September 27, 2021

RFI Response: National AI Research Resource
White House Office of Science and Technology Policy and National Science Foundation
**86 FR 39081; Document Number 2021-15660**

The Center for Security and Emerging Technology (CSET) offers the following submission for consideration by the Office of Science and Technology Policy and the National Science Foundation. OSTP and NSF requested information to support the work of the National Artificial Intelligence Research Resource (NAIRR) Task Force, which has been directed by Congress to develop an implementation roadmap for a shared research infrastructure that would provide Artificial Intelligence (AI) researchers and students across scientific disciplines with access to computational resources, high-quality data, educational tools, and user support.

Our submission recommends:

- **The NAIRR should facilitate access to compute through NSF's new CloudBank program**.
- **The NAIRR should provide computational resources for undergraduate and graduate students in the U.S. through cloud access beyond what is freely available through Google's Colab or similar industry resources.**
- **The NAIRR should provide computational capacity comparable to the compute used to train the private sector's largest models for specific research by PhD researchers.**
- **The Task Force should engage with the Department of Commerce and state commerce departments to determine an appropriate mechanism that allows entrepreneurs to access NAIRR compute resources.**
- **The NAIRR should create a data consortium, potentially through a public-private partnership (PPP), that cultivates, enriches, and maintains datasets for public use.**
- **The NAIRR should facilitate the sharing of data brought by researchers to CloudBank and enable sharing by default.**

202100261.A) **Goals for establishment and sustainment of a National Artificial Intelligence Research Resource and metrics for success:**

The NAIRR should nurture the development of AI to advance U.S. economic competitiveness and national security in areas that are underserved by the current ecosystem. In particular, the NAIRR should support research irrespective of potential commercial implications; support research into the safety, security, privacy, and equity of AI systems; and provide resources to underserved communities of developers.

Since the NAIRR is not empowered to change market incentives, policymakers should determine what it can *provide* or *aggregate* for the AI research community. Data, computational capacity, and talent--the people who create algorithms and design ML systems--are three main inputs for AI development; capital underpins access to all three. NAIRR's greatest impact will be to

improve the quality of, and access to, these inputs for a wide collection of actors that the market does not adequately support.

In order to establish the NAIRR's structure, oversight, and metrics for success, CSET conducted a gap analysis of the AI development ecosystem using the matrix below. We considered the matrix for four different prototypical AI researchers who might benefit from the NAIRR: an undergraduate student, university PhD faculty or graduate student researcher, a PhD faculty with exceptional computational requirements, and an entrepreneur.[1] Actors in the horizontal axis shape the ecosystem of AI research using the inputs on the vertical axis. For example, the federal government provides computational capacity (infrastructure) to some PhD and graduate student researchers through the 18 National AI Research Institutes funded by NSF. **The analysis determined that undergraduate students and entrepreneurs rely on free computational resources provided by industry and that all four typical users suffer from a paucity of usable data.**

Figure 1. Matrix of U.S. R&D actors and inputs[2]

| | | Actors | | | | | |
|---|---|---|---|---|---|---|---|
| | | Federal Government | State & Local Government | Industry | Philanthropy | Academia | International Partners |
| Inputs | Funding | | | | | | |
| | Human Capital | | | | | | |
| | Infrastructure | | | | | | |
| | Policy & Regulation | | | | | | |

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

The Networking and Information Technology Research and Development Program (NITRD) oversees Cloudbank, a program that negotiates rates and finance of cloud compute for researchers. The Task Force should leverage the NITRD's existing Cloudbank stewardship by giving it the responsibility and necessary authorizations to manage NAIRR resources. Placing the NAIRR under NITRD, and subsequently CloudBank, will speed the rollout of NAIRR cloud resources, facilitate access by NSF-funded researchers already eligible for CloudBank, and eliminate unnecessary duplication of similar efforts.

[1] Completed framework analysis available upon request.

[2] Melissa Flagg and Paul Harris, "System Re-engineering: A New Policy Framework for the American R&D System in a Changed World" (Center for Security and Emerging Technology, September 2020). https://doi.org/10.51593/20200050

1.D) **Capabilities required to create and maintain a shared computing infrastructure to facilitate access to advanced computing resources for researchers across the country, including provision of curated data sets, compute resources, educational tools and services, a user-interface portal, secure access control, resident expertise, and scalability of such infrastructure.**

**Compute:**

**The NAIRR should facilitate cloud access through NSF's new CloudBank program**. Rather than creating another duplicative management structure, the Task Force could bolster the CloudBank program, providing accounts and credits to students and researchers across the U.S.

