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

MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

October 1, 2021

Response to "Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource", RFI 86 FR 39081 (Document number 2021-1566)

As research computing professionals working at member institutions of the Massachusetts Green High Performance Computing Center (MGHPCC), we are pleased to be able to provide comments in response to the above-referenced RFI issued by the National Science Foundation and White House Office of Science & Technology Policy. Our responses here are being made collectively as individuals[1] with deep experience with research computing infrastructure who collaborate on MGHPCC activities, rather than as institutional responses by our universities or the MGHPCC itself.

Q1: What options should the Task Force consider for any of roadmap elements A through I

Q1.A Goals for establishment and sustainment of a NAIRR

The NAIRR should provide research infrastructure for both *foundational AI* (developing AI theory and methods that are independent of any particular application domain), and *use-inspired AI* research in specific application domains. Such use-inspired research should go beyond simply applying existing AI techniques and add new knowledge and understanding in both foundational AI and use-inspired domains. The virtuous cycle that exists between foundational and use-inspired AI research is a core tenet of NSF's National AI Research Institutes [2].

The scope of the NAIRR should be carefully defined, with a clear distinction between funding for development and use of research infrastructure and services (such as the NAIRR) and funding for other aspects of AI research (e.g., faculty, research assistant, staff time; travel). Funding for non-research-infrastructure AI research already exists in numerous Federal agencies and should not be the goal of the NAIRR. The budget for NAIRR should supplement, but not be drawn from, existing agency AI research budgets.

Q1.B A plan for ownership and administration of the National Artificial Intelligence Research Resource

A "non-siloed" infrastructure. The current national computing research infrastructure – funded primarily by the DOE and NSF, but also by NOAA, NIH and other agencies are rather stove-piped systems, with agency-funded resources primarily targeted towards agency-sponsored researchers. *As a national resource targeted at research in a specific area (rather than research funded by a particular agency), the NAIRR should not be partitioned or siloed according to funding agency.*

A1.C A model for governance and oversight to establish strategic direction, make programmatic decisions, and manage the allocation of resources

Collaborative, multi-agency oversight will be needed for NAIRR. Multiple agencies should have an oversight and governance role within the NAIRR. Agencies with significant investment in AI research (e.g., NSF, DOE, NIH, NIST, NOAA) might be considered as lead agencies in a governance structure. However, since AI techniques are critical in so many application areas, other agencies must also have a seat at the table, in at least an advisory role.

A community-based technical advisory committee. Given the NAIRR will exist to provide critical infrastructure services to the research community, a research-community-based advisory committee should advise the NAIRR on technical and policy issues and direction. In addition, since the results of AI research will impact individuals and organizations far beyond the AI research community, civil society should also participate on this advisory committee.

NAIRR Resource Allocation Committee (NRAC). The NAIRR will certainly be oversubscribed, with resource demands exceeding resource capacity. An allocation process will be needed that grants access based on an assessment of the relative readiness and appropriateness of allocation requests. There are several national models for providing such resource allocation, including NSF's Large Resource Allocations Community (LRAC) and XSEDE Resource Allocations Committee (XRAC), and similar allocation mechanisms for resources funded via the U.S. Department of Energy (e.g., INCITE, NERSC). These merit-based review mechanisms have generally served the community well. One significant shortcoming of the NSF process, however, is that decisions regarding funding for basic research proposals that require HPC resources are made *independently* of decisions regarding the allocation of research computing resources. These two decisions should be coupled, allowing the merit and cost of a research project to be considered as a whole. Certain NASA programs and NSF's CloudBank program can serve as models here, awarding computing resource credits at the time when a research project is reviewed and awarded.

In addition to NAIRR NRAC allocations for specific research projects, NAIRR allocations (and supporting services) will also be needed for educational purposes, for developmental and exploratory projects, and for novel meritorious projects that demonstrate the need and appropriateness for NAIRR resources. In the case of NSF-funded computational resources, a certain percentage of that resource (e.g., 20%) is set-aside for such uses. A community-based committee, perhaps part of the NRAC or perhaps separate, with a makeup that broadly reflects the research community can advise on these allocations and ensure democratized access to NAIRR's AI R&D capabilities.

