DATA&
SOCIETY

June 27, 2025

Response to the American Science Acceleration Project (ASAP) Request for Information

We welcome the opportunity to respond to the ASAP Request for Information on behalf of Data & Society, an independent, nonprofit research institute that studies the social implications of automation and AI. Our comment draws on ongoing research that explores how AI is transforming everyday work in scientific laboratories — not only by accelerating technical workflows, but by reconfiguring how scientific facts are produced, how expertise is recognized, and how epistemic authority is distributed across people and machines. While AI companies advertise that their technology will accelerate certain kinds of discovery, our findings underscore that this acceleration is neither neutral nor uniform. AI adoption in science creates new forms of dependence, new sites of friction, and new questions about accountability, reproducibility, and the value of human judgment.

We commend the ASAP initiative for foregrounding the need for new infrastructure, metrics, and institutional models to accelerate science in the public interest. **But acceleration alone is not a sufficient goal.** If the United States is to lead the next era of scientific innovation, it must invest not only in speed but in epistemic integrity, continuous maintenance of data infrastructures, and accountable systems of knowledge production. W*e use the term "epistemic" to refer to practices of knowledge production — how we come to know things, what counts as valid knowledge, and how we evaluate whether something is true or trustworthy*. Our comment highlights how AI mutually shapes scientific workflows and offers recommendations to ensure that acceleration does not come at the cost of validity of scientific claims.

Below, in responding to Questions 3, 4, and 6 of the RFI, we make three points:

1. **Scientific acceleration must prioritize epistemic integrity.** AI-assisted scientific workflows challenge traditional norms of transparency, reproducibility, and human interpretability. Without continued investments in explicit protocols for validation and contestability, scientific knowledge may become faster but less trustworthy. Federal scientific agencies should leverage their funding and institutional authority to establish researcher-led consensus on epistemic integrity in using AI for science.
2. **Reproducibility in the age of AI requires standards.** Probabilistic outputs from generative AI systems challenge traditional methods of verification. We propose institutional innovations such as community-managed repositories of reference outputs and deliberative standards for acceptable variability in AI outputs. Such shared scientific infrastructures require federal investment and sustained support from agencies such as NIST, NSF, NIH, and DOE.

3. **Data infrastructures are epistemic institutions.** AI-trained models often obscure data provenance, rely on uneven metadata, and potentially lock researchers into proprietary platforms. We call for federal investments in data auditability, provenance standards, and support for datasets that reflect a broad range of scientific approaches, perspectives, and knowledge systems.

Our recommendations aim to ensure that the acceleration of science also advances public trust, interpretability, and inclusive participation in knowledge production.

Q3. How do we ensure appropriate design of new scientific workflow models that offload certain tasks to AI while keeping human scientists at the center of the discovery process?

Using AI to accelerate scientific discovery presents significant opportunities for advancing knowledge across disciplines. However, accelerating science is not merely a matter of computational efficiency. Without careful attention to epistemic integrity — the trustworthiness, interpretability, and legitimacy of scientific knowledge — we risk generating outputs that are faster but neither reliable nor meaningful.

Although AI systems are increasingly being used to generate hypotheses, analyze data, and interpret results, they continue to operate as epistemic black boxes, delivering outputs whose internal logic scientists find difficult to fully verify or replicate. This problem is not unprecedented. Sociologist Donald MacKenzie's analysis of the 1976 computer-assisted proof of the four-color theorem underscores how the introduction of computational methods created profound epistemic discomfort — not because the math was wrong, but because the community could no longer inspect every step of the proof by hand.[1] Mathematicians questioned the legitimacy of proofs that relied on computational processes inaccessible to direct human verification. Ultimately, trust in such proofs had to be socially negotiated through independent replications, institutional endorsements, and revised norms for acceptable evidence.

