RFI Response: National AI Research Resource Interim Report

Submitted by the Software & Information Industry Association to the Office of Science and Technology Policy and the National Science Foundation

On behalf of the Software & Information Industry Association (SIIA), we appreciate the opportunity to provide feedback on the interim report of the National Artificial Intelligence Research Resource (NAIRR) Task Force (the Task Force).

I. Preliminary remarks

SIIA, a non-profit organization, is the principal trade association for the software and digital information industries worldwide. Our members include over 450 companies reflecting the diversity of the information landscape, from creation to dissemination to productive and responsible use. They include digital content providers and users in academic publishing, education technology, and financial information, along with creators of software and platforms used by millions worldwide, and companies specializing in data analytics and information services. Our members support policies that foster innovation and a healthy digital ecosystem, including consumer privacy protections, responsible and ethical AI, and diversity, equity, and inclusion (DEI) initiatives.

We congratulate the Task Force for its excellent report. The report demonstrates robust engagement with the challenges of expanding access and pathways to ensure a more diverse, equitable, inclusive, and accessible AI R&D ecosystem in the United States. We view the establishment of a NAIRR as a critical means to achieve these goals.

We endorse the objectives and vision as the Task Force has presented them in Chapters 1 and 2. Strengthening the U.S. innovation ecosystem to realize the potential of AI applications to advance "science, economic growth, national security, and the ability to meet pressing societal challenges" while protecting privacy, civil rights, and civil liberties. Our collective ability to unpack this potential most fully will depend on our ability to educate and foster talent across socio-economic lines, particularly, as the report notes, among "traditionally underrepresented groups in AI R&D" (Rep. at 2-1), and to provide access to data and compute resources. This approach will help to advance DEI goals with respect to AI expertise and will provide a stronger foundation for ensuring that values-based assessments are built into the datasets and algorithms that drive AI applications.

We recognize that the Task Force intends the report as an outline for the NAIRR project. As the Task Force works to develop implementation steps to realize the NAIRR, we provide the following comments in the spirit of assisting in that process and address only a handful of the recommendations.

II. Establishment and sustainment of the NAIRR (Chapter 3)

The report demonstrates careful consideration of alternative options for structuring the NAIRR. Each of the alternatives presented in Recommendation 3-3 has benefits and downsides. Considering these alternatives in light of the Task Force's vision and objectives for the NAIRR, we recommend that the Task Force propose to establish the NAIRR within an existing FFRDC. We believe this approach would provide the NAIRR with the strongest path to achieve its objectives in a manner that promotes diversity and democratization of AI R&D in the United States. It will speed the process from legislation to execution by leveraging the FFRDC's existing back-office infrastructure and expertise in managing complex government and public-private projects.

Locating the NAIRR in the Federal government has appeal although will present extraordinary challenges with respect to basic organizational requirements. These include appropriations restrictions, limited ability to obtain resources and funding from non-governmental sources, and restrictive hiring authorities. These challenges are likely to delay creation of the NAIRR and generate ongoing complexities in execution.

A university-based approach also has appeal. However, as the report notes, there is a growing divide in AI resources concentrated in large private-sector firms, well-resourced universities, and national laboratories." (Report at 1-1.) The approach taken by the National Science Foundation (NSF) AI Institutes program can serve as a model to democratize access to AI R&D resources, although the decentralized nature of the program will present other challenges in executing the NAIRR vision.

Assuming the Task Force will recommend establishment of the NAIRR outside the Federal government, we would encourage the Task Force to provide additional guidance on the role of Federal agencies in the establishment and ongoing operations of the NAIRR. Recommendations 3-1 and 3-2 call for involvement by multiple Federal agencies, including the Department of Energy, the NSF, the National Institutes of Health, and the National Institute of Standards and Technologies (NIST). This "federated approach" is critical to facilitate oversight and guidance, expertise, funding, access to government data, and other resource needs.

We believe attention to the role of the Federal government is essential because the Federal government has an unmatched ability to catalyze diverse actors, including research institutions, private industry, civil society organizations, and state and local governments. This convening power is what distinguishes the NAIRR from other efforts to advance and democratize AI R&D.

