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

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government nor those of the National AI Research Resource Task Force., and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality, or content of all external links included in this document.

1. What options should the Task Force consider for any of roadmap elements A through I, and why?

[A] Goals for establishment and sustainment of a National Artificial Intelligence Research Resource and metrics for success.

There are two broad goals the Task Force has already rightly identified a NAIRR can help achieve: More AI innovation and bolstered U.S. competitiveness in AI.[3] However, it is important for the Task Force to recognize that even though many use these terms interchangeably, AI innovation and competitiveness mean different things and optimizing the NAIRR implementation roadmap to maximize for one can lead to different solutions than maximizing for the other.

First, consider how these terms differ. Competitiveness refers to the ability of an economy to compete effectively in global markets for traded goods and services in the absence of subsidies and government protections.[4] By enabling an economy to export more in value added terms than it imports, competitiveness increases a nation's standard of living. In contrast, innovation refers to developing an improved product, production process, or organizational method. If this innovation occurs in traded sectors, a nation's economy will become more competitive. But innovation in non-traded sectors will have less impact on competitiveness because by definition their output is not sold outside local borders.

These distinctions matter for implementing, operating, and administrating a NAIRR. To see why, consider two AI researchers, one of whom is pursuing research into AI models for construction and the other for manufacturing. Both research pursuits support AI innovation but only the latter would bolster U.S. competitiveness in AI because manufacturing is a traded sector from the perspective of the U.S. economy while construction is not. A NAIRR whose primary goal is to promote AI innovation should seek to support both types of research equally whereas a NAIRR whose primary goal is to bolster U.S. competitiveness should prioritize AI research for manufacturing over construction.

To be clear, boosting access to research tools can serve as means to both ends. But the Task Force should consider having well-articulated and distinct mechanisms to achieve each. One way to do this is by implementing separate support mechanisms for academic

[3] The White House, "The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force," news release, June 10, 2021, https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.

[4] Robert D. Atkinson, "The Competitive Edge: A Policymaker's Guide to Developing A National Strategy," (ITIF, December 2017), https://www2.itif.org/2017-competitive-edge.pdf.

researchers and private sector researchers to access the NAIRR. For instance, one mechanism could provide support for—and only for—eligible academic and government researchers who are conducting AI research that promotes AI innovation in any field, with research proposals reviewed through a competitive process. Another mechanism could provide eligible firms with innovation vouchers they can use to "buy" AI compute time and expertise at certain supercomputing centers, with the size and type of the voucher determined by its relevance to a national competitiveness strategy (e.g., focused on solving specific challenges and facilitating commercialization breakthroughs).

The role of government in increasing access to AI resources for academic and private sector researchers are different. Academic researchers typically conduct crucial early-stage AI research that provides foundational, generic knowledge that everyone—including industry—can draw on for ideas and innovation. However, only well-resourced institutions provide access to expensive AI resources, such as powerful AI compute. The government's role is to ensure as many qualified academic researchers as possible have access to AI resources in order to expand the pool of general AI knowledge for the benefit of everyone. Private sector researchers typically conduct later-stage R&D, which is important in bringing innovations to market. The private sector already has incentives to invest in AI resources. The role for government is to ensure the private sector's incentives to invest in R&D for AI are sufficient to maximize overall economic welfare.