This issue is magnified today in AI-driven research, particularly at the cutting edge of science, where knowledge is unsettled and the criteria for success are still in formation. In these frontier contexts, scientists frequently encounter what sociologist of science Harry Collins has identified as the *experimenter's regress* — a circular uncertainty in which there is no way to determine whether an experiment was done correctly without knowing the correct result and no way to know the correct result without doing the experiment correctly.[2] Consider, for example, an AI-driven medical imaging scenario where a model flags early signs of disease that clinicians have missed. Without an independent way to verify the model's output, clinicians face a similar challenge: accepting the model's judgment requires presuming its correctness, yet validating

[1] Donald MacKenzie, "Slaying the Kraken: The Sociohistory of a Mathematical Proof," *Social Studies of Science* 29, no. 1 (February 1, 1999): 7–60, https://doi.org/10.1177/030631299029001002.
[2] Harry Collins, *Changing Order: Replication and Induction in Scientific Practice* (London: Sage, 1985).

that correctness depends on interpretive judgment and corroboration outside the model's logic. The only resolution to this regress is trusted social mechanisms: rigorous peer review, reproducibility, and shared epistemic standards that scientists must establish and train to utilize. No matter how prominent AI becomes in scientific workflows, *AI cannot stand alone because no method has ever stood alone*. Therefore, investment in AI requires investments in human expertise and the social conditions necessary for scientific rigor.

Without transparent methods of validation and community oversight, scientists may defer too readily to outputs that "look plausible" but cannot be fully justified. The risk here is not just error; it is **epistemic distortion**. Predictive accuracy can be mistaken for theoretical insight. As Arvind Narayanan and Sayash Kapoor warn, this can produce "illusions of progress" in the absence of genuine understanding.[3] Complementing this concern, Lisa Messeri and MJ Crockett have argued that reliance on AI systems can foster "illusions of understanding": researchers may overestimate their own comprehension of scientific problems when deferring to systems perceived as more knowledgeable or objective than they truly are.[4]

Current AI-integrated workflows face particular vulnerabilities in this regard, notably reproducibility failures due to subtle but common issues like data leakage, where testing datasets inadvertently influence model training.[5] Such issues result in overly optimistic performance estimates and conclusions that become difficult to replicate in a different scientific work setting. Additionally, widespread use of generative AI tools has led to unreliable "hallucinated" outputs, further exacerbating scientists' skepticism around AI-driven research.[6] This challenge is only intensified by the rapid proliferation of AI-generated scientific papers on databases like Google Scholar, threatening to "overwhelm the scholarly communication system and jeopardize the integrity of the scientific record."[7]

These challenges are only the beginning. Workflows designed around AI's capabilities may inadvertently prioritize data-rich, quantifiable research questions at the expense of qualitative, exploratory, or speculative inquiries. Furthermore, without intentional safeguards, the adoption

[3] Arvind Narayanan and Sayash Kapoor, "Why an Overreliance on AI-Driven Modelling Is Bad for Science," *Nature* 640, no. 8058 (April 2025): 312–14, https://doi.org/10.1038/d41586-025-01067-2.
[4] Lisa Messeri and M. J. Crockett, "Artificial Intelligence and Illusions of Understanding in Scientific Research," *Nature* 627, no. 8002 (March 2024): 49–58, https://doi.org/10.1038/s41586-024-07146-0.
[5] Sayash Kapoor and Arvind Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," *Patterns* 4, no. 9 (September 2023): 100804, https://doi.org/10.1016/j.patter.2023.100804; Henry Han, "Challenges of Reproducible AI in Biomedical Data Science," *BMC Medical Genomics* 18, no. 1 (January 10, 2025): 8, https://doi.org/10.1186/s12920-024-02072-6.
[6] Van Noorden, Richard, and Jeffrey M. Perkel. "AI and Science: What 1,600 Researchers Think." *Nature* 621, no. 7980 (September 27, 2023): 672–75. https://doi.org/10.1038/d41586-023-02980-0.
[7] Haider, Jutta, Kristofer Rolf Söderström, Björn Ekström, and Malte Rödl. "GPT-Fabricated Scientific Papers on Google Scholar: Key Features, Spread, and Implications for Preempting Evidence Manipulation." *Harvard Kennedy School Misinformation Review*, September 3, 2024. https://doi.org/10.37016/mr-2020-156.

