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

Executive Summary

Establishing a National Artificial Intelligence Research Resource (NAIRR) is an ambitous and admirable goal. With the NAIRR's envisioned computing and data infrastructure to democratize AI, collaboration between various stakeholders will lead to better research, innovation, and, ultimately, citizen services. SAS Institute Inc. (SAS) welcomes the chance to provide suggestions to the NAIRR Task Force to help shape its roadmap strategy. In our more than 45+ years as a leader in the field of advanced analytics, SAS strongly agrees with the goal of democratizing access to actionable AI.

It may be surprising that SAS' view is that NAIRR should embrace both a variety of open source and commercial tools and methods as a commercial software vendor. In our view, democratizing analytics requires data scientists to collaborate across coding languages (e.g., SAS, Python, R, Lua, etc.) as well as low-code analysis techniques. Successfully accomplishing this means NAIRR should pursue a strategy built upon choice rather than one around a prefered technology solution. The Task Force should consider deploying solutions that have capabilities to enable researchers to detect and mitigate bias (both in data and in models), explain and interpret the output of their models at a global and local level, and employ privacy preserving techniques.

Unleashing the power of an open analytics ecosystem must balance this choice of techniques with the need for a common framework for collaboration, governance, data management, and explainability. This layer of control will allow researchers with different skill sets to collaborate on a common problem seamlessly, ensure the data being run through AI algorithms is high quality, and ensure that models being run across different languages can be explained effectively. Governance of data and analytic models is also a key component to reinforce principles of data ethics and responsible research.

Blending this strategy together at scale requires interconnected initiatives related to people, process, and technology; something best done through public-private partnerships (PPP). SAS alone is investing $1 billion in AI over three years through software innovation, education, expert services related to topics ranging from advanced analytics, machine learning, deep learning, NLP and computer vision. Leveraging these types of investments across the private sector will increase the speed to reaching a successful implementation of the NAIRR strategy. SAS is eager to make meaningful contributions to this effort and welcomes any follow-up to our recommendations.

SAS Feedback on NAIRR Roadmap

1. *What options should the Task Force consider for any of roadmap elements A through I above, and why? [Please take care to annotate your responses to this question by indicating the letter(s) of the item (A through I in the list above) for which you are identifying options.]*

Roadmap Element D – Capabilities to create/maintain shared computing infrastructure

This element will require consideration of the people, processes and technology used to enable the NAIRR.

People considerations

To enable research, especially in traditionally underserved communities, institutions, and regions, the Task Force should consider using solutions that:

- Are well supported by training resources, online communities, and technical support.
- Have a significant user base so that developed skills are easily transferable.
- Provide graphical user interfaces in addition to coding interfaces – enabling domain experts, citizen data scientists, and users who may be new to the field to take advantage of AI capabilities faster, without having to become full-fledged programmers. In fact, experienced researchers may find it beneficial to be able to test ideas in a graphical interface without having to write code.
- Enable use and provide support for people with disabilities.

NAIRR leaders and administrators should have a firm understanding of data sources utilized, and possess and/or have access to those with domain expertise in AI, machine learning, data governance and infrastructure knowledge associated with deploying AI solutions (e.g., hardware, networking, cloud architecture, etc.).

Process considerations

Processes are critical to consider roadmap element D. The Task Force should:

- Develop processes for open-source package management.
- Consider how easy it will be to apply system updates for COTS and open-source solutions.
- Establish processes for ensuring that the NAIRR is restricted to use by the intended audience and appropriately shareable amongst those with access.
- Leverage automation wherever possible to promote repeatability and ensure all resources are being used efficiently.

Technology considerations

Finally, as technology is the core of NAIRR operations, the Task Force should consider:

- The variety of software and hardware used to develop AI applications. There are

many open source and COTS products that can be used to develop AI applications. Many times, researchers may want to use multiple programming languages and hardware as part of an overall development approach. Similarly, teams of researchers may have different skill sets on the same team, yet still need to collaborate on a common problem. Given this, the NAIRR should be a platform in which researchers can seamlessly apply a variety of languages and techniques using both open source and COTS products. This will have the additional benefit of not locking AI researchers into any specific technology.