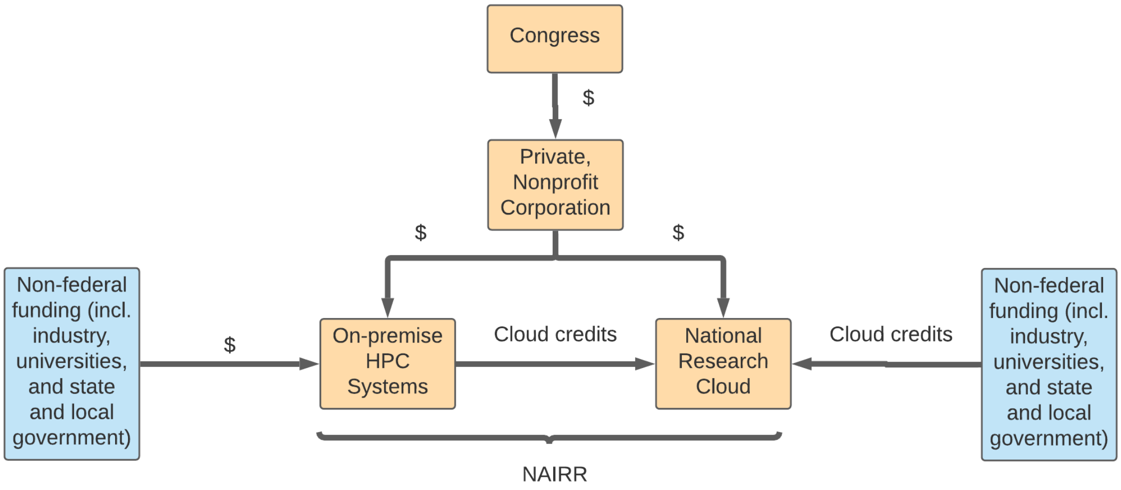
The Task Force has also rightly recognized that democratizing access to AI compute for academic researchers can help ensure all individuals have equal opportunity to succeed in becoming the next generation of AI researchers. The Task Force could introduce a third support mechanism that specifically supports the allocation of resources at Minority-Serving Institutions (MSIs) that include Historically Black Colleges and Universities, Hispanic-Serving Institutions (HSIs), and Tribal Colleges and Universities (TCUs) to help achieve this end.

[B] On a plan for ownership and administration of the National Artificial Intelligence Research Resource.

One option the Task Force should consider is for a private, non-profit corporation to allocate federal funds to the NAIRR. This entity would be created by Congress and act as a steward of the federal government's investment in national AI research resources. To see how this might work, consider the Corporation for Public Broadcasting (CPB), a private, non-profit corporation established by the Public Broadcasting Act of 1967 to act as a steward of the federal government's investment in public broadcasting. The mission of the CPB is to "ensure universal access to non-commercial, high-quality content and telecommunications services."[5] It does not produce or distribute programs, nor does it own, control, or operate any broadcast stations. Instead, CPB allocates federal funding to local radio stations in each of the 50 states, which broadcast national content to their local communities and broadcast local programs they create themselves too.

We propose the Task Force consider a comparable model for the administration of the NAIRR. Through the appropriations process, Congress would enact federal payments to a private, non-profit corporation, just as it does for federal agencies that fund high-performance computing (HPC) at supercomputing centers and universities (see Figure 1). The corporation would not own, operate, or control any HPC systems itself but instead be charged with facilitating geographic diversity of AI compute, the development and expansion of HPC for AI, and providing funding to local HPC systems.

Figure 1: Proposed model for the NAIRR.


Congress
$
Private, Nonprofit Corporation
$
$
Non-federal funding (incl. industry, universities, and state and local government)
$
On-premise HPC Systems
Cloud credits
National Research Cloud
Cloud credits
Non-federal funding (incl. industry, universities, and state and local government)
NAIRR

The activities of the corporation could be twofold: 1) to allocate federal funds to local, on-premise HPC systems at universities, colleges, and research institutes across the country; and 2) to provide funding for nationally accessible resources such as a National Research Cloud.

Regarding the former, local systems could be owned and maintained through public-private partnerships, which is discussed further in section 4. And to understand the latter, let us

[5] "About CPB," last accessed September 10, https://www.cpb.org/aboutcpb.

return to the example of public broadcasting. National programming producers like NPR, APM, and PRI are independent entities that are funded through a number of sources including corporate sponsorships, funds from CPB, and fees from locally owned and operated radio stations that pay to be their members and distribute their programming. A similar set up could work for nationally accessible HPC as part of the NAIRR. For instance, one nationally accessible resource could be a National Research Cloud (NRC), set up as publicly and privately funded non-profit with member institutions across the country. The members (both public and private) would make some level of AI compute available in the cloud and gain access to government datasets and other incentives from the NRC in return. Because it would be a public-private non-profit, the NRC could partner with private companies to obtain cloud services from existing vendors for AI researchers, which would be particularly valuable in the short-term as it gets established. In addition, local institutions would have a choice. They could choose not to participate in the NRC and exclusively provide local, on-premises AI compute, which will be important for some researchers who require on-premises resources for reasons such as data security, application performance, or teaching purposes.

