Harvard Data Science Review • Special Issue 4: Democratizing Data

Enabling Responsible Artificial Intelligence Research and Development Through the Democratization of Advanced Cyberinfrastructure

Manish Parashar[1]

[1]Scientific Computing and Imaging (SCI) Institute, University of Utah, Salt Lake City, Utah, United States of America

Published on: Apr 02, 2024

DOI: https://doi.org/10.1162/99608f92.9469c089

ABSTRACT

Artificial intelligence (AI) is driving discovery, innovation, and economic growth, and has the potential to transform science and society. However, realizing the positive, transformative potential of AI requires that AI research and development (R&D) progress responsibly; that is, in a way that protects privacy, civil rights, and civil liberties, and promotes principles of fairness, accountability, transparency, and equity. This article explores the importance of democratizing AI R&D for achieving the goal of responsible AI and its potential impacts.

Keywords: responsible AI, data democratization, FAIR data, open and equitable access, NAIRR, National Data Platform

1. Introduction

Artificial intelligence (AI) is rapidly becoming the transformative technology of the 21st century, driving innovations, enabling new discoveries, and spurring economic growth. It is impacting everything from routine daily tasks and services to addressing societal-level grand challenges—it has the potential to revolutionize solutions to many scientifically and societally important problems and to improve the lives of every individual, especially those with special needs (such as elderly people) and who have been traditionally underserved. At the same time, there are growing concerns that AI could have negative social, environmental, and even economic consequences. To realize the positive and transformative potential of AI, it is imperative to advance AI and its applications responsibly; that is, in a way that achieves societal good while also protecting privacy, civil rights, and civil liberties, and promotes principles of fairness, accountability, transparency, and equity.

Democratizing AI research and development (R&D) so that all researchers from every background can participate in foundational, use-inspired, and translational AI R&D and contribute to the AI R&D ecosystem has been highlighted as essential to mitigating these negative impacts by the National AI Research Resource (NAIRR) Task Force (TF)[1] as outlined in their final report (NAIRR, 2023). Recognizing that, today, advances in AI R&D are very often tied to access to large amounts of computational power and data, the TF highlighted the importance of a widely accessible AI research cyberinfrastructure (CI) that brings together computational resources, data, testbeds, algorithms, software, services, networks, and expertise that can help democratize the AI R&D landscape and enable responsible AI R&D that benefits all.

This article explores the importance of democratizing AI R&D to achieve the goal of responsible AI and the resulting imperative of democratizing access to state-of-the-art advanced CI. It further highlights some of the key barriers preventing such democratizing access, and the vision of the NAIRR TF for addressing these barriers. It then introduces the National Data Platform, a recently launched project aimed at catalyzing an open, equitable, and extensible data ecosystem, as one effort toward democratizing data. While these efforts are U.S.-

based, they can provide a means for engaging and cooperating with similar (and complementary) efforts internationally toward a more global CI.

2. The Importance of Democratizing Access to Advanced CI

Progress at the current frontiers of AI is increasingly tied to access to large amounts of computational power and data, and such access today is too often limited to researchers at well-resourced organizations (e.g., large companies, well-funded institutions, etc.). This unequal access is resulting in a large and growing resource divide that can limit and adversely skew our AI R&D ecosystem. For example, this imbalance can lead to data sets, AI models, and applications that are biased and lack fairness and can threaten our ability to cultivate a diverse AI research community and workforce.

A widely accessible CI that brings together computational resources, data, testbeds, algorithms, software, services, networks, user training and support, and expertise is thus critical to addressing these concerns, and for democratizing the AI R&D landscape. Such an accessible CI would ensure that all researchers have the opportunity to contribute to the research enterprise, broadening the range of researchers who are able to be involved in AI R&D and at the same time, growing and diversifying approaches to, and applications of, AI. It would also open new opportunities for AI-enabled progress across all scientific fields and disciplines, as well as support critical research areas such as AI auditing, testing and evaluation, trustworthy AI, bias mitigation, and AI safety. Similarly, broad, equitable access to diverse AI-ready data repositories (as associated computing and software) is essential to develop, validate, and deploy fair and responsible AI models and reduce bias. Increased CI access would also result in a diversity of perspectives, which would, in turn, lead to new ideas that may not occur otherwise, and to AI systems that are responsible and inclusive by design.