- Scalability will be critical. AI projects typically ingest massive amounts of data and are computationally intensive. Computational constructs that leverage containers and Kubernetes can enable elastic scaling and can provide workload management features, so researchers are not competing for compute resources. These constructs can also provide fault tolerance in the event of a hardware failure.
- Access to specialized hardware and optimized software. Some types of AI algorithms, such as those used for deep learning, run more efficiently on specialized hardware chips like GPUs (as opposed to traditional CPUs), accelerating the time to train models. Many AI applications also require millions of complex transactions per second which benefit from optimized software that takes advantage of techniques such as in-memory computing. The Task Force should consider providing a platform that allows for access to the specialized hardware and software required for modern AI research.
- Centralized administration and governance. The Task Force should consider solutions that provide activity tracking of various aspects of the NAIRR environment, such as servers, job content, and usage. Aside from making the environment easier to maintain, this will facilitate a chargeback model if that is something the Task Force chooses to implement.
- Ecosystem integration. Given the wide variety of technologies that might be included in the NAIRR, the Task Force should consider how these will interoperate. Open software that enables communication by REST APIs will enable technologies to be more easily integrated.
- Security. The technology should have authentication and authorization mechanisms to ensure appropriate access to technical capabilities and data sources. All-or-none access permissions placed on datasets may protect sensitive information – including personally identifiable information (PII) and personal health information (PHI) – but unnecessarily restrict the utility of data available for AI applications. The Task Force should pursue a solution that includes automated detection and masking of sensitive information and/or row-level security permissions to ensure a larger, more representative amount of data is available for researchers. The Task Force may also want to consider whether to encrypt sensitive data and ensure its secure transmission.

- **Data ethics capabilities.** The Task Force should consider deploying solutions that have capabilities to enable researchers to detect and mitigate bias (both in data and in models), explain and interpret the output of their models at a global and local level, and employ privacy preserving techniques.
- **Data curation.** An AI application is only as good as the data that goes in it – and it provides the best return when it is supported by a well-governed data management program. AI systems do not merely extract insights from the data they are fed (as traditional analytics do) – they actually change the underlying algorithm based on what they see in the data. The more data they are fed, the more tightly they define the algorithm and the more confidently they make classifications or predictions. Because of this feedback loop, errors can multiply upon themselves if bad data and/or biased data (see Q3) is fed into the AI application. The dangers in that are obvious: inconsistency, inaccurate insights, loss of trust and decisions made that are misaligned with values and policy.
- **Data management.** In addition to providing curated data, the Task Force should consider data cataloging solutions to make it as easy as possible for the research community to find curated data, identify data quality issues and other attributes of the data that may need to be addressed, and evaluate its fit for purpose – including the potential to introduce or amplify bias in the resulting models.

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

In terms of prioritizing the items above, we would recommend ensuring people with the requisite skills are in place first. It would be valuable as well to coordinate and share learnings that emerge from bringing researchers together from academia, the private for-profit and nonprofit sectors, and the public to use those learnings to adapt to the technological and human infrastructures that advance multi-, inter-, and trans-disciplinary research. With these inputs, the Task Force would then be able to develop thoughtful processes and implement the required technology.

Once operational, NAIRR administrators should monitor usage to determine whether there are topic areas that are more addressed than others, commonly encountered problems, and other aspects of the environment for insights into subsequent prioritization of capabilities, services, and the curation of data sources.

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?*

AI and related technologies do not exist in a vacuum. They are inextricably linked to society and the natural world around us – often influencing one another in unexpected ways. Researchers leveraging the NAIRR platform should be provided with the necessary resources to practice developing AI technologies responsibly. The NAIRR task force

should consider a four-pronged approach to reinforce principles of data ethics and responsible research:

1. *High Quality Data / Data Governance*

Many instances of AI products and insights that have gone awry stem from the utilization of data that do not appropriately reflect the breadth of experiences within the larger population, or that perpetuate biases, or are not fit for purpose. The Task Force should consider the following:

- The data made available to researchers on the NAIRR platform should be carefully considered, managed, and governed appropriately with the goal of improving quality over time.
- Data should be representative of the diversity of the country as a whole, be granular enough for insights to be derived at a local level, and include a rich variety of high-quality data sources that could cover a broad range of research interests.
- Researchers should have access to sophisticated data matching techniques that would help them connect and enrich data sources.
- Data sets should be accompanied by datasheets that provide transparency to researchers regarding how the data was collected, its limitations, and any other ethical considerations.