Such a set-up would be adaptable, allowing for the incorporation of new resources and novel computing capabilities. One important and related question the Task Force raised in a recent workshop was which regions should it prioritize for on-premises systems?

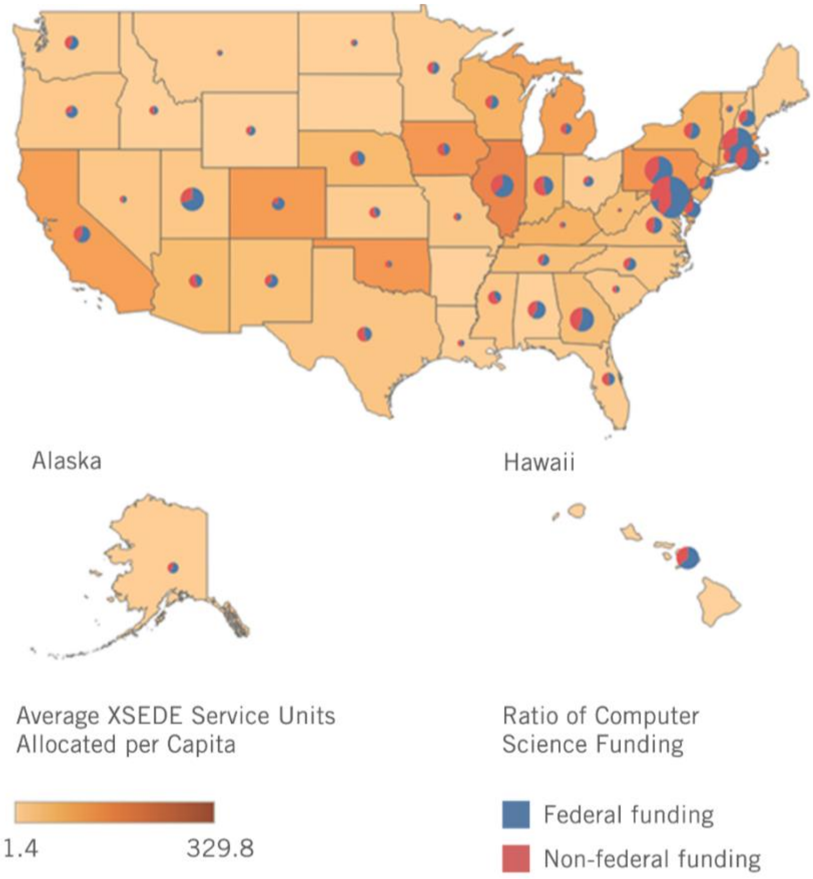
Which regions should the Task Force target for providing local systems?

The NAIRR should prioritize providing local resources in regions wherein the gap between AI compute demand and supply is greatest.

Some communities, institutions, and regions already have high access to HPC availability while others are conducting high levels of AI research but have little access to powerful systems. In our 2020 report *How the United States Can Increase Access to Supercomputing*, we provided an estimate of access to HPC resources per capita across the United States. We used data on compute time researchers requested in 2017, 2018, and 2019 from NSF's Extreme Science and Engineering Discovery Environment (XSEDE), a platform that coordinates the national sharing of supercomputing, as well as data on the researcher's organization and the state in which the organization is located (see Figure 2).[6]

[6] Hodan Omaar, "How the United States Can Increase Access to Supercomputing," (Center for Data Innovation, December 2020), https://www2.datainnovation.org/2020-how-the-united-states-can-increase-access-to-supercomputing.pdf.