of AI may disrupt essential laboratory roles and practices, reconfiguring opportunities for apprenticeship, collaboration, and creative insight. Collectively, these changes risk creating a **scientific monoculture**, privileging certain data-driven methodologies while undermining other equally critical forms of reasoning, experimentation, and interpretation vital to a diverse and robust scientific ecosystem.

Effectively embedding AI in scientific research requires rethinking not just *what* counts as valid evidence, but also *how* its credibility is established. Without continued investments in explicit protocols for interpreting, validating, and contesting AI-generated knowledge, we risk epistemic acceleration that surpasses our capacity for thoughtful deliberation. To proactively address these challenges, scientific workflows integrating AI should adhere to three foundational principles:

- **Interpretability by design**. AI tools must include transparent mechanisms that enable scientists — not just developers — to understand and interrogate their outputs.[8] Interpretability, in this context, means that explanations for AI outputs must be not only technically transparent but also socially actionable.[9] This requires ensuring that AI outputs are responsive to the practical demands scientists face, including contestability, and fit within their own organizational contexts. Thus, interpretability is integral, not peripheral, to the responsible deployment of AI technologies in scientific settings.
- **Social infrastructures for validation**. AI-driven scientific claims must be subject to rigorous, community-based validation processes.[10] Independent replication, peer review, collaborative verification, and open platforms for contestation are essential.
- **Preserving human judgment**. Scientific workflows must enhance, rather than replace, human expertise and critical interpretation.[11] AI systems should serve as collaborative tools that augment human insight, creativity, skepticism, and judgment, to ensure that scientists remain central to the organizational structure of science.

---

[8] An illustrative effort in this direction is the SHAP (SHapley Additive exPlanations) framework, which helps make complex AI models more transparent by turning their predictions into clear, understandable explanations. It estimates how much each input — like age, income, or location — contributed to a decision, using a consistent and theoretically grounded approach. This can potentially help scientists and decision-makers trust and verify the model's behavior. See, Scott Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions" (arXiv, November 25, 2017), https://doi.org/10.48550/arXiv.1705.07874 for methodological insights into the framework.

[9] Upol Ehsan et al., "Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI," *Proc. ACM Hum.-Comput. Interact.* 7, no. CSCW1 (April 16, 2023): 34:1-34:32, https://doi.org/10.1145/3579467.

[10] See our response to Question 4 for further discussion of validation and reproducibility of AI outputs.

[11] Recent critiques by scientists themselves illustrate how AI systems risk displacing — rather than supporting — scientific intuition and creativity. As one researcher commented, "Why would I want to outsource my fun to a computer, and then be left with only the hard work to do myself? In general, many generative AI researchers seem to misunderstand why humans do what they do, and we end up with proposals for products that automate the very part that we get joy from." Such reflections point to the need for workflows that respect and preserve the aspects of research that are driven by curiosity and joy in discovery. Kyle Wiggers, "Experts Don't Think AI Is Ready to Be a 'Co-Scientist,'" TechCrunch, March 5, 2025, https://techcrunch.com/2025/03/05/experts-dont-think-ai-is-ready-to-be-a-co-scientist/.

Faster science is possible, but only if we invest as deeply in the integrity of knowledge production as we do in its acceleration. Redesigning scientific workflows must sustain the social and epistemic conditions that make scientific discovery trustworthy in the first place.

Q4. In order to measure the success of ASAP, we need to have objective metrics that measure the speed of scientific innovation. What metrics already exist and what ones need to be created? What information should the federal government have to understand the health and productivity of our innovation ecosystem, and what tools, processes, or institutions should be used to do so?