2. Which capabilities and services (see, for example, item D above) provided through the NAIRR should be prioritized?

NAIRR resources and the public cloud. Whenever possible, NAIRR should leverage commercial cloud offerings, and only develop its own specialized computing resources following a detailed and expansive/inclusive cost/benefit analysis. A 2018 NSF workshop report on *Enabling Computer and Information Science and Engineering Research and Education in the Cloud* [3] provides a thoughtful discussion of the possible advantages and disadvantages of this approach. NSF's Cloudbank project and NIH STRIDES program are important existing programs piloting the notion of cloud-based computation and data resources for the research community (including AI researchers). A valuable question to consider is *can the NAIRR developers envision an NAIRR hosted entirely in commercial clouds - with appropriate high-level support and tailored interfaces for seamless and straightforward use by a wide range of end-users and organizations?*

NAIRR resources should also be able to interoperate seamlessly with campus-based research infrastructure, allowing researchers to migrate code and data among campus-based, NAIRR-based, and commercial-cloud-based services.

The "data resources'' hosted by NAIRR should be broadly defined to include higher-level, "synthesized" data products (e.g., knowledge graphs; see "Open Knowledge Network" NITRD Big Data IWG workshop report, November 2018[4]) and related toolsets. These capabilities are often only available at scale within industry settings; similar open resources and tools will serve to broaden and democratize AI research.

To support the rapidly growing volume of data resources that are envisions to be offered by NAIRR, there will need to be a renewed and increased focus on leveraging new and novel storage technologies and architectures to support this pace. AI workloads are inherently IO-heavy, so will require new architecture and hardware/software solutions to deliver these resources at scale.

Community code (open source, freeware, and purchased) that enhances and manipulates datasets, and open models built on such datasets should be considered part of the NAIRR. The NAIRR should also consider hosting a "model commons" similar to NIH's notion of a "data commons".

People. Raw computational capabilities and datasets are a necessary, but far from sufficient, set of resources to meet NAIRR goals. Perhaps most importantly, people-centered services - educational, outreach and training activities, expert consulting, and distributed teams of local, sufficiently resourced, "champions" (perhaps similar to the "campus champions" program that was started by XSEDE and has evolved into community of interest that shares information frequently and widely) - will be key components of an accessible NAIRR that can serve a broad and diverse set of AI researchers.

4. What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?

A cloud-backed NAIRR provides an opportunity for public private partnership in which the physical infrastructure is provided by commercial cloud vendors. In this case, the true value-added by the NAIRR is the coordination and funding for resource alignment, access and use, and myriad services layered on top of the physical infrastructure – similar to the approach being taken by Cloudbank, and STRIDES. Multiple cloud providers have collaborated to provide resources to researchers funded under the NSF BIGDATA and Data Hubs programs; commercial cloud services have also hosted NOAA and NASA datasets.

There are a wide array of specific activities that could contribute to a NAIRR network and that do exist today. The NAIRR planners may want to consider strategically collaborating with, strengthening and leveraging some/all of these entities, rather than creating a duplicate structure. As an illustrative example we list below a non-exhaustive set of projects/activities that a NAIRR effort could help bring into a more coordinated whole.

1. All commercial cloud providers have public dataset programs that are curated and organized collections with practical value for applied AI. The Azure "Planetary Computer" initiative is one (of many) examples of the sort of remarkable resources available this way.

2. Commercial cloud providers also support low-powered, free to use compute and software resources (for example Google colab) that can be powerful platforms for basic educational infrastructure.

3. Commercial cloud providers also support access to large scale compute resources and to ad-hoc data storage for a fee.

All these commercial resources are a tremendous potential resource to leverage in some way. There are some caveats to these resources that a NAIRR could try and improve.

- Public datasets (that are freely hosted) are selected by cloud providers and held under terms that are set by providers. The eligibility and storage duration of a potential public dataset is not guaranteed.
- Datasets are subject to metered egress charges between providers and/or between providers and external networks (including academic networks). No provider has a mechanism to uncategorically waive all metered egress charges.
- Cloud-based compute costs can be sub-optimal for some modes of use, an issue that presents challenges for both academic and industry users[5].
- Efficient use of commercial cloud resources requires cost engineering and financial management skills that (today) rarely the academic department-level, where researchers are accustomed to submitting allocation proposals for resources that are operated by agencies such as NSF, DOE, NIH, or by campus research computing groups.

- The commercial providers offer powerful but proprietary AI tool sets. On one hand these tool sets accelerate scientific progress; but on the other hand, they serve as obstacles to the goal of being able to move flexibly from one provider to another. The NAIRR could play a role in ensuring that these proprietary differences represent healthy competition of ideas, and not artificial barriers (e.g., via proprietary data formats).
- General cloud use is treated as services and subject to F&A under current Federal sponsor practices. The Financial Accounting Standards Board[6] has recently made some moves toward capitalization allowances for some cloud expenses. The NAIRR planners should carefully consider the myriad issues around how F&A is charged for cloud services.