Figure 2: Proportion of XSEDE service units allocated per capita and size of research funding from high-research institutions in computer science per capita in each state.


Alaska
Hawaii
Average XSEDE Service Units Allocated per Capita
1.4
329.8
Ratio of Computer Science Funding
Federal funding
Non-federal funding

The key insight from this figure is more access to powerful HPC resources is found in states like Massachusetts, Pennsylvania, and Illinois that have leading academic institutions, which can either stand up their own HPC centers or partner with other leading research institutions in their state to create multi-institutional centers.[7] Federal investments in more HPC resources in regions where HPC availability is already high will not be the most effective way to close the gap between HPC demand and supply because institutions either already have baseline AI compute and are using it for research, or they don't have research funding which means access to HPC is not the problem, research funding is.

[7] Top institutions are defined by whether they are ranked among the top 500 research institutions. The data is limited to R1 (very high research activity) and R2 (high research activity) universities.

By contrast, little access to powerful HPC resources is found in states like South Dakota and Utah that have few leading research academic institutions that have the capacity to support HPC systems.

What is important though, is that all regions that lack access to HPC are not the same. Some are doing more AI research than others. For example, while Utah's academic supercomputers are neither particularly large nor particularly powerful, the state is home to the Scientific Computing and Imaging (SCI) Institute, a research institute that focuses on conducting application-driven research in new scientific computing and visualization techniques and tools. The SCI Institute's faculty and alumni are recognized around the world for their contributions to scientific computing and research. South Dakota also has few HPC systems. But unlike Utah, South Dakota has no research facilities identified as conducting high-level research in any field.

The point is, there should be demonstrable evidence that providing access to AI compute in a community, institution, or region will result in an increase in AI research because as explained earlier, democratization is a means to an end, not an end in itself. In cases wherein HPC availability and AI research is low, the Task Force should consider requiring institutions to first increase funding for AI research, prove that they have sought partnerships with industry, or that increasing resources will support AI education and training for underrepresented groups, because there is a risk that investments in AI compute may not return increases in AI research.

We acknowledge that this map is limited because it only shows demand for a subset of academic researchers, not for all researchers. However, as several individuals in the Task Force's workshops have noted, there is little literature on what level and type of compute AI researchers need. Our report offers a starting point, but the Task Force should seek to work with federal agencies and private sector companies, where possible, to obtain additional data on HPC supply and demand.

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

Governance, oversight of strategic direction, and the allocation of federal funds should be made centrally by the private, non-profit corporation. It could have a board of directors appointed by the President of the United States which, after confirmation by the Senate, could serve a fixed length term. In turn, the board would appoint roles for leadership of the corporation, such as president and chief executive officer.

The National Research Cloud could have its own board of directors that oversees day-to-day operations and manages the NRC budget, which could be elected by its member institutions and organizations. Similarly, if the Task Force decides to coordinate the sharing of on-premises systems, this could be governed by a collaborative partnership of participating institutions just as XSEDE is.[8]

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

The NAIRR will have a number of distinctive attributes that will make its security somewhat distinct from general-purpose computing architecture.

First, the primary purpose of the NAIRR is to provide researchers with access to advanced, high-performance computing systems, and obtaining time on these systems will be highly valuable. However, stakeholders are likely to disfavor security protocols that impede collaboration or usability.

Second, because the NAIRR may bring together disparate computing systems and distributed data with varying levels of reliability and provenance, there is a risk that the responsibility of cybersecurity will be left to institutions, resulting in a patchwork of security protocols across the country. At the same time, computer security is context- and mission-dependent. A security mechanism designed to enforce a particular policy considered essential for security by one site might unnecessarily block legitimate users of another site.