Objective metrics for measuring the speed and quality of scientific innovation require deeper reconsideration in the context of integrating AI systems into research workflows. Historically, reproducibility — the capacity to verify results independently — has been central to the credibility and measurement of scientific progress.[12] Yet, as computational processes become more complex and opaque, the expectation of **reproducibility faces profound challenges**.

As discussed earlier, an instructive early example of such challenges is the 1976 computer-assisted proof of the four-color theorem.[13] The four-color theorem posits that no more than four distinct colors are required to color any map in such a way that no two adjacent regions share the same color. Though the statement seems intuitive, it remained unproven for over a century and resisted the best efforts of many of the world's leading mathematicians. The eventual proof relied heavily on a computer to exhaustively check thousands of map configurations, a task impossible for humans to verify manually. Initially, mathematicians were confronted with significant unease precisely because the computational process could not be directly inspected, raising concerns around trustworthiness and reproducibility of the proof. To address these concerns, they adopted rigorous independent replication: multiple teams rewrote and re-executed the computational algorithms on different hardware and software setups, effectively building trust through redundancy and transparency.

---

[12] It is important to note here that Science and Technology Studies (STS) scholarship has consistently challenged the idea that reproducibility is simply a matter of exact replication. Steven Shapin and Simon Schaffer, for instance, argue that establishing credibility for experimental results required trusted witnesses, standardized practices, and institutional validation — demonstrating that reproducibility depends as much on shared norms and social infrastructure as on experimental design. See, Steven Shapin and Simon Schaffer, *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton, NJ: Princeton University Press, 2017); Along similar lines, Samir Passi and Steven J. Jackson, "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects," *Proc. ACM Hum.-Comput. Interact.* 2, no. CSCW (November 2018): 136:1-136:28, https://doi.org/10.1145/3274405 argue that trust in data science is a collaborative accomplishment achieved through situated practices such as algorithmic witnessing — the technical reproduction and iterative scrutiny of model outputs by practitioners.

[13] MacKenzie, "Slaying the Kraken."

Today, however, the reproducibility challenge posed by generative AI, especially large language models (LLMs), is fundamentally different and even more complex. Unlike the deterministic computer algorithms used in the four-color theorem, **LLM outputs are inherently non-deterministic**: *they can vary subtly (or dramatically) with each execution of the same prompt*.[14] This variability arises from the probabilistic nature of model sampling, where multiple plausible outputs can emerge from a single input; it is the source of the value and flexibility of AI, and its most significant weakness. The same feature that allows for generative innovation also poses a serious challenge to scientific reproducibility. Even if another research team replicates the exact experimental setup — same model, same parameters, same prompt — they may produce materially different results each time.

This challenge of non-deterministic outputs has critical implications for measuring scientific innovation. If each replication yields different outcomes, traditional replication-based verification no longer establishes confidence or credibility and metrics such as replication counts or benchmarking become challenging to operationalize.[15] Thus, federal agencies seeking to understand and foster a robust scientific ecosystem need to develop new ways to measure innovation and productivity that explicitly account for the variability inherent in generative AI.

Given this profound shift, new metrics, protocols, and institutions are necessary to understand and validate innovation in AI-augmented science. Several promising approaches could help address these reproducibility challenges:

- **Transparency in prompt and parameter disclosure**. Mandating comprehensive disclosure of all experimental prompts, parameters, and training datasets used in generative models would significantly enhance interpretability. Disclosure practices also create the conditions for epistemic accountability, allowing other researchers to understand how results were produced and assess their credibility.[16] Federal scientific

[14] See, for example, Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

[15] Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, ed. J. Vanschoren and S. Yeung, vol. 1, 2021, https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

[16] Many LLM-related papers now include appendices with the exact prompt templates used for experimental runs (e.g., in few-shot learning, reasoning tasks, or translation). These details — along with temperature settings, number of tokens, and sampling methods — enable other researchers to re-run experiments, compare results across models, and identify sources of variance. This growing norm in academic publishing highlights how prompt and parameter transparency functions as both a scientific method and an accountability mechanism. Xiaoming Liu et al., "StablePT: Towards Stable Prompting for Few-Shot Learning via Input Separation" (arXiv, October 3, 2024), https://doi.org/10.48550/arXiv.2404.19335; Devichand Budagam et al., "Hierarchical Prompting

agencies can lead the way on such norms by requiring that grantees adhere to them and register their outcomes in standardized formats on open repositories.