While the Task Force recommendations make clear that Federal agencies will have ongoing roles in providing access to government data and agency expertise, we encourage the Task Force's implementation plan to include recommendations on the following:

- Identifying a lead agency or office. The success of the NAIRR in the short term will require dedicated guidance and participation of the Federal government. Though a federated approach makes sense from the perspective of resourcing, coordination will be essential to assisting NAIRR in coordinating among different stakeholders and marshalling the resources of the Federal government. An entity such as the Office of Science and Technology Policy (OSTP) or the National Science Foundation (NSF) has requisite experience to take on this role.
- Anticipated role of Federal agencies in supporting the NAIRR at establishment and in an ongoing manner. Members of the Task Force have the expertise and experience to provide a concrete vision for how the Federal government will engage with NAIRR at inception and over time. Providing guidance to implementers with further detail on agencies' anticipated financial and resource contributions will help to guide Congress, the Executive Branch, and the NAIRR management entity in advancing the NAIRR proposal.
- Anticipated need for specific Federal roles to support NAIRR. The NAIRR should seek to leverage government expertise in critical areas. For example, we would recommend that NIST lead the effort to provide standards for assessing the quality of data pools contributed to and created by the NAIRR. The NAIRR should leverage NIST's expertise in developing standards for test, evaluation, verification, and validation procedures and building a risk-management framework for responsible AI.

In addition, Recommendation 3-1 contemplates that Congress will appropriate funding to individual Federal agencies that will support the efforts of NAIRR and that NAIRR management will explore additional, presumably non-governmental, revenue sources. We support the approach to funding NAIRR through multiple sources. As funding (and other resource support) will be fundamental to success, we would urge the Task Force to consider funding needs over a five-to-ten-year period with recommendations about the level of funding that may be required from Congress and from private sources.

Consistent with our remarks regarding the unique convening power of the Federal government, the Task Force should consider what ongoing role Congress may have in sustaining the NAIRR infrastructure and providing an ongoing appropriations source – either directly or, as currently proposed, through individual Federal agencies. Recommendation 3-16 would require dissemination of reports to the public, Congress, and supporting Federal agencies. Beyond this, the report does not contemplate a role for Congress. While it may be left to the management entity to determine funding needs and sources, on the assumption that regular appropriations from Congress will be needed, we

encourage the Task Force to consider what ongoing relationship the NAIRR will have with Congress in terms of oversight and potentially direct appropriations.[1]

Recommendations 3-13 and 3-14 address contributions from the private sector, which will be essential to ensure the success of NAIRR in resource intensive areas (such as data and compute, covered in Chapter 4) regardless of congressional appropriations. We support involvement of the private sector and note that several firms, during the initial RFI period, indicated willingness to support the NAIRR through different types of contributions. We encourage the Task Force to provide further guidance about anticipated private sector financial and in-kind contributions, including anticipated needs and recommendations to address potential conflicts while allowing NAIRR to leverage private sector resources to support talent, data, and compute necessary for the NAIRR program.

We strongly endorse Recommendation 3-7's call "to build a DEIA focus into the system and operational plan from the beginning, rather than as an afterthought." We support elevating and elaborating on this recommendation as it is critical to the success of the NAIRR in achieving its overall objectives. The Task Force may consider moving this recommendation into Chapter 2. As the Task Force develops its implementation plan, we urge it to consider recommendations specific to cultivating talent from underrepresented groups including from minority serving institutions.

III. NAIRR resource elements and capabilities (Chapter 4)

Our feedback on Chapter 4 focuses on the findings and recommendations with respect to data. Access to robust, reliable, and trustworthy data is a key impediment to the democratization and diversification of AI innovation and to the quality of AI innovation. Developing robust datasets that meet the standards for responsible AI and minimize privacy concerns is extremely costly for most researchers, state and local government agencies, and companies. The alternative of relying on poor quality data increases the likelihood of unintentional bias and faulty predictions. Datasets that do not comport with standards of accuracy, reliability, trustworthiness, and bias present significant societal risk.[2]

We strongly endorse the Task Force's findings on data. As Findings 4-1 through 4-3 accurately claim, the curation and aggregation of robust, high-quality datasets is one of the leading challenges that

[1] Congress has in the past created non-governmental entities and has funded them in different ways. The Corporation for Public Broadcasting (created by the Public Broadcasting Act of 1967), for example, continues to receive Federal appropriations, while the National Constitution Center (created by the Constitution Heritage Act of 1988) received "seed" funding and now relies exclusively on philanthropic support, ticket sales, and membership.

[2] See, e.g., Joshua New, "AI Needs Better Data, Not Just More Data," Center for Data Innovation (Mar. 20 2019), https://datainnovation.org/2019/03/ai-needs-better-data-not-just-more-data/; Tasha Austin, et al., "Trustworthy Open Data for Trustworthy AI," Deloitte Insights (Dec. 10, 2021), https://www2.deloitte.com/us/en/insights/industry/public-sector/open-data-ai-explainable-trustworthy.html.