2. *Technology*

The NAIRR platform should provide researchers access to analytical tools to reinforce principles of ethical and responsible innovation which may include, but are not limited to, the following:

- Detect and Mitigate Bias. Capabilities to detect and mitigate bias in data and in models are becoming more widely available. These include, but are not limited to, the ability to assess appropriate representation in data, flag sensitive features, identify proxies for sensitive features, assess model performance differences by sensitive features and their proxies, assess model performance differences within the feature space (independent of sensitive features), offer a range of fairness definitions, and bias mitigation algorithms for selected fairness definitions.
- Explain and Interpret. Capabilities to understand how complex models behave overall (globally) and at the individual observation level (locally) could include, but are not limited to, surrogate model interpretability, explainable machine learning models, natural language explanation of model results, and causal inference.
- Consider Privacy and Security. Capabilities to empower researchers to respect the privacy of data subjects should be offered. These may include the ability to automatically flag protected data and apply privacy preserving techniques to the data

(e.g., differential privacy, encryption, synthetic data generation, etc.).

3. *Ethics training for researchers*

Technology alone cannot solve for equity and other ethical considerations. Researchers should not only ask whether they *can* build data driven systems and insights, but rather whether they *should* in the first place. It requires thoughtful consideration of the impact their work can have if replicated at scale. Familiarizing researchers with data ethics principles of human-centricity, transparency, inclusivity, accountability, robustness, and privacy will be a critical first step to empower them to acknowledge blind spots and build data-driven insights responsibly. Mandatory training should be provided to researchers that emphasize these points before leveraging this platform. Training should also focus on how to translate principles into practice by providing tangible best practices on how to interpret results, detect and mitigate potential biases, and consider privacy and accountability throughout the entire data and model life cycle.

4. *Network of diverse perspectives*

The onus of developing AI technologies and insights responsibly and ethically should not solely fall on the shoulders of researchers leveraging the NAIRR platform. Researchers should have access to a wide and diverse set of perspectives in an effort to mitigate their own blind spots and biases. This will be especially important for researchers who intend to work on high-risk use cases (i.e., researchers could be required to submit a proposal that covers what they intend to do within the platform and a method of determining risk would need to be developed). The NAIRR Task Force should consider providing access to multi-disciplinary and diverse voices to researchers using the platform. Offering researchers access to diverse stakeholders can be enabled through public-private partnerships, partnerships with minority serving institutions (e.g., HBCUs), non-profit entities focused on advancing data ethics and responsible AI, and organizations that aim to bring the voices of impacted communities to the table. The following areas of expertise are recommended:

- Subject matter experts of the data being provided on the platform
- Social scientists
- Data ethics practitioners
- Ethicists
- Equity researchers at the intersection of various domains (e.g., health equity, transportation equity, education equity, etc.)
- The voice of impacted community and stakeholder groups

Researchers should be encouraged to consult with the network of diverse stakeholders as they formulate their research topics, identify candidate data sets for analysis, clean and analyze data, build models, and derive insights from their analysis.

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

Given one of NAIRR's goals is to enable all of America's diverse AI researchers to participate in advancing AI, the Task Force should consider partnerships with minority serving institutions like Historically Black Colleges and Universities (HBCUs) that have developed data science and machine learning curricula at both the undergraduate and graduate levels. Providing these students and researchers with access to rich data sources and NAIRR's robust AI R&D infrastructure will be a critical ingredient to ensure the AI workforce reflects the diversity of America.

There are several examples of research platforms (powered in part by SAS) that facilitate the types activities, resources, and services desired for NAIRR. Two include:

- The National Opinion Research Center (NORC) (out of the University of Chicago) manages a data enclave that is a high-performance computing environment with cutting-edge statistical, analytical, visualization, data management, and reporting tools. Since 2006, state and federal agencies, research institutions, foundations, and universities have used the enclave to securely house and provide remote access to confidential data. Enclave-based research informs a wide spectrum of public and private sector decision-making, as well as journal articles, books, position papers, conference presentations, dissertations, etc. At any given time, the enclave supports over 1,000 researchers via contracts and grants with a wide variety of government, academic, nonprofit, and commercial clients.
- Project Data Sphere is an independent initiative of the CEO Roundtable on Cancer. It leverages experts spanning industry, academia, and government to achieve mutual goals of improving cancer trials in order to expedite drug discovery. As the leading oncology open access data sharing platform, Project Data Sphere hosts de-identified patient-level data contributed by industry, academia, and PDS research programs. By openly sharing data, convening world class experts, and collaborating across industry and regulators to catalyze new scientific insights, Project Data Sphere accelerates delivery of effective treatments to patients. An open-access data-sharing model gives researchers rapid and ready access to data sets and analytical tools.