- **Community-managed repositories[17] of traceable reference outputs**. Rather than attempting to eliminate variability, we should recognize and curate it. Institutions or platforms that steward collections of AI-generated reference outputs along with reasoning traces and prompt histories can serve as epistemic anchors.[18] These repositories would allow scientists to not only compare outputs under controlled conditions but also examine the reasoning pathways models followed to arrive at those outputs, providing additional windows into understanding AI behavior.[19] Federal scientific agencies should fund the creation and maintenance of repositories that support such traceable inference in the face of non-determinism.
- **Multi-stakeholder deliberation and social infrastructure**. Reproducibility in AI-driven contexts may rely less on exact duplication of results and more on community consensus around acceptable variability.[20] We need deliberative spaces — such as peer review boards, expert panels, or reproducibility forums — to interpret and standardize acceptable scientific practices that could foster trust in variable outputs. Federal scientific agencies should fund such convenings, including with international partners.

To ensure that scientific acceleration remains credible and trustworthy, we must design metrics and infrastructures that reflect not only what AI can produce, but what scientific communities can reliably interpret, validate, and build upon. Reproducibility must evolve from a technical benchmark to a collective practice.

Taxonomy: A Universal Evaluation Framework for Large Language Models" (arXiv, June 18, 2024), https://doi.org/10.48550/arXiv.2406.12644.

[17] AlphaFold DB provides predicted protein structures generated by DeepMind's AI models, shared openly with the scientific community. While not non-deterministic in the same way as LLM outputs, this kind of community-accessible repository illustrates the practice of publishing AI-generated outputs in ways that enable downstream reuse, validation, and refinement. A similar model could be adopted for textual or interpretive scientific AI outputs. See: https://alphafold.com/

[18] Projects like BigBench and HELM evaluate LLMs on a wide variety of tasks using consistent input prompts. While their current outputs are primarily used for benchmarking, archiving these responses over time — as reference generations — could create stable comparison points. This could help scientists contextualize their own model behaviors and assess whether surprising outputs are truly novel or reflect known variability. Aarohi Srivastava et al., "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models" (arXiv, June 12, 2023), https://doi.org/10.48550/arXiv.2206.04615; Percy Liang et al., "Holistic Evaluation of Language Models" (arXiv, October 1, 2023), https://doi.org/10.48550/arXiv.2211.09110.

[19] While reasoning traces provide valuable insights into AI decision-making, models may not always be truthful in their reasoning logs, particularly when exhibiting misaligned behaviors. See, for example, Anthropic, "Reasoning Models Don't Always Say What They Think," Anthropic | Alignment, April 3, 2025, https://www.anthropic.com/research/reasoning-models-dont-say-think.

[20] See, for example, the ML Reproducibility Challenge (https://reproml.org/), which invites participants to independently reproduce results from accepted ML papers. Importantly, it doesn't aim to *replicate* results in a rigid sense, but instead creates a deliberative space to discuss why results differ and what counts as a reasonable threshold for "acceptable" variation.

Q6. How can America build the world's most powerful scientific data ecosystem to accelerate American science?

The integration of AI into scientific research builds on — and amplifies — earlier challenges posed by big data. The big data paradigm introduced a potential shift in scientific reasoning: away from asking specific, hypothesis-driven questions and toward finding patterns in large datasets,[21] even when the meaning or reliability of those patterns isn't clear.[22] In the context of AI-assisted science, these concerns become even more acute; it becomes harder for scientists to judge whether AI-generated insights are meaningful, or simply the result of spurious correlations. As we have previously noted, AI systems trained on heterogeneous, poorly documented, or domain-mismatched datasets can produce **highly plausible but epistemically unstable outputs** — outputs that appear scientifically meaningful but may lack evidential grounding or reproducibility.