AI researchers and experts face when conducting their work. Differences in data labeling and data curation hinder the widespread adoption and deployment of AI in a variety of fields. In tandem with data challenges, many AI experts (particularly those in underserved and underrepresented areas) struggle with the acquisition of necessary computational resources. Finding 4-8 aptly describes a common occurrence in which AI experts may be hindered in their work without access to sufficient compute resources. Within a more democratized ecosystem that the NAIRR will provide, AI experts will be able to surmount these challenges, as resources–not skill–are the greatest limiting factor in the national AI R&D environment.

By their very nature, AI and ML technologies benefit from access to vast datasets. Whether through an independent aggregation of a large dataset or a multilateral conglomeration, AI stands to gain massively from diverse data. Independently aggregating data, however, can be extremely costly and difficult to accomplish. The alternative, while simpler with regards to the actual gathering of data, poses significant challenges that could be addressed by the NAIRR. Commonly, multilateral data aggregation suffers from issues such as the storage of data, differences in data labeling, and a lack of high quality, specialized data all of which originate from the decentralized nature of this approach.[3] The NAIRR would aid in surmounting this problem by providing a central entity in which contributors could store and share data to facilitate a cooperative effort on research fronts.[4] Furthermore, with a more unified structure, the NAIRR would offer the opportunity to present labeling standards to which data contributions must adhere, resolving the issue of heterogeneity. With contributing entities able to focus on specific data of their choosing rather than being concerned with quantity, this specialization could increase the overall quality of the NAIRR's stored data. Experts in the field of AI allege that in recent years, enormous investments have been diverted away from AI R&D to other financial ventures.[5] The proposed establishment of the NAIRR would thus aid in reinvigorating AI research and overcoming present roadblocks.

The research opportunities that the establishment of the NAIRR poses are significant. The opportunity for contributors to focus solely on their area of data expertise will yield greater quality, more reliable data. This refining of data presents an excellent opportunity for researchers and companies alike to conduct their own research on a large, robust dataset. Within the medical field, for

[3] Sara Brown, "Why it's time for 'data-centric artificial intelligence'," MIT Sloan (June 7, 2022), https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence.

[4] Connor Wright, "Our Top-5 takeaways from our meetup 'Protecting the Ecosystem: AI, Data and Algorithms'," Montréal AI Ethics Institute (Sept. 20, 2021), https://montrealethics.ai/our-top-5-takeaways-from-our-meetup-protecting-the-ecosystem-ai-data-and-algorithms/.

[5] RE•WORK, "Experts Predict The Next Roadblocks in AI" (Aug. 20, 2020) https://blog.re-work.co/experts-explain-the-next-roadblocks-in-ai/.

instance, there presently is insufficient usable medical data for research.[6] Some of the largest datasets available to medical experts are strewn across a mélange of "national government-sponsored studies, insurance claims, large clinical trials, cohort studies, and individual institutional registries."[4] To centralize the needed data in the NAIRR, actualization of Recommendation 3-13 would be an excellent mechanism to do so by linking an entity's access to NAIRR resources to its data contribution levels. This approach would augment data available through the NAIRR while also providing much-needed computational resources. By amassing data not only across medical fields but also all disciplines while simultaneously encouraging that quality be upheld, the NAIRR presents a unique opportunity for large amounts of reliable data to be available, a key component in the democratization of artificial intelligence.

One of the NAIRR's most effective tools in ensuring high quality data is contained within Recommendation 4-6. The provided technical infrastructure and support staff are integral to the NAIRR's success, as they serve to educate and cultivate "community-driven standards and improvements to data quality as are determined by the relevant domains." Taking the medical field as an example once more, there is no standard model or procedure by which medical experts gather data.[4] To successfully deploy and scale AI research on a national level, leadership is required to standardize disease diagnoses (data categorization and labeling) in order to establish "ground truths" or "gold standards" for classification.[7] These "ground truths" are of the utmost importance when running supervised AI models which are trained to predict and classify based on said truths.[5] The NAIRR's permanent technical support staff would be able to standardize data practices within appropriate domains, thus negating the need for post hoc data curation.

This robust, high-quality data that is aggregated within the NAIRR can then be analyzed and used by others. It is this form of multilateral data collaboration that can empower professionals to perform research previously found to be extremely costly. By democratizing and opening access to this data, the NAIRR could enormously expand the number and range of studies and research conducted.[5] AI models themselves benefit from having a plethora of data sources, and a plethora of researchers would benefit from these new AI possibilities. The outline to incentivize data contributions as a collective (as alluded to in Recommendations 3-14 and 4-1) provides the NAIRR the chance to overcome a sort of collective action problem and to have the widest possible impact on the AI community.

