August 9, 2019

Russell T. Vought
Office of Management and Budget
725 17th St NW
Washington, DC 20503

Dear Mr. Vought,

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit comments in response to the Office of Management and Budget's (OMB) request for information on identifying priority access or quality improvements for federal data and models for artificial intelligence (AI) research and development (R&D) and testing.[1]

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C., and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as important data-related technology trends. The Center is a non-profit, non-partisan research institute affiliated with the Information Technology and Innovation Foundation.

Federal data and models can serve as an incredibly important public resource for the advancement of AI and it is encouraging that OMB is exploring how the federal government provides this resource. To maximize the effectiveness of this initiative, OMB should institutionalize a demand-driven framework that establishes a permanent channel for industry, academia, and other AI stakeholders to help identify and prioritize the publication of and improvements to the highest-value models.

Which models are most important for agencies to focus on, and why?

The most important models agencies should focus on will likely be those with either broad applicability, such as general purpose object detection models, or specialized applicability in sensitive and high-value contexts, such as in the key domains this RFI identifies, including healthcare and transportation. It is encouraging that OMB is interested in identifying the highest-priority models to focus on, however OMB can substantially bolster this effort by creating a permanent, dedicated

[1] "Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing; Request for Information," Federal Register, July 10, 2019, https://www.federalregister.gov/documents/2019/07/10/2019-14618/identifying-priority-access-or-quality-improvements-for-federal-data-and-models-for-artificial.

channel to solicit feedback about in-demand data and models. While OMB can and should act on the recommendations about which models are the most important that others identify in their responses to this RFI, AI R&D is rapidly evolving, and the needs of the AI development community will likely similarly evolve and look considerably different two years from now than it does today. Thus, to maximize the value the federal government can provide to AI R&D, it should consider the identification of high-priority models a continuous endeavor.

There is some precedent for this continuous, demand-driven approach in the federal government's provision of open data. In 2015, the Department of Health and Human Services (HHS) launched its Demand-Driven Open Data (DDOD) program to improve how the agency prioritizes the release of high-value datasets.[2] Though the federal government is required to treat all of its data as open data, many agencies must prioritize the release of certain datasets rather than publish everything at once, due to resource constraints. However, this caused data owners in federal agencies to prioritize datasets that were the easiest and least-risky to publish, rather than the highest-value ones.[3] HHS launched the DDOD program to remedy this by creating a channel for data users to request the publication of certain datasets by articulating valuable use-cases for them, ensuring that HHS could maximize the impact of the limited resources available for the publication of open data.[4]

OMB should provide federal agencies guidance on how to replicate this demand-driven approach for the publication of models useful for AI R&D across the federal government. This would ensure that agencies can better meet the needs of the AI development community on a continuous basis. Additionally, on a macro level, this would allow for OMB to identify trends in demand for certain kinds of models, enabling the agency to develop more concrete, specific policies about the publication of these resources.

What are key gaps in data and model availability that are slowing progress in AI R&D and testing? Which areas of AI R&D and testing are most impacted?

There are a variety of areas where a lack of publicly available datasets hinder progress in AI development. One notable example is the lack of a common, authoritative facial recognition training dataset for the public and private sectors. While many U.S. companies developing facial recognition technology (FRT) invest heavily in developing this resource for proprietary use, historically, the training data available to developers overwhelmingly consists of white, male faces, causing many

[2] Amanda Ziadeh, "HHS On a Mission to Liberate Health Data," *GCN*, June 5, 2015, https://gcn.com/articles/2015/06/05/hhs-data-liberation.aspx.
[3] Ibid.
[4] "Demand-Driven Open Data," U.S. Department of Health and Human Services, https://www.hhs.gov/cto/projects/demand-driven-open-data/index.html (Accessed August 2, 2019).

facial recognition systems to underperform for minorities and women.[5] Though the private sector has an incentive to make facial recognition algorithms as accurate and reliable as possible, developing a representative dataset of hundreds of thousands of faces requires considerable resources, and even if a company were to do that, it has little incentive to share this data with potential competitors. Recognizing this challenge, IBM announced in June 2018 that it would publish the world's largest annotated dataset of faces specifically for the purposes of studying bias in facial analysis.[6] This is encouraging, however not only is this insufficient to solve the problem of biased facial recognition systems but overcoming this challenge should not just be the responsibility of the private sector.