3. Barriers to Equitable CI Access and Use

Despite global investments and the growing availability of CI and services, significant barriers still limit broad and equitable access to this CI ecosystem, especially for individuals and institutions that are resource constrained and for communities that have been traditionally underrepresented. Open and equitable access to CI and its integration into research workflows can be challenging, particularly causing inequities for under-resourced communities and researchers. Knowledge, technical and social barriers, and challenges were explored in the *Missing Millions* report (Blatecky et al., 2021) and were highlighted in recent papers (Parashar, 2022; Parashar & Altintas, 2023). Key barriers include:

Knowledge barriers refer to the lack of a broad awareness of available CI resources, services, and expertise; how they can be used; and the critical need for support structures often missing at the local level, especially at under-resourced institutions. A related barrier is the recruitment, retention, and cultivation of a skilled, diverse, and agile workforce, which needs opportunities and mechanisms for training (including reskilling and upskilling), mentoring, recognition, and professional development.

Technical/Procedural barriers include processes, protocols, mechanisms, and infrastructure for getting access to and using CI and that are often limited by local infrastructure and capabilities. Local resources and capabilities often are not equitably available across the full range of institution types, preventing certain segments of the research community from accessing and using even freely shared CI resources and services.

Social barriers at the institutional and regional levels impact how access to data and data services are viewed, funded, and supported. For example, the importance of access to CI and the support needed to make it happen may not be appreciated at an institutional level, resulting in a lack of mechanisms and structures needed to support researchers. This lack of appreciation of CI can further perpetuate and amplify the impacts of the knowledge and technical barriers, and once again, can disproportionately affect under-resourced institutions and communities.

4. Democratizing Access to CI

Democratizing access to the advanced CI ecosystem, including the computing resources and data sets that are essential to AI R&D, and ensuring every researcher has fair and equitable access to the resources are key to achieving the positive potential of AI in a responsible manner. As a result, strategic investments in resources, services, and expertise at local, regional, and national levels are needed to systematically address barriers to CI access while at the same time providing adequate training and support structures.

In the United States, democratizing access to CI is emphasized as central to the country's competitiveness in science and technology and is essential for critical science-driven decision-making for addressing important and urgent national and global issues, such as climate change and environmental sustainability (Parashar et al., 2022). For example, the Nelson Memo from the White House Office of Science and Technology Policy (OSTP; 2022b) sets the ambitious goal of providing free, immediate, and equitable access to U.S. federally funded research, including publications, and underlying scientific data. Similar goals have been set by other nations and groups, such as the Plan S initiative for Open Access publishing, launched by cOAlition S, a group of national research funding organizations, with the support of the European Commission and the European Research Council (ERC).[2]

Addressing the barriers and challenges noted above requires increasing awareness and access by deploying more equitable mechanisms and services for CI discovery (e.g., user-friendly portals and gateways) and access (e.g., flexible allocation mechanisms and access modes that support a diversity of users and application needs); integrating users and locally accessible CI resources as part of a shared fabric of national CI resources through high-speed frictionless data networking; establishing more accessible support structures and integrating and embedding these support structures within communities (e.g., easily accessible, and scalable networks of experts responsive to local needs), creating methods for education and training, on-ramping, mentoring, and support that target the spectrum of CI users and skills, promotes success and advancement, and facilitates exchanges between communities and the dissemination of best practices.

5. The Vision and Goals of the NAIRR

An urgent vision for U.S. leadership in responsible AI and a strategic implementation plan for strengthening and democratizing AI innovations were presented in a January 2023 final report of the NAIRR TF (2023). As outlined in this report, the NAIRR is envisioned as a widely accessible, national cyberinfrastructure that will advance and accelerate the US AI R&D environment and fuel AI discovery and innovation in the United States by empowering a diverse set of users across a range of fields through access to computational, data, software, and training resources, along with testbeds. Created by leveraging, linking, and augmenting the nation's existing cyberinfrastructure resources, the NAIRR would support cutting-edge explorations in AI R&D and improve the ease of collaboration across disciplines. If realized, it would create opportunities to train the future AI workforce, support, and advance trustworthy and responsible AI, and catalyze the development of ideas that can be practically deployed for societal and economic benefits.

