aids—refining survey questions, identifying ambiguous or leading wording, translating instruments, and conducting exploratory pilots to surface low-consensus items—where they can function as cost- and time-efficient scaffolds for human-led research if treated as directional signals and validated against non-LLM sources. By contrast, the authors caution against using LLMs as substitutes for human respondents, especially on political, divisive, or cognitively complex topics: models often misrepresent opinion distributions, collapse within-group variation, exaggerate or suppress polarization, and rely on stereotypes, while exhibiting non-human cognitive patterns such as hyper-accuracy and lack of uncertainty. Even fine-tuning and retrieval augmentation yield mixed gains. The authors therefore advocate a hybrid approach in which LLMs support survey design and exploratory analysis, while human samples remain the gold standard (albeit a flawed – and becoming increasingly so as response rates plummet – gold standard, which raises very important questions for future research) for inference and decision-making, backed by transparency, logging, and systematic validation.

# Appendix J Background on AI-Led Qualitative Interviews at Scale

Recent work develops and evaluates a scalable framework for conducting AI-led qualitative interviews using LLMs [111]. Its central goal is to bridge qualitative and

quantitative social science by enabling thousands of short, semi-structured interviews that follow established principles of qualitative interviewing, such as non-directive questioning, follow-up prompts, and cognitive empathy. The authors design a single-agent LLM system driven by a flexible prompt that incorporates best practices from sociology and can be easily adapted across research domains, allowing large-scale deployment with minimal technical overhead.

Methodologically, the paper benchmarks LLM-led interviews against human-led interviews using multiple evaluation strategies, including expert ratings by trained sociologists, respondent-based quality metrics, and comparisons to open-ended survey questions. Across a range of applications, eliciting subjective mental states, political preferences, decision-making processes, and mental models of public policy, the authors find that LLM-led interviews are rated as comparable to average human experts in online settings and, in some cases (especially with voice input), approach face-to-face interview quality. Respondents provide substantially richer responses than in standard open-text surveys, and both experts and participants report high satisfaction with the interview process. The findings suggest that AI-led interviews can serve as a powerful complement to traditional qualitative methods, particularly when scale, speed, and systematic analysis are required, while not replacing in-depth human interviews.

# Appendix References


[A1] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[A2] Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.

[A3] Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. Citeme: Can language models accurately cite scientific claims? *Advances in Neural Information Processing Systems*, 37:7847–7877, 2024.

[A4] Minyang Tian, Luyu Gao, Shizhuo Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems*, 37:30624–30650, 2024.

[A5] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[A6] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

[A7] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

[A8] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[A9] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

[A10] David P. Woodruff, Vincent Cohen-Addad, Lalit Jain, Jieming Mao, Song Zuo, MohammadHossein Bateni, Simina Br
uanzei, Michael P. Brenner, Lin Chen, Ying Feng, Lance Fortnow, Gang Fu, Ziyi Guan, Zahra Hadizadeh, Mohammad T. Hajiaghayi, Mahdi JafariRaviz,

Adel Javanmard, Karthik C. S., Ken-ichi Kawarabayashi, Ravi Kumar, Silvio Lattanzi, Euiwoong Lee, Yi Li, Ioannis Panageas, Dimitris Paparas, Benjamin Przybocki, Bernardo Subercaseaux, Ola Svensson, Shayan Taherijam, Xuan Wu, Eylon Yogev, Morteza Zadimoghaddam, Samson Zhou, and Vahab Mirrokni. Accelerating scientific research with gemini: Case studies and common techniques, 2026.

[A11] Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4(1):43–56, 2024.

[A12] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[A13] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025.

[A14] Yuxuan Lu, Jing Huang, Yan Han, Bennet Bei, Yaochen Xie, Dakuo Wang, Jessie Wang, and Qi He. Beyond believability: Accurate human behavior simulation with fine-tuned llms. *arXiv preprint arXiv:2503.20749*, 2025.

[A15] Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30084–30099, 2025.

[A16] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, pages 1–8, 2025.

[A17] Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*, 2025.

[A18] Sayash Kapoor and Arvind Narayanan. Could ai slow science?, 2025.

[A19] Johan SG Chu and James A Evans. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118, 2021.

[A20] Jay Bhattacharya and Mikko Packalen. Stagnation and scientific incentives. Technical report, National Bureau of Economic Research, 2020.

[A21] Daniel Goroff, Neil Lewis Jr, Anne M Scheel, Laura Scherer, and Joshua A Tucker. The inference engine: A grand challenge to address the context sensitivity problem in social science research. 2018.

[A22] Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *Advances in Neural Information Processing Systems*, 37:98180–98212, 2024.

[A23] Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025.

[A24] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

[A25] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[A26] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[A27] James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[A28] Julien Boelaert, Samuel Coavoux, Etienne Ollion, Ivaylo Petev, and Patrick Präg. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, page 00491241251330582, 2025.

[A29] Alex Lyman, Bryce Hepner, Lisa P Argyle, Ethan C Busby, Joshua R Gubler, and David Wingate. Balancing large language model alignment and algorithmic fidelity in social science research. *Sociological Methods & Research*, 54(3):1110–1155, 2025.

[A30] John Wihbey and Samantha D'Alonzo. Ai simulations of audience attitudes and policy preferences:"silicon sampling" guidance for communications practitioners. 2025.

[A31] Friedrich Geiecke and Xavier Jaravel. Conversations at scale: Robust ai-led interviews with a simple open-source platform. *Available at SSRN 4974382*, 2024.