The National Institutes of Standards and Technology (NIST) conducts evaluations of FRT systems from a wide variety of developers using a series of benchmarking datasets as part of its Ongoing Facial Recognition Vendor Test (FRVT).[7] The FRVT datasets are very large, ranging from 1,000 to 1,000,000 images and are specialized to evaluate an FRT algorithm's fitness for use in various high-value contexts, including identifying faces in child exploitation images, visa images, and mugshot images.[8] NIST did not make these datasets available to developers as doing so could cause developers to train their FRT systems on just those images, allowing them to score highly on the FRVT but perform worse for more general use.[9] This is understandable. However, NIST should develop a comparable resource that it makes publicly available, to the extent reasonable. (NIST should not be making images of child exploitation publicly available, for example.)

A similar gap exists in the field of genomics, as a lack of large amounts of easily accessible, representative genetic data limit the potential for AI R&D in the space. Fortunately, the federal government is already working to address this problem. In 2016, the National Institutes of Health (NIH) launched its Precision Medicine Initiative (PMI) and established its All of Us Research Program, which aims to establish a cohort of 1 million people who donate data about their genetics, medical

[5] Steve Lohr, "Facial Recognition Is Accurate, if You're a White Guy," The New York Times, February 9, 2018, https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.
[6] IBM, "IBM to Release World's Largest Annotation Dataset for Studying Bias in Facial Analysis," News Release, June 27, 2018, https://www.ibm.com/blogs/research/2018/06/ai-facial-analytics/.
[7] Patrick Grother, Mei Ngaan, and Kayee Hanaoka, *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification* (Washington, DC: National Institute of Standards and Technology, April 2019), https://www.nist.gov/sites/default/files/documents/2019/04/04/frvt_report_2019_04_04.pdf.
[8] Ibid.
[9] "NIST Evaluation Shows Advance in Face Recognition Software's Capabilities," NIST, News release, November 30, 2018, https://www.nist.gov/news-events/news/2018/11/nist-evaluation-shows-advance-face-recognition-softwares-capabilities.

histories, lifestyles, and other factors to accelerate precision medicine research.[10] NIH is currently developing a research protocol to ensure broad access to its data in ways that meets researchers' needs while still protecting privacy and security.[11] Other agencies should look to initiatives like the All of Us Research Program as a model for voluntary data collection and sharing in areas where there is a need for more data.

Finally, OMB should track agency progress in implementing the OPEN Government Data Act. As of August 2019, it appears that the Department of State, the Department of the Interior, the Department of the Treasury, and the Small Business Administration had missed their deadline to appoint a chief data officer.[12] Many federal agencies collect data that may have value for AI, but if they do not release open data then it will not be available for this or other purposes.

Overall, the most effective way for the federal government to identify gaps in data and model availability would be able to implement a demand-driven approach described in this filing by creating a process that allows stakeholders to regularly submit requests for new data sets, as described in the DDOD program.

What data ownership, intellectual property, or data sharing considerations should be included in federally-funded agreements (including, but not limited to, federal contracts and grants) that results in production of data for R&D?

While considerations about data ownership, intellectual property, and data sharing in federally-funded contracts, grants, and other agreement will necessarily vary depending on context, as a rule, federally-funded agreements that result in the production of data, for R&D or other purposes, should maximize the availability and reuse of this data, ideally by assigning it an open license and using a non-proprietary format whenever reasonable.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
dcastro@datainnovation.org

Joshua New
Senior Policy Analyst
Center for Data Innovation
jnew@datainnovation.org

[10] "About the *All of Us* Research Program," National Institutes of Health, Accessed July 26, 2019, https://allofus.nih.gov/about/about-all-us-research-program.
[11] "Researchers as Partners," National Institutes of Health, Accessed July 26, 2019, https://www.researchallofus.org/about/researchers-as-partners/.
[12] "Agencies are now requited to have a chief data officer. Do they?" *Workscoop,* August 5, 2019, https://workscoop.com/2019/08/05/federal-chief-data-officer-evidence-based-policymaking-deadline.