Fortunately, there are several security solutions that can overcome these constraints and several groups have been thinking about them for a long time. In a 2021 paper titled "Trustworthy Scientific Computing," Sean Peisert, who leads computer security R&D at Lawrence Berkeley National Laboratory, proposed a model called hardware-based trusted execution environments (TEEs). As Peisert explains, TEEs increase HPC security at a minimal cost to performance by isolating computation and "preventing even system administrators of the machine in which the computation is running from observing the computation or data being used or generated in the computation."[9] This paper is part of a larger project that Peisert leads at the Berkeley Lab Computational Research Division, a national laboratory operated by the University of California, to take a broad look at several aspects of security and scientific integrity issues in HPC systems. Since this project has already begun testing and identifying the security requirements of national HPC resources, the Task Force should

[8] "XSEDE Governance," last updated May 17, 2021, https://www.xsede.org/about/governance.

[9] Sean Peisert, "Trustworthy Scientific Computing," *Communications of the ACM*, 64(5), (May 2021), DOI: 10.1145/3457191.

seek to work with this group and others like it to get a fuller understanding of what the security requirements of the NAIRR will likely involve.

2. Which capabilities and services provided through the NAIRR should be prioritized?

The Task Force should prioritize the development of a service-oriented architecture, which would integrate widely divergent components in the NAIRR by providing users with a common interface and a set of standard protocols for them to efficiently access the tools they need.

On one hand, the distributed framework we have proposed for the NAIRR offers an operating model that is flexible enough to adapt to new scenarios, resources, and computing capabilities. Resource diversity is important to ensure AI researchers can remain competitive. Indeed, research and advisory firm Gartner predicts that by 2025, "traditional computing technologies will hit a digital wall, forcing the shift to new computing paradigms such as neuromorphic computing."[10] There is also already a growing market for emerging AI chips that are specialized to best support different AI capabilities and services. For instance, field programmable gate arrays (FPGAs), which are AI chips mostly used to apply trained AI algorithms to new data inputs, and application-specific integrated circuits (ASICs), which can be used for either training or inference tasks, have seen considerable adoption recently.[11]

However, a single resource made up of heterogenous computing systems and data with different architectures, interconnects, memory, and authentication policies presents practical challenges to researchers trying to execute services on the NAIRR and technical developers of the NAIRR who will need to create portals, gateways, and workflow engines for it. Fortunately, many of these problems are not new—just more difficult to solve at scale. XSEDE presents a promising example of how to enhance interoperability and cross-platform functionality. As a "single virtual system that scientists can use to interactively share computing resources, data, and expertise," XSEDE uses service-oriented architecture to guide users through the different services and capabilities NSF's resources offer, enabling them to efficiently access their desired functionality.

A tougher problem with heterogenous computing systems that was raised in a recent Task Force workshop is that it will be more difficult for very computationally intensive problems to be executed because users will have to deal with load balancing over different systems,

[10] Kasey Panetta, "Gartner Top 10 Strategic Predictions for 2021 and Beyond," *Gartner blog*, October 21, 2020, https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-predictions-for-2021-and-beyond.

[11] Saif M. Khan, "AI Chips: What They Are and Why They Matter," (CSET, April 2020), https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

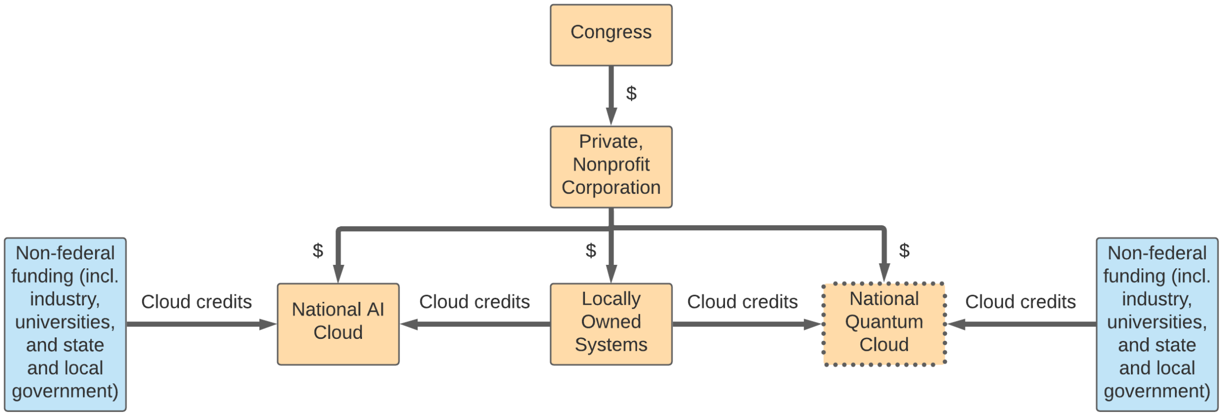
interoperability, resource selection, among other challenges. However, the Task Force should consider that the ultimate goal of the NAIRR is to democratize access to spur AI innovation, bolster U.S. competitiveness in AI, and bridge the "compute divide." The "long tail" of AI researchers that have more modest computational needs represent, in aggregate, the majority of AI researchers and a significant portion of AI advances. The NAIRR should therefore prioritize capabilities and services that meet the majority of AI researcher needs.