Democratizing AI R&D has economic impact as well. The inherent nature of data is nonrival, and thus benefits can be derived from data aggregation at a large scale. Some experts at the Joint Research Centre (JRC) in the European Commission allege that from a societal perspective, "it may therefore be

[6] Kobayashi, Y., et al., "How will 'democratization of artificial intelligence' change the future of radiologists?," *Japanese Journal of Radiology* 37, 9–14 (2019), https://doi.org/10.1007/s11604-018-0,793-5.

[7] Wang, Sophia Y et al. "Big data requirements for artificial intelligence," *Current opinion in ophthalmology* vol. 31,5 (2020): 318-323, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8164167/.

better to share data as widely as possible" as could be done through the NAIRR.[8] Furthermore, the economic nature of data appears to align with economies of scope and scale. Those of scope focus on aggregation of data across a "variety of situations and observations" into a single dataset, while those of scale focus on large, in-depth datasets.[6] The combination of these two (breadth and depth) are captured within the structure of the NAIRR in its appeal to varying disciplines and its dedication to quality.

We encourage the Task Force's implementation plan to build out Recommendations 4-1 through 4-5 with concrete methods for the NAIRR to obtain and develop large, high-quality, and privacy-protected datasets. Despite the richness of government data and the thorough recommendations to make government data available through the NAIRR (see Recommendations 4-10 and 4-11), there remains a significant amount of data that is not within government control. The NAIRR can serve a critical function by gathering and making accessible data and incentivizing the creation of new datasets.

Specifically, we recommend that the Task Force include methods for leveraging private sector contributions, creating new public-private initiatives to gather and curate data, and leveraging Federal government expertise to ensure that datasets made available to researchers through the NAIRR meet high-quality standards. Examples of such methods include:

- Leveraging private sector resources to generate large synthetic data pools. Synthetic datasets can enable algorithms to run on data that reflect, rather than rely on, real-world data. This approach would allow for the creation of a robust data lake that can be vetted to ensure accuracy, reliability, fairness, and so on. Moreover, it would not present privacy and individual rights concerns that may arise from the collection, retention, sharing, and use of datasets that are built directly from personal information. We understand there is interest in the private sector to work with the government on this sort of initiative.
- Incentivizing private sector companies to provide unique data in a non-proprietary form. Many potential AI applications rely on proprietary data that private sector entities are understandably reluctant to make available. The NAIRR should explore methods to incentivize collection of such data on a voluntary basis and leverage appropriate privacy enhancing technologies (PETs) to ensure protection of proprietary information.
- Fostering the creation of large open datasets of personal information collected through enhanced notice and consent procedures. Personal information remains critical to many potential AI applications yet the collection and use of such information raise privacy concerns. Pilot projects to gather new forms of data (such as voice samples) from

[8] Martens, Bertin, "The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning," JRC Digital Economy Working Paper 2018-09 (Dec. 2088), https://ssrn.com/abstract=3357652.

individuals who have received notice and consented would help to avoid these concerns.

IV. Privacy, civil rights, and civil liberties (Chapter 6)

We endorse the finding and recommendations in Chapter 6 of the report. We offer the following suggestions for consideration by the Task Force in developing the implementation plan.

First, we encourage the Task Force to consider proposing a framework for assessing privacy, civil rights, and civil liberties issues with respect to data and algorithms. Recommendation 6-2 nicely outlines an ethics vetting process. While there is wide agreement on the need for ethics assessments, there is variance on what benchmarks or criteria should be used to guide evaluation, particularly with respect to civil rights and civil liberties concerns. We recognize that the Task Force recommends that the NAIRR management entity have responsibility for developing these criteria. Given the extraordinary expertise and experience on the Task Force, this task would benefit from additional guidance from Task Force members.

Second, with respect to the vetting process outlined in Recommendation 6-2, we encourage the Task Force to examine what role the Federal government and other actors should have in ensuring that the data used within the NAIRR framework meet high-quality standards for reliability, trustworthiness, and bias. We encourage the Task Force to explore ways to incorporate NIST's standards and expertise into a vetting process.

V. Ideas for developing a roadmap to establish and build out the NAIRR in a phased approach, and appropriate milestones for implementing the NAIRR.

As the Task Force develops its implementation plan, we recommend preparing draft legislation and, as appropriate, text of proposed executive orders. We have found that providing lawmakers and policymakers with draft text is generally welcome and helps to facilitate the path from concept to realization.

* * *

Thank you for the opportunity to provide feedback on the interim report of the Task Force. We would be pleased to discuss any of these issues in further detail. Please direct any inquiries to **Paul Lekas, SIIA Senior Vice President for Global Public Policy** (plekas@siia.net).