**The NAIRR should provide computational resources for undergraduate and graduate students in the U.S. through cloud access beyond what is freely available through Google's Colab or similar industry resources.**
The NAIRR can achieve this aim by creating a student interface on CloudBank. NSF CloudBank could afford all student accounts a basic budget for compute, and build a mechanism that lets them purchase more time, lets their university credit their accounts, and lets them receive scholarships for time on the platform. If adequately liberalized, such a mechanism would allow for outside funding to support students' learning and research on the platform. This system would also enable students to augment classes both at their institution and through MOOCs by performing computationally intensive exercises that may have been previously out of reach. This structure would also subsidize university costs to teach AI classes, since students could access such a platform for their work rather than the university paying for compute. Students should also be able to file patent applications based on research conducted on NAIRR cloud resources, not subject to Bayh-Dole regulations that currently guide patent protections for university faculty. Such a comprehensive system would nurture self-guided learning, innovation, and democratize computational access to students attending universities without the means to afford similar resources.

Student activity should be co-signed by faculty to avoid misuse of computational resources, like mining cryptocurrencies. Misuse of the cloud resources could result in the loss of federal aid to the student, just as those charged with possession of illegal substances lose access to FAFSA support.

At the other end of the spectrum, **some researchers may need computational capacity comparable to the compute used to train the private sector's largest models--the NAIRR should meet that need.** Owing to the expense of offering such computational capacity, applications for these extra-large models should be evaluated by the NSF as it does with other research grants. Providing enough compute may be easy, if a little costly. Meeting or in some cases exceeding the largest private sector models would cost as much as several million dollars which pales in comparison to other high profile research efforts like the National Ignition Facility ($3.5B) or CERN ($23B). High uptake of the NAIRR's cloud compute will lead to increasing costs, but this should be considered an indicator of success—a signal that NAIRR resources are supporting previously unmet demand. Private sector cloud solutions including AWS, Microsoft Azure for researchers, and Google TPU Research Cloud already offer

researchers access to high performance computing (HPC). As the financial intermediary, CloudBank can negotiate more competitive rates for these services as usage increases overtime.

Separately, CloudBank could use existing private-public infrastructure to quickly increase access to high performance compute. For example, Frontera, an NSF-supported supercomputer located at University of Texas at Austin, already provides access to HPC to some AI researchers and could be incorporated into CloudBank. Incorporating existing public infrastructure into CloudBank, like Frontera, would create a centralized marketplace for researchers and students to access compute. We support a proposal by Dr. Vince Kellen, CIO of UC San Diego and a co-PI for CloudBank, that would increase cloud compute resources. According to Kellen, many researchers and academic institutions maintain on-premises compute capacity; how much is unknown. Some of this capacity is likely underutilized, as there can be downtime between experiments and computational cycles. Dr. Kellen identifies this as an opportunity to create a marketplace for compute: a single, unified mechanism where academic institutions can list and lease their excess capacity. This would maximize the use of existing computational capacity, by allowing researchers at other institutions to conduct research remotely. If such a system proved successful, the high performance computing facilities of the national labs may similarly benefit from such a market—though appropriate guardrails would be necessary given the value of such assets.

**The Task Force should engage with the Department of Commerce and state commerce departments to determine an appropriate mechanism that allows entrepreneurs to access NAIRR compute resources.** Entrepreneurs are a particularly difficult group to reach with resources. Unmoored by commitments to academic or other public institutions, entrepreneurs also represent the riskiest group to serve. Ethical questions answered by IRBs in academia are unanswered in independent business ventures, so a careful administrative process to evaluate individuals and their actions of NAIRR cloud resources is necessary. The Department of Commerce and state commerce departments may find it easier to negotiate access if a system of ethics review boards are put in place.[3]

But, it is also necessary to provide this access. Start-ups rely on capital, either from lenders or private equity, to purchase compute and data access. Although free data and compute are available, adequate amounts of either can be expensive. This pushes entrepreneurs into a position of needing a proof of concept before seeking investments or loans. Such a scheme could be paid for overtime if a small equity share of companies developed on the platform is held by an independent trust, which uses the proceeds of equity sales or revenue earned from dividends to pay for more computational resources later on or reimburse past expenses. Many start-ups may fail; therefore, any such revenue would only subsidize some of the operational costs.