This introduces three critical vulnerabilities in current data ecosystems:

- **Opacity of evidence generation.** AI systems trained on big data often obscure the link between input data and scientific claims, complicating the evidential status of outputs. AI-assisted scientific research practices can detach results from the conditions of their production, undermining the scientist's ability to interrogate or trace scientific inferences.[23]
- **Loss of contextual metadata.** Scientific validity depends on more than volume: it hinges on understanding how that data was generated. Yet AI models trained at scale often strip away key contextual details — such as who collected the data, using what instruments, under what conditions, and with what assumptions. When this provenance is lost, researchers cannot reliably assess whether a model's outputs are appropriate for

[21] Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2000, https://www.wired.com/2008/06/pb-theory/.
[22] Sabina Leonelli, "Scientific Research and Big Data," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Summer 2020 (Metaphysics Research Lab, Stanford University, 2020), https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.
[23] For example, AI models used in ecological forecasting often fail to show how different input variables — such as land use, temperature, or species traits — shape their predictions. This makes it difficult for scientists to assess whether the results are credible or spurious. To address this, researchers are developing new tools that make AI predictions more transparent and easier to scrutinize, helping restore scientific trust in model-driven forecasts. See, Robin Zbinden et al., "MaskSDM with Shapley Values to Improve Flexibility, Robustness, and Explainability in Species Distribution Modeling" (arXiv, March 17, 2025), https://doi.org/10.48550/arXiv.2503.13057.

their domain.[24] Even plausible results can mislead if built on mismatched or untraceable foundations.

- **Platform dependencies and infrastructural lock-in.** Science relying on big data is frequently dependent on computational infrastructures and standards maintained by a small number of commercial actors. In AI-assisted research, this dependency deepens, as teams must align with the tooling, APIs, and update cycles of proprietary platforms to run or fine-tune models, compromising long-term scientific autonomy.[25]

Addressing these concerns requires treating data infrastructures not as neutral pipelines, but as **epistemic institutions** — embedded systems of expert judgment, authority, and accountability. In practice, this means:

- **Investing in data provenance infrastructure.** Federal agencies should fund tools and standards that ensure data used to train or evaluate AI models in science include detailed provenance, instrumentation details, collection protocols, and curation histories. Data without context cannot support meaningful inference.
- **Supporting epistemically diverse datasets.** Dominant datasets often reflect what is most available, not what is most relevant. Federally funded data commons should prioritize the inclusion of datasets representing under-represented questions, species, populations, and modalities — counteracting the homogenizing effects of "data-intensive" science.
- **Mandating data interoperability and auditability.** Data infrastructures must allow for inspection, comparison, and challenge — not just passive use. Standardizing audit logs, versioning protocols, and open APIs for scientific AI models will support contestability and reduce epistemic lock-in. This means supporting not just their initial development, but also resourcing the **ongoing maintenance of these data**

[24] For example, many public medical imaging datasets — such as those used for training diagnostic AI models — lack consistent metadata about how images were acquired, including details about equipment type, imaging settings, and clinical context. This missing information can lead AI systems to learn spurious patterns or fail to generalize across clinical settings, undermining their scientific and clinical validity. See, Davood Karimi et al., "Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis," *Medical Image Analysis* 65 (October 1, 2020): 101759, https://doi.org/10.1016/j.media.2020.101759.