In the long run, a distributed framework could also enable the NAIRR to expand to include different technologies. Most importantly, the Task Force should consider in its roadmap how such a resource could incorporate resources for quantum computing. Because quantum computers are very specialized and expensive to develop, few universities provide access to these systems to support research activities. Instead, most academic researchers access these systems through quantum clouds—services that provide remote access to quantum systems through existing Internet infrastructure. Companies such as Amazon and Microsoft have already begun to make access to quantum computers available through their quantum computing-as-a-service (QCaaS) offerings, which are fully managed services that enable researchers and developers to begin experimenting with systems from multiple quantum hardware providers in a single place. Even with declining computing costs though, the costs and know-how for using advanced computing, including QCaaS solutions, will remain out of reach for many academic researchers.[12]

While AI and quantum computing differ, the crux of the problem is the same: How can the United States provide academic researchers with affordable access to high-end computing resources in a secure environment? Rather than reinventing the wheel, Figure 3 below illustrates how the scope of the NAIRR could be adapted to include additional resources to support quantum computing research.

[12] Hodan Omaar, "The Case for a National Quantum Computing Research Task Force in the United States," June 9, 2021, https://datainnovation.org/2021/06/the-case-for-a-national-quantum-computing-research-task-force-in-the-united-states/.

Figure 3: Expanding the scope of the NAIRR to include quantum computing.


Congress
$
Private, Nonprofit Corporation
$
$
$
Non-federal funding (incl. industry, universities, and state and local government)
Cloud credits
National AI Cloud
Cloud credits
Locally Owned Systems
Cloud credits
National Quantum Cloud
Cloud credits
Non-federal funding (incl. industry, universities, and state and local government)

3. How can the NAIRR and its components reinforce principles of ethical and responsible research and development of AI, such as those concerning issues of racial and gender equity, fairness, bias, civil rights, transparency, and accountability?

The ethical considerations regarding the NAIRR fall into two main buckets. One is how to ensure the allocation of resources in the NAIRR are fair and the other is how to ensure those resources are used to advance ethical and responsible AI research.

The first question is essentially a cake cutting problem, which is the challenge of allocating a single divisible, continuous, resource in a fair and equitable manner.[13] The "cake" in this case is the NAIRR and individuals have different preferences over difference pieces (because they will be pursuing different types of research and different systems within the NAIRR will be better suited to their needs). How should one split the cake so that it is fair in the sense of distributional fairness, understood as maximizing everyone's utility, and in the sense of not having disparate impact across protected groups?

Such practical problems are studied in mechanism design, a field of economics that studies the mechanisms through which a particular outcome or result can be achieved. Mechanism design can help bring analytic clarity to policy goals. Consider the statement: "The NAIRR should provide AI compute to the greatest number of AI researchers." This directive could be operationalized multiple ways. One way would be to minimize the total number of AI researchers that have less than some threshold of AI compute. Another way would be to first

[13] Rediet Abebe et al., "Fair Division via Social Comparison," AAMAS '17: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, (May 2017), pages 281–289, DOI: 10.5555/3091125.