The NAIRR would accelerate these outcomes by enabling U.S.-based researchers to access the digital resources that enable AI R&D. These resources would be made available through an integrated user portal with key functionalities such as single sign-on access to resources, collaboration tools, search tools for resource discovery, detailed resource specifications and user guides, an interface for computational job submission, and consolidated accounting of resource use. User support services and interactive training modules would support users new to the field, which, along with clearly defined policies and standards of practice, would promulgate best practices for trustworthy AI model development and responsible data use by design. A publicly accessible NAIRR user portal would provide curated catalogs that list commonly used AI data sets, testbeds, educational resources, and relevant metadata, serving as a clearinghouse for the AI R&D community. Through a tiered-access model, vetted researchers would be able to conduct research on sensitive or restricted data in secure enclaves.

The NAIRR TF recommended that the NAIRR be established with four measurable goals in mind: (1) spur innovation, (2) increase diversity of talent, (3) improve capacity, and (4) advance trustworthy AI. The NAIRR would meet these goals by supporting the needs of researchers and students from diverse backgrounds who are pursuing foundational, use-inspired, and translational AI research, with the overarching goal of lowering barriers to participation in the AI research ecosystem and increasing the diversity of AI researchers.

The NAIRR is envisioned as a shared national cyberinfrastructure—a broadly accessible federated mix of computational and data resources, testbeds, software, and testing tools. These resources would be made available through an integrated portal, with available training tools and user support services to facilitate their use. Computational resources would include conventional servers, computing clusters, high-performance computing, and cloud computing, and would support access to edge computing resources and testbeds for AI R&D. Open and protected data would be made available under tiered-access protocols and co-located with computational resources, and an ‘AI commons’ where data resources could be contributed by NAIRR users or stakeholders would be explored. When fully implemented, the NAIRR would address both the capacity (ability to support a large number of users) and capability (ability to train resource-intensive AI models) needs of the AI research community. Note that the exact balance of resources across the different resource types, that is, compute (capacity and capability), data, services and tools, testbeds, training, and so on, will involve many tradeoffs that should be driven by the overarching goal of the NAIRR of democratizing AI R&D, the needs of the user community, as well as other resources that may be available across the broader AI R&D ecosystem, for example, at academic institutions or industry, and gaps therein. For example, access to AI-ready data sets and data services remains a challenge in the academic community and may be prioritized by the NAIRR. Similarly, ensuring resource access for users who would not otherwise have access to resources would need adequate capacity for computing and training resources. Optimizing these tradeoffs would be an iterative process and would change as the user community evolves. Additionally, incentives for resource sharing by users and other stakeholders, such as the AI Commons mentioned above can also help increase available resources.

The NAIRR would establish multiple allocation processes based on the nature, size, and scope of the requests, agency-driven allocations, peer-reviewed research allocations, and expedited startup allocations. Users would be able to discover and access resources through an integrated NAIRR portal. For more advanced users, opportunities to access resources directly from providers would also be made available.

To protect the resources it makes available, the NAIRR would implement system safeguards using government-applicable security guidelines, particularly those established by the National Institute of Standards and Technology (NIST). Most importantly, the NAIRR TF recommended that the NAIRR should set the standard for responsible AI research through the design and implementation of its governance processes. The TF emphasized that the NAIRR must be proactive in addressing privacy, civil rights, and civil liberties issues by integrating appropriate technical controls, policies, and governance mechanisms from the outset, including criteria and mechanisms for evaluating proposed research and resources for inclusion in the NAIRR from a privacy, civil rights, and civil liberties perspective. Regular training should be required to build NAIRR users' awareness about rights, responsibilities, and best practices related to privacy, civil rights, and civil liberties in AI research, in accordance with the *Blueprint for an AI Bill of Rights* published by the White House OSTP (2022a), and well as the AI Risk Management Framework developed by NIST (U.S. Department of Commerce, 2023).