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

Public-private partnerships (PPPs) bring forward the benefits of public-sector priorities to coordinate collective action together with the innovation capacities and efficiencies of the private sector, including both for-profit and nonprofit organizations. Within the Aritificial Intelligence space, the National Science Foundation-funded National Artificial Intelligence Research Institutes demonstrate the abilities and willingness of publicly funded agencies to work in partnership with private sector organizations to jointly fund research to be done by consortia of public and private institutions of higher education and nonprofit research and development organizations. The Institute of Education

Sciences has funded state agencies to work with private sector service providers to develop statewide longitudinal data systems, some of which SAS has worked with and incorporated AI- and machine language-powered data management and analytics. Additionally, SAS served to advise states on developing new governance, including ownership, structures, and processes, that oversee and operationalize new public-public partnerships between state agencies. There are just a few of the countless examples of federal agencies directing collective interests and action through grants, contracts, and cooperative agreements.

PPPs have the potential to create new capabilities and capacities, particularly through multi-, inter-, and trans-dsciplinary collaboration to better identify and define outstanding questions and the lenses by which solutions are ideated and pursued. SAS' experiences in advising new partnerships, facilitating those partnerships, and being a PPP participant have taught us that effective PPPs have clearly defined common objectives, mutual benefits to participants, and shared resources (e.g., financial, human, physical). Our PPP experiences in education, energy, transportation, water, agriculture, defense and public safety, and health sectors were guided by clear principles that set the directions of the partnership in the form of enabling constraints that maintained flexibilities to better allow for emergent work, behaviors, and dynamics. Less successful PPPs have often been characterized as having highly prescriptive rules that served to overly constrain the activities of the partners and limited the potential innovation and novelty that comes from connecting expertise of the different partners. SAS has continued to accumulate the learnings from our guidance, facilitation, and participation in PPPs across several disciplines and industries, including learnings that serve to accelerate and amplify the benefits of PPPs while also dampening or avoiding those factors that lead to more unproductive pathways.

6. *Where do you see limitations in the ability of the NAIRR to democratize access to AI R&D? And how could these limitations be overcome?*

According to a McKinsey survey of 1,000 leading executives managing analytics initiatives, only 8% had successfully scaled practices from the pilot to production stages. Although the NAIRR platform is intended for research and will not be a true "production" environment, the underlying infrastructure must be production-level quality to mitigate any risks to a truly democratized AI platform through successful management of the shared analytic resources (e.g., curated datasets, models) and scalability of the compute resources.

In SAS' extensive experience working with academia and government, the false dichotomy pitting open source analytics tools, such as Python and R, against commercial off the shelf (COTS) platforms, such as SAS, DataBricks, and IBM, creates unnecessary choices. The dichotomy drives the disconnect between the promise of democratized AI and successful, trusted results. Failed AI projects typically choose a strategy heavily

dependent on a specific set of tools and then face the consequences when the results are either disjointed and ungoverned or closed off and vendor dependent.

SAS has learned that successful projects embrace the dynamism, access and choice offered through open source tools while also gaining the governance and decreased time to value offered by commercial platforms. This combined industrialized analytics approach can offer democratized AI in a manner that benefits the field quickly and effectively.

Another limitation is access to the compute infrastructure required by AI research. Underserved communities may lack the financial resources to procure and manage appropriate infrastructure, thereby inequitably limiting access to developmental opportunities for budding researchers. NAIRR can partially overcome this limitation by providing and managing the infrastructure, and by hosting the AI research platform in a web-accessible location so that a more diverse range of researchers can access AI methods without needing their own infrastructure.

In summary, NAIRR can help democratize access to AI by developing a technology solution which:

- Prioritizes multi-model machine learning of various programming languages as well as non-coding approaches to accommodate the various skill sets and preferences of the research community.
- Takes advantage of various high performance computing techniques so that models can be processed quickly and that researchers have the ability to validate innovative ideas on sufficiently large datasets.
- Automates performance tracking and re-training of models to ensure optimal results as additional data are accumulated and keeps the focus on innovation.
- Includes or integrates with a model inventory to maintain documentation, versioning and model lineage within a governed platform.
- Provides built-in model interpretability capabilities to help researchers explain results and uncover bias.
- Incorporates centralized governance to control access to and usage of data and models across various stakeholders.

To contact your local SAS office, please visit: sas.com/offices