Federal Register Notice 86 FR 46278, https://www.federalregister.gov/documents/2021/08/18/2021-17737/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence, October 1, 2021.

Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource: Responses

RFI Response: National AI Research Resource

Input from Argonne National Laboratory

October 1, 2021

Contact: Ian Foster, [redacted]

Introduction

Artificial Intelligence (AI) research at Argonne National Laboratory dates back to 1963, when logician Larry Wos first started to explore the use of computers to prove mathematical theorems. Today, hundreds of Argonne scientists and engineers are engaged in the development and application of AI methods to scientific problems, as documented for example in a recent Department of Energy (DOE) AI for Science report[1]. This experience has convinced us that the opportunities for application of AI methods to scientific problems are enormous and as yet largely untapped. We are thus delighted to respond to this RFI relating to a proposed National AI Research Resource (NAIRR).

The RFI requests input on many different aspects of the proposed NAIRR. We focus our response on a few areas where we believe that our experience at the intersection of AI and science allow us to make particularly pertinent and actionable recommendations

Some of the issues discussed here have previously been discussed in a Computing Community Consortium white paper[2].

Q1: Options to Consider

A. Goals

We propose the following goals for a NAIRR.

- Provide researchers in academia, national laboratories, and industry access to appropriately large and diverse **computing resources** to support large-scale research into both new AI methods and the innovative applications that are enabled by these new methods.
- Assemble **large, curated datasets** that, in a manner akin to ImageNet for image classification, advance the state of the art in developing AI methods and models across diverse scientific application areas.

[1] AI for Science, https://www.anl.gov/ai-for-science-report

[2] A National Discovery Cloud: Preparing the US for Global Competitiveness in the New Era of 21st Century Digital Transformation, https://arxiv.org/abs/2104.06953

- Provide a competitive, nationwide, **peer-review process** for selecting which projects have access to these resources.

- Establish a high-quality, nationally recognized process and machinery for the systematic and sustained **intercomparison** over time of DL-based solutions to important scientific and engineering problems.

In addition to the metrics for success often used to evaluate the impacts of scientific facilities, we propose the following:

- Number of substantial AI-ready datasets assembled, and the use made of those datasets.

- Number and diversity of students and junior researchers introduced to AI methods via NAIRR.

B. Ownership and Administration

The NAIRR requires an operator with deep experience and expertise, and a deep bench of talented staff, in establishing and operating large facilities with large and diverse user communities. These factors suggest that the **Department of Energy could be the appropriate agency responsible for the implementation, deployment and administration of the NAIRR**. The DOE already has extensive, multi-decade experience in running large-scale high-performance computing centers for the national research community. More broadly, DOE has a substantial and proven experience with a sophisticated project management structure (governed by DOE Order 413.3b) that spans the lifetime of large scientific projects, including experimental facilities, large-scale computing facilities, and more recently the Exascale Computing Project. This process has been well tested and tuned over time and could be applied to run the NAIRR successfully, with appropriate controls to ensure on-time delivery within approved budget and then transition to continued successful operations. We also note that colocation with existing facilities (e.g., supercomputers) will enable new scientific methodologies such as AI-driven simulations.

D. Infrastructure capabilities

While different AI applications necessarily have varying infrastructure requirements, we wish to emphasize the needs of applications that combine AI methods for computational simulation—as when, for example, AI methods are used to guide the choice of the next simulation or to accelerate simulation kernels, or when computational simulation is used to generate training data for AI models. The growing importance of these methods in many areas of science (e.g., climate change, drug design, materials design) suggests a need for infrastructure that can **support close coupling between AI and HPC**.