Recognizing the global nature of the AI R&D ecosystem and the benefits of international cooperation and sharing in advancing AI research, the NAIRR TF recommended exploring mechanisms for cooperation with similar resource infrastructure efforts around the world, in alignment with guidelines and frameworks setup by the U.S. government for such cooperation and sharing. The NAIRR TF also recommended leveraging existing international forums such as the International Science Council’s Committee on Data and the Global Partnership on AI to support ongoing international collaborations and foster new opportunities.

6. Toward Democratizing Data: Envisioning a National Data Platform

The overarching goal of the recently funded National Data Platform (NDP)[3] (Parashar & Altintas, 2023) is to respond to the barriers and challenges in accessing data, data CI, and data services, and to enable data providers, data users and educators to generalize data pipelines so that they can be equitably used by all researchers at the national scale. Specifically, NDP aims to bridge existing gaps in the foundational data CI to: (1) federate often siloed data repositories into a unified discovery and access platform; (2) integrate them with advanced computing CI; and (3) provide open and equitable access via standardized processes and customizable services for ingestion, indexing, curation, and analysis. Through a `removing-the-barriers' approach combining needs assessment, co-design, and user capacity building with existing ready-for-scale data CI capabilities, NDP aims to enable the necessary data discovery, wrangling, and knowledge management services. These services will be available to a wide community of researchers, enabling them to collaborate effectively through unified platforms. More importantly, they will enable researchers to move away from one-off ad hoc solutions. NDP will aim to answer the following questions from systems, services, and open data perspectives:

- What are the foundational data abstractions and services that can serve as multipurpose and expandable building blocks for data-driven and AI-integrated application patterns, and how can everyone effectively access and utilize these abstractions and services?
- How can such abstractions and services be developed and deployed on top of existing production-ready CI from storage to the edge-to-high-performance-computing computing continuum to ensure equity of access and use?
- What are the governance and open science, open data, and open CI requirements and challenges, and what are the required guardrails for protecting privacy, civil rights, and civil liberties that will ensure a more equitable use of such data systems and services for everything from education to new AI training and application development?

Architecturally, NDP is envisioned as a federation to diverse data sets on top of the existing and evolving national CI capabilities. NDP will provide the necessary discovery, wrangling, and knowledge management services for data across open national CI with an end goal of increasing availability of trusted open AI-ready data sets and model benchmarks to enable AI advancement. These data and model services will be equitably accessible by a wide community of researchers following the FAIR (Wilkinson, 2016) and CARE (Carroll et al, 2020) principles along with the necessary education and training, to enable them to use the data and collaborate effectively.

7. Conclusions

AI is rapidly becoming an integral part of all aspects of science and society, driving discovery, innovation, and economic growth, and it has the potential to have a transformative impact. Consequently, it is critical that we ensure that AI R&D progresses responsibly, in a way that protects privacy, civil rights, and civil liberties, and promotes principles of fairness, accountability, transparency, and equity. This article highlighted the importance of democratizing AI R&D for achieving the goal of responsible AI, the resulting imperative of democratizing access to advanced cyberinfrastructure, and the key barriers to such democratized and equitable access to CI resources, services, and expertise. It then summarized the vision of the NAIRR TF for addressing these barriers and democratizing AI R&D and introduced the National Data Platform as an example of democratizing CI. While these projects are U.S.-based, they can provide a basis for international cooperation and sharing.

Acknowledgments

The author would like to acknowledge the tremendous efforts and dedication of the NAIRR Task Force members. The author is also grateful for the discussions and thoughtful feedback from a wide range of colleagues.

Disclosure Statement

Manish Parashar has no financial or non-financial disclosures to share for this article.