4. In the non-commercial space there are many exemplar infrastructure activities that an NAIRR could build on. Some of these include

- NIH STRIDES and Bridge2AI- these programs have been successful in improving data availability across the broader NIH community
- NSF Open Storage Network - this pilot program has goals to provide a managed storage fabric for very large data collections that have value but may not be eligible for cloud provider public dataset hosting (for example very large corpuses or collections for which the science audience is valuable but small in number).
- The Dataverse community - a mature network of curated long-term data storage for research reproducibility and extension with a strong heritage in social and economic research data.
- NSF CloudBank - a streamlined program for allocating cloud resources within grant allocations

5. In the human resources space some example infrastructure activities NAIRR could leverage and strengthen include the:

- NSF Campus Champions, Cyberteam and ask.ci network - these activities support training and knowledge sharing across technical domains
- NSF XSEDE extended support program - this provides technology expertise to help domain projects leverage advanced computing well
- Northeastern University led AI Jumpstart initiative aimed at proactively training small and medium scale industries in AI leveraging state-of-the-art technology solutions.
- MIT and Harvard Airforce AI Accelerator programs aimed at connecting cutting edge AI research with strategic national security needs
- The DARPA Colosseum 5G simulator at Northeastern, a platform to enable AI researchers to explore applications of AI to next-generation wireless.
- BU SAIL program aimed at embedding AI researchers within domain research groups
- US Research Software Engineer Association network. This is a self-organizing group of research software engineering professionals, including technical AI/ML practitioners in academia, that is working to build a professional national community focussed on sustaining careers of people who sustain software infrastructure that underlies all manner of research, including AI and ML.

5. What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?

We believe that deliberate public-private partnering has excellent potential for creating an infrastructure that meets the broad national needs that a NAIRR initiative ought to aspire to, if it is to play a major role in national research and economic growth. We have discussed a number of public-private partnerships above.

The COVID19 HPC Consortium may provide aspects of an interesting exemplar on the potential of a collaborative multi-stakeholder public-private grouping that works effectively as a whole. Although the Consortium came about in very unusual circumstances, the activity demonstrated an entity that spanned NASA, DOE, NIH, NSF and DoD, universities, commercial cloud participants AWS, Azure, Google Compute, and IBM. The participants developed effective mechanisms for researchers from all sectors (industry, academic and federal) in need of resources to access a diverse mix of capabilities that included compute, data, and technical expertise all within a single overarching virtual organization. While the exceptional circumstances meant that important issues around financial and business bookkeeping were largely set aside, many other practical governance and operational issues that a close public-private NAIRR partnership might need to manage were addressed effectively. The NAIRR planners might want to leverage the lessons learned from this activity on how to reach a broad research community, provide relatively seamless access to resources, and proactively support and guide projects to the correct scale and type of resources.

A significant public-private partnering around NAIRR would be very much in keeping with more deliberate Federal, industry, and academia coordinated innovation strategies envisioned in the PCAST report "Recommendations for Strengthening American Leadership in Industries of the Future"[7].

Respectfully,

---

[1] Contributors to this document are *Wayne Gilmore,* Director of Research Computing Services, Information Services & Technology, Boston U.,; *John Goodhue,* Executive Director, Massachusetts Green High Performance Computing Center; *Christopher N. Hill,* Principal Research Scientist, MIT; *David Kaeli,* College of Engineering Distinguished Professor of Electrical and Computer Engineering, Northeastern U.; *Eric Kolaczyk,* Professor of Mathematics and Statistics, Director of the Hariri Institute for Computing and Computational Science & Engineering, Boston U.; *Jim Kurose,* Distinguished Professor of Computer Science and Associate Chancellor for Partnerships and Innovation, U. Massachusetts Amherst; *Scott Yockel,* Director for Research Computing, Faculty of Arts and Science, Harvard U.

[2] https://www.nsf.gov/pubs/2020/nsf20604/nsf20604.htm

[3] https://cra.org/cloud-access-for-nsf-cise-research/

[4] https://www.nitrd.gov/pubs/open-knowledge-network-workshop-report-2018.pdf

[5] https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/

[6] https://www.fasb.org/

[7] https://science.osti.gov/-/media/_/pdf/about/pcast/202006/PCAST_June_2020_Report.pdf