[25] Critical scholars have shown the structural convergence of artificial intelligence with Big Tech, demonstrating how companies like Amazon, Microsoft, and Google dominate the infrastructure, resources, and investments underlying the industrialization of AI. For example, in Fernando van der Vlist, Anne Helmond, and Fabian Ferrari, "Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence," *Big Data & Society* 11, no. 1 (March 1, 2024): 1–16, https://doi.org/10.1177/20539517241232630, the authors use a technographic approach to show that AI's deployment at scale is inseparable from these firms' cloud platforms, creating dependencies and lock-in effects that shape who can build, access, and benefit from AI across sectors.

infrastructures.[26] Without sustained investment in maintenance, scientific infrastructures become brittle over time, undermining the long-term accessibility, auditability, and relevance of research outputs.

In sum, the promise of AI-assisted discovery cannot be realized without parallel investments in **scientifically meaningful, socially governed data ecosystems**. Big data infrastructures mutually shape what questions can be asked, what evidence counts, and who gets to know. As such, they must be designed as part of science policy, and not outsourced to proprietary platforms or abstracted into technical backends.

Conclusion

The acceleration of American science must be matched by an equally substantive investment in the epistemic foundations that make science credible, accountable, and contestable.[27] As this response makes clear, the integration of AI into research workflows is not simply a technical upgrade; it is a transformation of how knowledge is produced, validated, and understood. To ensure that acceleration enhances rather than erodes the integrity of science, policy must center interpretability, provenance, human judgment, and institutional safeguards for trust.

To realize this goal, federal agencies should focus on three strategic priorities, and make the appropriate investments to fulfill them:

1. **Design AI systems for interpretability and scientific accountability.**
   Support the development and deployment of AI tools that make their reasoning transparent and contestable — not just to developers, but to the scientists who use them.
2. **Strengthen the social infrastructure of validation.**
   Invest in institutions — peer review, reproducibility forums, and open repositories — that translate AI-generated results into trusted scientific knowledge.
3. **Build public data ecosystems for epistemic diversity and auditability.**
   Treat data infrastructure as a public good by funding provenance-rich, interoperable, and audit-ready datasets that reflect diverse scientific domains and reduce dependency on proprietary platforms. This includes sustained support for the **maintenance and updating** of public datasets and metadata systems to ensure they remain accessible, trustworthy, and relevant over time.

[26] For a detailed account of the work that goes into maintaining data infrastructures, see Steven J. Jackson et al., "Maintaining Data Infrastructures," in *The Sage Handbook of Data and Society*, ed. Tommaso Venturini et al. (Thousand Oaks, CA: Sage Publishing, 2025), 19–35, https://sk.sagepub.com/hnbk/edvol/the-sage-handbook-of-data-and-society/chpt/2-maintaining-data-infrastructures.
[27] Alondra Nelson, "Ten Times Faster Is Not 10 Times Better," *Science* 388, no. 6754 (June 26, 2025): 1353–1353, https://doi.org/10.1126/science.adz9545.

Cutting-edge advancements in AI technology are playing an important role in the scientific process, but to meaningfully lead in scientific innovation, the federal government must also focus on building the social infrastructures that allow science to remain trustworthy and responsive in an AI-driven research landscape.

Respectfully submitted,

Ranjit Singh, PhD
AI on the Ground, Data & Society

With contributions from:
Jacob Metcalf, PhD
AI on the Ground, Data & Society

References:

Anderson, Chris. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired*, June 23, 2000. https://www.wired.com/2008/06/pb-theory/.

Anthropic. "Reasoning Models Don't Always Say What They Think." Anthropic | Alignment, April 3, 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Budagam, Devichand, Sankalp KJ, Ashutosh Kumar, Vinija Jain, and Aman Chadha. "Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models." arXiv, June 18, 2024. https://doi.org/10.48550/arXiv.2406.12644.

Collins, Harry. *Changing Order: Replication and Induction in Scientific Practice*. London: Sage, 1985.

Ehsan, Upol, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. "Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI." *Proc. ACM Hum.-Comput. Interact.* 7, no. CSCW1 (April 16, 2023): 34:1-34:32. https://doi.org/10.1145/3579467.