References

Blatecky, A., Clarke, D., Cutcher-Gershenfeld, J., Dent, D., Hipp, R., Hunsinger, A., Kuslikis, A., & Michael, L. (2021). *The missing millions: Democratizing computation and data to bridge digital divides and increase access to science for underrepresented communities*. National Science Foundation.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1), Article 43. https://doi.org/10.5334/dsj-2020-043

National Artificial Intelligence Research Resource Task Force. (2023, January). *Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a National Artificial Intelligence Research Resource* [Final Report]. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

Parashar, M. (2022, September). Democratizing science through advanced cyberinfrastructure. *Computer*, *55*(9), 79–84. https://doi.org/10.1109/MC.2022.3174928

Parashar, M., & Altintas, I. (2023, October 9–13). Toward democratizing access to science data: Introducing the National Data Platform. In *2023 IEEE 19th International Conference on e-Science*, Limassol, Cyprus (pp. 1–4). https://doi.org/10.1109/e-Science58273.2023.10254930

Parashar, M., Friedlander, A., Gianchandani, E., & Martonosi, M. (2022). Transforming science through cyberinfrastructure. *Communications of the ACM*, *65*(8), 30–32. https://doi.org/10.1145/3507694

U.S. Department of Commerce. (2023, January). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

White House Office of Science and Technology Policy. (2022a, October). *Blueprint for an AI bill of rights: Making automated systems work for the American people* [White paper]. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

White House Office of Science and Technology Policy. (2022b). *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research* [Memo]. Executive Office of the President of the United States. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf

Wilkinson, M., Dumontier, M., Aalbersberg, I. G. Appleton, M. Axton, A. Baak, N. Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

Footnotes

1. The NAIRR Task Force was a federal advisory committee that ran from June 2021 through April 2023, with the overarching goal of strengthening and democratizing the U.S. AI innovation ecosystem in a way that protects privacy, civil rights, and civil liberties. Additional information about the task force and its work is available at https://ai.gov/nairrtf ↩

2. https://www.coalition-s.org/ ↩

3. https://www.nationaldataplatform.org/ ↩

References

- Blatecky, A., Clarke, D., Cutcher-Gershenfeld, J., Dent, D., Hipp, R., Hunsinger, A., Kuslikis, A., & Michael, L. (2021). *The missing millions: Democratizing computation and data to bridge digital divides and increase access to science for underrepresented communities*. National Science Foundation.

  ↩
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, *19*(1), Article 43. https://doi.org/10.5334/dsj-2020-043

  ↩
- National Artificial Intelligence Research Resource Task Force. (2023, January). *Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a National Artificial Intelligence Research Resource* [Final Report]. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

  ↩
- Parashar, M. (2022, September). Democratizing science through advanced cyberinfrastructure. *Computer*, *55*(9), 79–84. https://doi.org/10.1109/MC.2022.3174928

  ↩
- Parashar, M., & Altintas, I. (2023, October 9–13). Toward democratizing access to science data: Introducing the National Data Platform. In *2023 IEEE 19th International Conference on e-Science*, Limassol, Cyprus (pp. 1–4). https://doi.org/10.1109/e-Science58273.2023.10254930

  ↩

- Parashar, M., & Altintas, I. (2023, October 9–13). Toward democratizing access to science data: Introducing the National Data Platform. In *2023 IEEE 19th International Conference on e-Science*, Limassol, Cyprus (pp. 1–4). https://doi.org/10.1109/e-Science58273.2023.10254930

  ↩

- Parashar, M., Friedlander, A., Gianchandani, E., & Martonosi, M. (2022). Transforming science through cyberinfrastructure. *Communications of the ACM*, *65*(8), 30–32. https://doi.org/10.1145/3507694

  ↩
- U.S. Department of Commerce. (2023, January). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

  ↩

- White House Office of Science and Technology Policy. (2022a, October). *Blueprint for an AI bill of rights: Making automated systems work for the American people* [White paper]. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

  ↩
- White House Office of Science and Technology Policy. (2022b). *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research* [Memo]. Executive Office of the President of the United States. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf

  ↩
- Wilkinson, M., Dumontier, M., Aalbersberg, I. G. Appleton, M. Axton, A. Baak, N. Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

  ↩