**Data:**
**The NAIRR should create a data consortium, potentially through a public-private partnership (PPP), that cultivates, enriches, and maintains datasets for public use.**

[3] James E. Baker, "Ethics and Artificial Intelligence" (Center for Security and Emerging Technology, April 2021). https://doi.org/10.51593/20190022

Procuring new data is time consuming and capital intensive. Well-financed companies can dedicate large amounts of resources to shaping data into usable formats for AI and machine learning research. Students, start-ups, and academic researchers cannot. The NAIRR must prioritize curating high-quality datasets for users—without it, computational resources may achieve relatively limited new discoveries as many researchers scour well-trodden data sources for new insights. The data consortium should own the following responsibilities:

1. **Identifying** datasets that are relevant to researchers' needs and represent a diverse set of topics. If AI research is to enable the development and growth of other sectors, data that reflects those fields is crucial to researchers' success.
2. **Acquiring** the data by scraping it from public-domain internet sources, buying it from vendors, manually recording data points contained in narrative sources such as news media, etc.
3. **Moving** the data from different endpoints (API, FTP, web downloads, etc.) into a computing environment where it can be processed and stored.
4. **Structuring and enriching** the data so that it has a consistent, meaningful format and can be efficiently sorted and queried. This may involve converting the data to a standard format, translating it from other languages, adding metadata, or developing and applying taxonomies and classification models. Outside specialists, such as annotators, technical experts, and translators, may be involved.
5. **Integrating** the data with other relevant datasets by mapping out common features, creating unique cross-connecting identifiers and fixing duplications and ambiguities. This "data fusion" process is a prerequisite for analyses involving more than one type of information (that is, most useful analyses).
6. **Validating** the data to ensure it reflects "ground truth" and has been processed correctly. This requires careful judgment and clear, consistent procedures.
7. **Documenting** the data so that others can understand what it is and how to use it.
8. **Hosting** the data through an online repository, likely data.gov or directly on the PPP's website. The data should be readily available on the CloudBank (or other selected) computing platform.
9. **Maintaining** the data as facts, user needs, vendor terms, etc. change over time.

**The NAIRR should integrate the data.gov website with the existing CloudBank platform and the proposed student access (above).** Integrating the two offerings may increase the use of government data.

**The NAIRR should facilitate the sharing of data brought by researchers to CloudBank and enable sharing by default.** Researchers should be able to limit the availability of their data with a provided justification, up to and including keeping it all private, but sharing data has many potential benefits. To facilitate this data sharing, CloudBank would need to host storage space for data and allow user accounts with appropriate permissions to access the data. Besides introducing replicability to research findings and publications, sharing data can facilitate other research, encourage the development of new benchmarks on shared data sets, and break down systemic moats that protect some particular researchers who benefit from close ties to particular

government agencies or data sources. Although all data sharing is ultimately the subject of agreement between the original party and the data provider, access to NAIRR cloud resources could facilitate more collaborative behavior.

**1.E) An assessment of, and recommended solutions to, barriers to the dissemination and use of high-quality government data sets as part of the National Artificial Intelligence Research Resource;**

While there are sometimes legitimate reasons for protecting government data, the government should identify non-sensitive data and share more of it. The NAIRR can help government agencies publish their data for use by researchers where feasible.

The Task Force should request an Executive Order or administrative rule requiring government agencies to 1) identify data that can be published without overriding classification or PII concerns, 2) fund NITRD to hire new FTEs able to conduct data quality analysis and standardization/cleaning of data sets in preparation for publication, and 3) require agencies work with these new FTEs to facilitate the publication of government data. NITRD will require new funding to hire these FTEs. As an alternative to new FTEs under NITRD, the Task Force could coordinate with the Presidential Innovation Fellows program, the U.S. Digital Service, and the U.S. Digital Corps. The end result will be new staff to help agencies clean and publish data for use by researchers of all stripes, not just AI researchers on CloudBank. The resulting curated data should be uploaded to data.gov. The mandate for agencies to work with new NITRD staff and publish those cleaned data will make available data which is currently held in an unusable format.

**1.F) An assessment of security requirements associated with the National Artificial Intelligence Research Resource and its management of access controls.**

The NAIRR should follow existing best practices for access management control utilized by private sector cloud service providers. Student's university accounts should be authorized by an email address from an accredited institution of higher education and require a cosigner from the institution.

