May 29, 2025

Faisal D'Souza
National Coordination Office
Networking and Information Technology Research and Development
215 Eisenhower Avenue
Alexandria, VA 22314

Dear Mr. D'Souza,

On behalf of the Center for Data Innovation (datainnovation.org), I am pleased to submit this response to the Office of Science and Technology Policy's (OSTP) and Networking and Information Technology Research and Development's (NITRD) for comments on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan.[1]

The Center for Data Innovation studies the intersection of data, technology, and public policy. The Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as open data, artificial intelligence, and the Internet of Things. The Center is part of the Information Technology and Innovation Foundation (ITIF), a nonprofit, nonpartisan think tank.

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the strategic plan and associated documents without attribution.

Yours sincerely,

Hodan Omaar
Senior Policy Manager
ITIF's Center for Data Innovation

[1] Federal Register, "Request for Information on the Development of a 2025 National AI R&D Strategic Plan," April 29, 2025, https://www.federalregister.gov/documents/2025/04/29/2025-07332/request-for-information-on-the-development-of-a-2025-national-artificial-intelligence-ai-research.

EXECUTIVE SUMMARY

We appreciate the administration's commitment to strengthening America's innovation capacity in AI and strongly support OSTP and NITRD's efforts. The National AI R&D Strategic Plan was first launched under President Obama, updated under President Trump, and expanded under President Biden, and has always reflected the priorities of the administration. A second Trump term should now refocus it on delivering on the administration's pledge to make AI work for the American people. In this submission, we outline three recommendations to that end:

1. Make unlocking AI the guiding goal of federal AI R&D
2. Prioritize research that links technical design to performance outcome of AI systems
3. Invest in research to generate more and better data for AI.

1. MAKE UNLOCKING AI THE GUIDING GOAL OF FEDERAL AI R&D

The current Strategic Plan spans a broad and appropriate set of topics for R&D, but across many of its nine strategies, a common thread emerges: risk reduction. While the plan does include efforts to advance AI capabilities, its overall posture reads as one primarily focused on preventing harm. The implicit message is that if risks are managed, benefits will naturally follow. But that assumption doesn't hold. Making AI safer does not guarantee it will be usable, scalable, or adopted where it's needed most. A posture centered on harm prevention is not the same as one designed to enable public benefit.

If the United States wants to drive innovation, strengthen the economy, and improve public services through AI, its R&D strategy should read as a roadmap for unlocking AI's potential. That means revisiting each of the plan's nine strategies to ensure they are guided not only by the need to avoid failure, but by the goal of enabling success.

Even European governments, which are known for precautionary AI stances, are re-writing their research playbooks to unlock AI. In France, the national research institute Inria is ensuring that public R&D is aligned with the country's goal of becoming an AI powerhouse.[2] In Germany, a new high-tech ministry is being created to take over research from the education ministry and technology and aerospace from the economics ministry in an effort to better leverage research for industrial competitiveness and strategic advantage.[3] And at the EU level, the European Commission has a

[2] Staff Writer, "France Aims for AI Leadership: A Look at the Nation's Ambitions and Challenges," *Elnion*, June 1, 2024, https://elnion.com/2024/06/01/france-aims-for-ai-leadership-a-look-at-the-nations-ambitions-and-challenges.
[3] Gretchen Vogel, "Germany to create 'super–high-tech ministry' for research, technology, and aerospace," *Science*, April 11, 2025, https://www.science.org/content/article/germany-creates-super-high-tech-ministry-research-technology-and-aerospace.

forthcoming Strategy for AI in Science.[4] While not a national R&D plan in the same sense, it is an effort to coordinate research investments around a clear objective: accelerating the uptake of AI across scientific domains, especially in mission-critical areas like health, climate, and clean tech.

The lesson is clear: economies competing in AI are repositioning public R&D as a catalyst for deployment and competitiveness. The United States should ensure it does the same—re-examining each of its nine strategies through that lens so federal investments do more than avert harm; they actively unlock AI's value in every sector that matters.

2. PRIORITIZE RESEARCH THAT LINKS TECHNICAL DESIGN TO PERFORMANCE OUTCOMES

For AI systems to be adopted effectively, organizations need research that links the outcomes they care about, such as fairness, reliability, or security, with the technical features that are most likely to produce those results.[5]

To see why, consider the Veterans Health Administration, which is piloting a tool called Pingoo AI to give diabetic veterans up-to-date health information using a retrieval-augmented generation (RAG) model.[6] The first thing the VA needs to do is decide what kind of performance outcome matters most in this context. Fortunately, the National Institute of Standards and Technology (NIST) has done important groundwork to define the key characteristics of trustworthy AI, including reliability, safety, privacy, fairness, and explainability. Using that menu, the VA might identify validity and reliability as the outcomes that matter most for Pingoo AI.