E. Government Data Sets

The US government maintains many datasets of high scientific value that are, however, not easily usable by AI researchers. The reasons for this situation are varied and no single solution can overcome all obstacles. However, the NAIRR can make a substantial contribution to science by establishing methods to allow **co-location of government datasets with suitably powerful AI computers**. To give just one example, environmental researchers can, today, easily retrieve individual files from satellite imagery datasets maintained by NASA, but cannot readily train neural network models on the entirety of these large (sometimes petabyte-scale) datasets. Deployment of these datasets on storage systems colocated with AI supercomputers would allow large-scale application of AI methods to addressing important environmental questions. For other important datasets, the establishment of secure data enclaves co-located with AI supercomputers will be important for progress.

Q2: Capabilities and Services to be Prioritized

We propose two specific capabilities that we believe will have a disproportionately large impact on research in AI methods and applications.

Implement Methods for DL Model Comparison, Evaluation, and Improvement

As is well known, what gets measured gets improved. As demonstrated, for example, via the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and the Intergovernmental Panel on Climate Change (IPCCs) Climate Model Intercomparison Project (CMIP), the establishment of codified processes, supported by appropriate tools, for comparing different approaches to a scientific or engineering question can drive great technical and scientific progress. The NAIRR can achieve broad impact on both AI research and the sciences by catalyzing and supporting similar such initiatives in areas of national priority.

To this end, we argue that NAIRR should incorporate a program that would encompass four overarching goals:

1. Development of a **general computational framework for DL model comparison, evaluation, and improvement**. This framework can include, for example, mechanisms for submitting, training, and (re)running many DL models on challenge problems, and for storing and analyzing results.

2. The **application of this framework** across specific priority research areas in close partnership with the domain science communities (e.g., cancer biology, materials discovery, synthetic biology, climate change impacts research, surrogate models).

3. **Data generation or acquisition** aimed at improving DL models in priority areas, a topic that will also couple well with research efforts in autonomous discovery. (See also below)

4. The development and incorporation into this framework of methods to improve the **interpretability, explainability, and fairness** of DL models—so that, for example, when evaluating DL models of climate change impacts, scoring metrics are evaluated with respect to their impacts on disadvantaged populations.

By promoting rigorous, data-driven comparison of alternative approaches, and enabling the systematic study of DL model effectiveness in real-world settings, this initiative will contribute to broad advances in AI methods and applications.

Create Large, AI-ready Datasets Relevant to National Priorities

The development of certain large, curated datasets such as ImageNet have been an important factor in recent AI advances. Similarly large datasets are needed to advance AI in areas of national priority, from climate science to healthcare and materials discovery. Significant effort will be needed to make such datasets AI-ready, especially with providing relevant metadata pertaining to issues such as data provenance, quality, and completeness. In many of the priority areas, sufficiently large datasets either do not exist (e.g., in materials science, despite good efforts such as the Materials Project and Materials Data Facility) and/or are not accessible in forms that permit application of AI methods (e.g., as noted above, climate science). New, substantial efforts are required to create new datasets suitable for AI research, and to deploy those datasets in ways that permit large-scale AI model training and inference.

The creation of such datasets will both drive advances in AI, by enabling AI researchers to experiment with existing and new methods in new contexts, and benefit the disciplines in which the datasets are created.

This activity will be expensive, and thus targets must be carefully chosen, with extensive community consultation. The process by which datasets are created or acquired will also require careful design and management to avoid (or at least characterize) bias.

Support the Creation of AI-for-Science Software

Modern AI benefits greatly from enormous investments by industry in powerful AI libraries and frameworks. However, these systems are inevitable designed to meet industry needs, and those needs do not always align with the needs of scientific communities. In areas where suitable software is lacking, support for its creation and maintenance will be important.

An important lesson from other scientific domains is that while graduate students can do exceptional work in exploring new methods, software produced by graduate students is rarely suitable for broad use. Building high-quality, broadly usable scientific software is hard, and requires people with specialized software engineering skills. Career paths for such people rarely exist in academic settings, but the national laboratories have a long history of constructing high-quality scientific software (e.g., mathematics libraries). Thus the construction of AI-for-science software could well be a task to be assigned to the national laboratories.

