Date: 23 May 2019

To: The National Institute of Standards and Technology

From: Dr. Michael Stumborg, Dr. Christine Hughes, Center for Naval Analyses

Subject: RFI: Developing a Federal AI Standards Engagement Plan

Background

Researchers from the Center for Naval Analyses, the Department of the Navy's Federally Funded Research and Development Center, hereby submit this response to the subject NIST RFI. The information herein is the opinion of the authors only. It does not necessarily represent the opinion of the Department of Defense, the Department of the Navy, or the Center for Naval Analyses.

The information provided pertains in general to "technical standards and guidance that are needed to establish and advance trustworthy aspects (e. g., accuracy, transparency, security, privacy, and robustness) of AI technologies," and specifically to "reference data and datasets" as AI tools.

Despite decades of effort by AI researchers, the current state of the art is limited mostly to machine learning algorithms that depend on training data. It is axiomatic then, that we cannot expect high quality AI applications without high quality (training) data.

AI is similar to many other analytic pursuits in that analysts spend upward of 80% of their time discovering, gaining access to, and "cleaning" data before it can be used by their machine learning algorithms. It is also axiomatic then, that we cannot expect to accelerate the adoption of AI without first making training data easily accessible to AI practitioners.

To secure American leadership in Artificial Intelligence, we propose that NIST focus considerable effort on developing Data Accessibility Standards and Data Quality Standards.

A thriving cottage industry has emerged over the past several years within academia, think tanks, and journalism highlighting and criticizing AI applications that have negative impacts on socio-economically disadvantaged groups. In many of these instances, it is poor quality training data, and not poor quality machine learning algorithms that are to blame. The inability of third parties to access training data to conduct oversight only exacerbates the fear of these wayward AI applications. Inaccessible, poor quality training data puts the acceptance of AI at risk in American society. A failure to address data accessibility and data quality may cede AI leadership to nations with a demonstrated low regard for their own citizens' rights and well-being. Given the vast potential economic and military advantages bestowed by AI leadership, and the relationship between AI advancement and the accessibility of high quality training data, establishing Data Accessibility Standards and Data Quality Standards rises to the level of a National imperative that NIST is in a position to both oversee and provide advocacy for.

Basic Purpose of the Proposed Data Standards

The basic purpose of the proposed data standards is to *measure and make known* the quality and accessibility of data sets, so that potential users can make informed decisions about the data's applicability to their purpose, and so that third party oversight can occur as a check on the potential for negative consequences of AI applications.

Characteristics of Data Access Standards

Data Access Standards support American AI leadership by making the training data needed for machine learning applications more visible and more accessible to all authorized users. These standards should therefore describe the attributes of data that define authorized use. Such attributes include, but are not necessarily limited to, US government security classification, the presence of law enforcement sensitive data, proprietary data, acquisition-sensitive data, personally identifiable information (to include biographic, biometric and contextual data for individuals), Freedom of Information Act (FOIA) exemptions, and even fees that might be required for data access.

Characteristics of Data Quality Standards

Data Quality Standards support American AI leadership by helping authorized users to rapidly evaluate data before investing time seeking access to it. Note that we have repeatedly referred to "poor quality" data without defining the attributes that make it so. Stakeholders from data consuming communities must identify and agree to these attributes and the standards to measure them. Each data quality attribute would preferably also have a quantitative value associated with it, thus providing a measureable standard.

Just as an example, Department of Defense policy uses five general attributes to describe data quality. In addition to being visible and accessible as described above, data should also be interoperable, understandable, and trustworthy. Each of these attributes can be subdivided further. Additional community-specific attributes can be defined to serve the unique needs of individual data consuming communities with data quality requirements not shared by other communities. In another example, the International Association for Information and Data Quality has a glossary of terms that might serve to further subdivide trustworthiness: To be considered trustworthy, data must be accurate, auditable, complete, consistent, credible, current, and timely.

These are just examples because it would be presumptuous to propose the exact content and structure of Data Quality Standards here. That is a task that must be undertaken by the stakeholder communities of data consumers, guided to consensus by an authoritative standards development organization such as NIST.

Implementation of Data Accessibility and Data Quality Standards

Once data standards are developed, they can be appended to data sets as metadata tags and recorded in data set registries. The registry would contain the metadata describing not just the content of the data set, but also the accessibility and quality compliance of the data sets within. The registry need not contain the actual data – so long as the metadata also contains information regarding the data

steward to contact to request access to the actual data set. Prospective users could query the data registry to identify data sets that they are permitted to access, and that are of a quality level that meets their needs. Once the available and appropriate data is identified, the prospective user can request and arrange data access.

Data set registries could be established, vetted, and populated by US Government organizations that fund the work that generates data. Receipt of government funding could be made contingent upon the data producer's agreement to populate the appropriate registry with data that adheres to the proposed Data Quality and Data Access Standards (once developed). The government agencies that make funding contingent on compliance will automatically meet many of the requirements of the OPEN Government Data Act (OGDA), and with minimal additional work.

Secondary Benefits

Recruiting and Retention of AI Talent

Currently, there is fierce competition for people with AI skills. They command high starting salaries that the federal government finds difficult to match under the restrictions of the GS schedule. The federal government must be competitive in this job market to realize American leadership in AI. Unable to compete fully on salary alone, the federal government often has one advantage over the private sector: important and meaningful problems to solve with AI and ownership of particular data. Adoption of Data Accessibility Standards and Data Quality Standards may provide the federal government with an additional advantage: a less menial work environment. As noted above, the typical analyst spends upwards of 80% of their time doing the "menial" tasks of discovering, accessing, and cleaning low quality data. Only then can they do the more exciting work of applying AI to important and meaningful problems. Developing and using the proposed data standards would relieve the typical analyst of much of this menial burden, allowing them to focus on the part of the job that motivates and excites them.

Cost-efficient AI Development

It makes little sense to hire highly skilled AI practitioners and then force them to conduct tedious tasks that do not require their advanced skill sets. Data standards promote the more efficient utilization of government AI talent for the same reasons that they improve recruiting and retention (reduction in menial tasking).

Big Data Analytics Acceleration

Because machine learning still represents the state-of-the-art in AI, and because machine learning is also a critical technique in Big Data Analytics, there is considerable overlap between these separate, but similar fields. The arguments above in favor of Data Accessibility Standards and Data Quality Standards for recruiting and retaining public sector AI practitioners, apply also to in-demand Big Data Analytics practitioners. The arguments on cost-efficiency also hold for Big Data Analytics.

Note that in all of the above discussion we refer to "data" standards, not "AI training data" standards. While the intent of this RFI is to support AI leadership, developing the proposed standards will support just about any data-intensive analytical activity, to include AI.

Compliance with Transparency Laws

The Freedom of Information Act (FOIA) has been in force since 1967. In 2019 OGDA mandated standards to facilitate public access to government data. Data Access Standards would "pre-define" the authorized users for data sets and/or data elements within data sets. In the absence of these standards, every public FOIA or OPEN data request must be adjudicated individually on a case-by-case basis. Pre-defining what data is exempt from FOIA and/or OGDA requests through the use of Data Accessibility Standards has the potential to at least partially automate the labor-intensive adjudication process.

Conclusion

AI applications are only as good as the data used to train them. AI standards at the application layer are necessary, but not sufficient to accelerate the use and facilitate the acceptance of AI. Standards to improve the accessibility and the quality at the data layer are also required. Data Accessibility Standards and Data Quality Standards should be an integral part to any standardization effort aimed at promoting a coherent path toward American AI supremacy.