Han, Henry. "Challenges of Reproducible AI in Biomedical Data Science." *BMC Medical Genomics* 18, no. 1 (January 10, 2025): 8. https://doi.org/10.1186/s12920-024-02072-6.

Jackson, Steven J., Jen Liu, Ranjit Singh, and Samir Passi. "Maintaining Data Infrastructures." In *The Sage Handbook of Data and Society*, edited by Tommaso Venturini, Amelia Acker, Jean-Christophe Plantin, and Tone Walford, 19–35. Thousand Oaks, CA: Sage Publishing, 2025. https://sk.sagepub.com/hnbk/edvol/the-sage-handbook-of-data-and-society/chpt/2-maintaining-data-infrastructures.

Kapoor, Sayash, and Arvind Narayanan. "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science." *Patterns* 4, no. 9 (September 2023): 100804. https://doi.org/10.1016/j.patter.2023.100804.

Karimi, Davood, Haoran Dou, Simon K. Warfield, and Ali Gholipour. "Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis." *Medical Image Analysis* 65 (October 1, 2020): 101759. https://doi.org/10.1016/j.media.2020.101759.

Leonelli, Sabina. "Scientific Research and Big Data." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University, 2020. https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.

Liang, Percy, Rishi Bommasani, Tony Lee, et al. "Holistic Evaluation of Language Models." arXiv, October 1, 2023. https://doi.org/10.48550/arXiv.2211.09110.

Liu, Xiaoming, Chen Liu, Zhaohan Zhang, et. al. "StablePT: Towards Stable Prompting for Few-Shot Learning via Input Separation." arXiv, October 3, 2024. https://doi.org/10.48550/arXiv.2404.19335.

Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." arXiv, November 25, 2017. https://doi.org/10.48550/arXiv.1705.07874.

MacKenzie, Donald. "Slaying the Kraken: The Sociohistory of a Mathematical Proof." *Social Studies of Science* 29, no. 1 (February 1, 1999): 7–60. https://doi.org/10.1177/030631299029001002.

Messeri, Lisa, and M. J. Crockett. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature* 627, no. 8002 (March 2024): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Narayanan, Arvind, and Sayash Kapoor. "Why an Overreliance on AI-Driven Modelling Is Bad for Science." *Nature* 640, no. 8058 (April 2025): 312–14. https://doi.org/10.1038/d41586-025-01067-2.

Nelson, Alondra. "Ten Times Faster Is Not 10 Times Better." *Science* 388, no. 6754 (June 26, 2025): 1353–1353. https://doi.org/10.1126/science.adz9545.

Passi, Samir, and Steven J. Jackson. "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects." *Proc. ACM Hum.-Comput. Interact.* 2, no. CSCW (November 2018): 136:1-136:28. https://doi.org/10.1145/3274405.

Raji, Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. "AI and the Everything in the Whole Wide World Benchmark." In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1, 2021. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Shapin, Steven, and Simon Schaffer. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton, NJ: Princeton University Press, 2017.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, et al. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." arXiv, June 12, 2023. https://doi.org/10.48550/arXiv.2206.04615.

Vlist, Fernando van der, Anne Helmond, and Fabian Ferrari. "Big AI: Cloud Infrastructure Dependence and the Industrialisation of Artificial Intelligence." *Big Data & Society* 11, no. 1 (March 1, 2024): 1–16. https://doi.org/10.1177/20539517241232630.

Wiggers, Kyle. "Experts Don't Think AI Is Ready to Be a 'Co-Scientist.'" TechCrunch, March 5, 2025. https://techcrunch.com/2025/03/05/experts-dont-think-ai-is-ready-to-be-a-co-scientist/.

Zbinden, Robin, Nina van Tiel, Gencer Sumbul, et. al. "MaskSDM with Shapley Values to Improve Flexibility, Robustness, and Explainability in Species Distribution Modeling." arXiv, March 17, 2025. https://doi.org/10.48550/arXiv.2503.13057.