At that point, the VA would turn to the technical literature to understand what contributes to validity and reliability in a system like this. It would find a wide range of methods that researchers have identified for addressing different aspects of those goals—such as improving document retrieval, adjusting source ranking, filtering hallucinations, flagging uncertainty, and more. But while all of these techniques may contribute to better performance, there is little research that helps determine which of them matter most in practice—or how to weigh their relative importance in a system like Pingoo AI.

---

[4] European Commission, "Artificial Intelligence (AI) in Science," accessed May 28, 2025, https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/artificial-intelligence-ai-science_en.
[5] Hodan Omaar, "Three Steps Trump Should Take to Advance Government AI Adoption," (Center for Data Innovation, April 2025), https://datainnovation.org/2025/04/three-steps-trump-can-take-to-advance-government-ai.
[6] Pingoo.AI, "Department of Veterans Affairs Chooses Pingoo.AI to Enhance Diabetes Education and Engagement for Veterans," press release, December 3, 2024, https://www.pingoo.ai/news/department-of-veterans-affairs-chooses-pingooai-to-enhance-diabetes-education-and-engagement-for-veterans.

This is the gap the National AI R&D Strategic Plan should address: identifying how different technical parameters map to measurable AI performance outcomes. By helping organizations understand which design choices are most likely to achieve their stated goals, the administration can support more effective and evidence-based AI adoption.

3. INVEST IN RESEARCH TO GENERATE MORE AND BETTER DATA FOR AI

The National AI R&D Strategic Plan rightly underscores the importance of data for AI but it treats data as something that already exists and merely needs to be unlocked, aggregated, or shared. In practice, AI innovation today is increasingly constrained not by inaccessible data, but by data that doesn't exist in the first place, and where data does exist, it often lacks the quality, representativeness, or interoperability required for high-impact AI applications. To realize AI's full potential, the next iteration of the R&D strategy should move beyond data access and actively invest in research for data generation—both more data and better data—as a core enabler of progress.

First, the United States should invest research to understand when, how, and in what proportions to mix synthetic and real data safely. A 2024 research paper from Epoch AI estimates that the supply of high-quality public web text could be used up for training AI models between 2026 and 2032.[7] In other words, AI developers will have scraped and reused just about all the useful human-written content on the Internet. This could happen even faster if companies keep "overtraining" their models, meaning feeding them more data than is needed for optimal performance during training. To work around this, many AI developers are now exploring the use of synthetic data—text or images generated by AI models themselves—to fill in the gaps. But using synthetic data effectively means overcoming three technical problems. One is that synthetic data tends to miss rare or unusual cases, because models generate outputs based on the most common patterns they've seen before. Without targeted research on how to preserve the "tails" of real-world distributions, AI systems will become less accurate in high-stakes or edge-case scenarios. Another is that synthetic content is prone to hallucination: large language models can produce factually incorrect or subtly distorted information, and when that becomes training data, the model may compound those errors over time. Research is needed to develop automated methods for detecting and filtering out hallucinated content before it pollutes the training set. Finally, we still don't fully understand how synthetic and real data interact during training. Early research shows that replacing real data entirely causes model collapse—a steep drop in performance as the model loses its grounding in real-world patterns but blending the

[7] Pablo Villalobos et al., "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data," *Epoch AI*, June 6, 2024, https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

two can preserve quality, if done carefully.[8] More work is needed to understand when, how, and in what proportions to mix synthetic and real data safely.

Research into synthetic data is especially promising for AI applications in healthcare.[9] Unlike public web text, real healthcare data is often sensitive, fragmented, or legally restricted. Synthetic data offers a way to simulate patient records or clinical scenarios without exposing private information. It can also be used to generate more training examples for rare conditions, where real-world data is sparse. Early studies have shown that synthetically augmented datasets can improve diagnostic performance for underrepresented groups, but only when the synthetic data is high quality and well-matched to real clinical patterns. Investing in this area could not only help extend the life of AI training pipelines, it could also accelerate progress in medical AI while protecting patient privacy.

Second, the United States should invest in research that enables better data collection in areas where information is missing, fragmented, or systematically poor. In sectors like education, infrastructure, and public health, the data needed to build reliable AI systems often doesn't exist in usable form. These gaps fuel the data divide: the social and economic inequalities that stem from uneven data collection, limited interoperability, and poor data quality.[10] To close that divide, the federal R&D strategy should focus on developing the tools, methods, and frameworks needed to generate high-quality, fit-for-purpose data in underrepresented domains. That includes supporting research into novel data collection techniques; methods for building and sustaining data partnerships, especially in commercially neglected areas; improving standards for data quality, interoperability, and documentation; and international research collaborations.

[8] AI Index Report 2025, Stanford Institute for Human-Centered Artificial Intelligence, April 7, 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report
[9] Ibid.
[10] Gillian Diebold, "Closing the Data Divide for a More Equitable U.S. Digital Economy," (Center for Data Innovation, August 2022), https://datainnovation.org/2022/08/closing-the-data-divide-for-a-more-equitable-u-s-digital-economy/