Q3: Reinforcing Ethical and Responsible AI R&D

The need for action relating to ethical and responsible AI R&D is clear. Others will surely speak to the need for robust education initiatives and for efforts aimed at creating AI-ready datasets, such as those envisioned by the recent NIH Bridge2AI program. We wish to emphasize the importance of robust and transparent processes for capturing how data are produced and curated, and models generated and shared. A useful step in this direction could be to extend the Data Management Plan that researchers must today provide with research proposals to encompass descriptions of how researchers propose to make data available with machine-readable labels or metadata.

Q4: Building Blocks

DOE national laboratories have for decades run large-scale **high-performance computing facilities** for use by the national research community, e.g., Argonne Leadership Computing Facility (ALCF), Oak Ridge Leadership Computing Facility (OLCF), and National Energy Research Scientific Computing Center (NERSC). These facilities run leading-edge supercomputers for scientific simulations, and projects are selected from an open, competitive, nationwide peer-review process. Many of these scientific applications are already starting to use AI methods, and the next-generation systems being deployed at present have been designed to support such applications and workloads, with for example 10,000s of GPUs. In other words, examples already exist of large-scale computing facilities at DOE laboratories being used for AI in science (although not exclusively). These facilities can play an important role in the NAIRR.

The DOE national laboratories also have a long history of deploying and providing community access to innovative computer systems: see for example, Argonne's Advanced Computing Research Facility[3] in the late 1980s and the ALCF AI Testbed[4] today. The national laboratories would be well

[3] https://digital.library.unt.edu/ark:/67531/metadc283049/

[4] https://ai.alcf.anl.gov

positioned to operate a **National AI Accelerator Laboratory** in support of NAIRR goals.

DOE national laboratories are also stewards of **important scientific datasets** that, when linked with AI computing infrastructure, can provide AI researchers with new challenges and engage them in advancing scientific goals. (To give just two examples, atmospheric radiation monitoring data in the ARM data center[5], climate simulation datasets in the Earth System Grid Federation[6].)

Q5: Role of Public-Private Partnerships

DOE national laboratories provide **exemplars of successful public-private partnerships in deploying leading-edge computing facilities**. Large supercomputing systems deployed at any DOE computing facility involve a multi-year partnership between the laboratory and the system vendor (in some cases, multiple vendors). Vendors are explicitly funded as part of the procurement to perform non-recurring engineering (NRE) activities to meet the ambitious goals of the system. These NRE components are often essential to ensure that the technologies needed to deliver the system are available in the timeframe needed for the system deployment. In other words, NRE accelerates the availability of new technologies in a vendor's product. Furthermore, DOE has a history of separately funding vendors to develop new technologies independent of any specific procurement. Examples of such programs include FastForward, DesignForward, and PathForward, each which funded multiple vendors to accelerate their technology roadmaps to enable exascale computing. As a result, three exascale systems will soon be available at Oak Ridge, Argonne, and Lawrence Livermore National Laboratories.

Q6: Possible Limitations to Democratization

Design resource allocation policies to promote democratization

Methods for allocating scarce scientific resources that are based purely on merit-based peer review can be exclusionary due to winner-takes-all effects. Experience with national scientific facilities such as supercomputers emphasize the importance of also implementing policies designed to combat that effort. Specifically:

1. Set aside a substantial chunk of the resources to explicitly address fairness issues in access. Thus, for example, any researcher or student from an educational institution might be able to obtain some limited "startup" allocation.
2. Also allocate substantial resources to provide help to those who need it. Online help, training courses, etc.
3. And support a substantial outreach program to engage institutions and communities who are not currently active in the use of the resource.

The latter two activities, in particular, are not cheap (it takes much more human labor to support 10,000 rather than 100 users), but are essential if the NAIRR is to engage a broad community.

[5] https://www.arm.gov/data/

[6] https://esgf.llnl.gov