Maintaining the integrity of datasets used on the NAIRR platform is an imperative. At a minimum, datasets uploaded to the NAIRR platform should be automatically hashed using a reputable hashing algorithm not known to be compromised (e.g., SHA256). Hashing datasets of data.gov by default, which is outside the purview of the Task Force, would also help authenticate the veracity of data before it is utilized. Though tedious, hashing datasets will prevent malign actors from tampering with data used to train models and conduct research. This step will help defend against data poisoning attacks and model backdoors.[4]

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

[4] Andrew Lohn, "Poison in the Well: Securing the Shared Resources of Machine Learning" (Center for Security and Emerging Technology, June 2021). https://doi.org/10.51593/2020CA013

The first priority should be NAIRR access for entrepreneurs and start-ups. These individuals and firms can increase market competition, drive innovation, and challenge incumbents. Disruptive start-ups are a politically neutral way of challenging the market dominance of a few firms and can lead to entirely new applications of AI. Moreover, these individuals are caught between not having enough compute and data to produce a viable concept, which stops them from attracting private equity or receiving bank loans.

The second priority for the NAIRR should be to create full-time staff positions that identify, curate, and facilitate the publication of government datasets. The positions can be organized under NITRD, Presidential Innovation Fellows, U.S. Digital Service, or U.S. Digital Corps; no matter which agency presides over the effort, data publication must be their top priority. There is no mandate and few incentives for government agencies to share their data nor are they adequately staffed to prioritize the curation of these data sets for public use. Rather than each agency spinning up their own staff to do this work, the NAIRR should create specialized teams of data scientists with the authority to work with across agencies to execute data sharing efforts. All data that qualifies for cleaning, formatting, and publication should be hosted on data.gov.

Notably absent from these priorities is access for academic researchers. Their absence reflects our conclusion that reaching different user groups requires different levels of intentionality. If compute and data are made available to entrepreneurs and students on CloudBank, academic researchers with access to research funds will be able to easily purchase compute on the same platform. Supporting access for entrepreneurs and facilitating data publication by the government will take far more work and resources than helping academic researchers get access, and thus they must receive higher prioritization when evaluating resource distribution. If CloudBank is sufficiently liberalized, academic researchers will utilize the platform as they see fit.

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

**Compute:**

The National Science Foundation supports CloudBank, an intermediary for NSF-funded researchers seeking to perform their research in the cloud. Though currently limited to 150 specific principal investigators, the platform could easily expand to other NSF-funded researchers and students. Because CloudBank negotiates financing for computational resources, more users will increase CloudBank's negotiating power.

- CloudBank's authorizations and responsibilities should expand to a wider set of researchers including students. This expansion should include infrastructure for account verification and management.
- CloudBank will need additional funding to allow student accounts to access these resources. Current CloudBank funding facilitates the purchase of cloud compute from companies with money from NSF-funded researchers' grants.

Another NSF-funded cloud program, Exploring Clouds for Acceleration of Science (E-CAS), also offers cloud services to 6 research programs. E-CAS is currently a smaller initiative than

CloudBank, but could serve as another pathway to providing cloud compute access to researchers.[5]

Frontera, currently the world's fifth fastest high performance computer, provides compute to AI-focused researchers.[6] Many national labs host high performance computers, including many of the 18 NSF-funded AI Research Institutes.[7] If the Task Force chose to follow a "market-maker" model of supporting AI research—in addition to supporting CloudBank, then the NAIRR could host a single unified registration system to reserve time on high performance computers. This approach would encounter significant organizational barriers, but would leverage existing resources to support the most computationally intensive research.

However, access to compute must be simplified for researchers, particularly students. They should make a single application for compute resources and, after approval, where they actually receive compute should be transparent to them. By creating a one-stop shop for compute under the auspices of the NAIRR, researchers can avoid having to apply for compute resources across many programs.

**Data Aggregation:**

The federal government's data.gov website publishes data from federal, state, and local government agencies. Data.gov should be a resource for researchers accessing the NAIRR. A website plug-in for data.gov on the NAIRR platform would suffice.

The private sector and research community compile public datasets with zeal.[8] NAIRR would do well to link these resources on the NAIRR platform.

[5] https://internet2.edu/cloud/exploring-clouds-for-acceleration-of-science/
[6] https://nsf.gov/cise/ai.jsp
[7] https://nsf.gov/cise/ai.jsp
[8] Google Cloud Public Datasets | AWS Public Datasets | Wikipedia Dumps | Kaggle | Harvard Dataverse | FigShare |