DATA
COALITION

Data Coalition
1003 K Street NW, Suite 200
Washington, D.C. 20001

@DataCoalition
info@datacoalition.org
datacoalition.org

August 9, 2019

Subject: Data Coalition Comments on AI Data and Model R&D RFI (FR Doc. 2019-14618)

Dear Director Vought:

Please accept the Data Coalition’s comments on the Office of Management and Budget’s Request for Information on “Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing” (FR Doc. 2019-14618).

The Data Coalition is America’s premier voice on data policy. The Data Coalition and its members advocate for government information to be high quality, accessible, and useful. In most cases, this means transforming data to be standardized, open, and machine-readable information. Based in Washington D.C., the Data Coalition members include a broad cross-section of the data industry, including technology and data analysis companies, public sector consulting firms, and non-profit organizations.

The comments from the Data Coalition—prepared with assistance from George Schoeffel and Doug Hummel-Price as well as feedback from the Coalition’s member companies—specifically encourage OMB to articulate a clear definition of artificial intelligence. Our comments also include suggestions for types of data sets to prioritize as well as opportunities to improve data quality and standards.

Should you have additional questions about our comments, please contact me at nick.hart@datacoalition.org.

Regards,

Nick Hart, Ph.D.
CEO
Data Coalition

DATA COALITION COMMENTS ON ARTIFICIAL INTELLIGENCE REQUEST FOR INFORMATION

1 – KEY DEFINITIONS FOR ARTIFICIAL INTELLIGENCE

DEFINITIONS FOR "DATA", "DATA SET", AND "MODEL"

Given that the *Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing; Request for Information* (RFI) hinges on the use of a number of key terms, it is important that the Data Coalition provides a definition and understanding for each of these terms before our response to the RFI. When discussing artificial intelligence (AI), the Data Coalition defines AI as proposed by draft legislation the AI in Government Act of 2019 (S. 1363):

> (5) the term "artificial intelligence" means any method implemented on a computer, including any method that is drawn from machine learning, data science, or statistics, to enable the computer to carry out a task or behavior that would require intelligence if performed by a human.

This definition reflects other industry definitions including ones utilized by Booz Allen Hamilton, SAP, and Deloitte which can be found in the footnotes below.[1]

There are other existing definitions of AI currently in use across the federal government beyond the scope of our comments. For instance, the John S. McCain National Defense Authorization Act for Fiscal Year 2019 (P.L. 115-232) defines AI as:

> (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
> (2) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
> (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
> (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.

[1] *The Artificial Intelligence Primer*, Booz Allen Hamilton, https://www.boozallen.com/s/insight/thought-leadership/the-artificial-intelligence-primer.html;, "What is Artificial Intelligence," SAP, https://news.sap.com/2018/03/what-is-artificial-intelligence/; *Artificial Intelligence*, Deloitte, https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/deloitte-analytics/deloitte-nl-data-analytics-artificial-intelligence-whitepaper-eng.pdf

> (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting. (Title II, Section 238 - JOINT ARTIFICIAL INTELLIGENCE RESEARCH, DEVELOPMENT, AND TRANSITION ACTIVITIES)

The Data Coalition recommends that the Administration clearly articulate the definition of AI being relied upon for official policy matters moving forward.

We define “data”, “data set”, and “model” in the following ways:

- “Data” - The Data Coalition defines data as codified in the OPEN Government Data Act, Title II (P.L. 115-435), Section 202:

  > (16) the term ‘data’ means recorded information, regardless of form or the media on which the data is recorded.

- “Data set” - The Data Coalition defines data set as a collection of data that can be manipulated as a unit.

- “Model” - The Data Coalition defines model as the articulation of a mathematical relationship between a group of data inputs and a resulting output.

Although this definition of model may be technical and axiomatic, the main idea behind it is simply an attempt to quantitatively connect data to real-world settings.

DISTINCTION BETWEEN “MODEL” AND “MODEL SPECIFICATION”

The definition of a model is distinct from the definition of “model specification,” which refers to the type of model and hyperparameters given to it (if any), without any reference to data fed into the model.

For example, in K-Nearest Neighbors (KNN), which is a non-parametric machine learning method to model new data based on the most similar existing observations in the data, the number of observations, *k*, used to classify a given data point is a model specification. In some contexts, the Office of Management and Budget (OMB) may be more interested in which *types* of models (model specifications) researchers intend to use, rather than distinct *instances* of models, which require reference to data fed into them. This is a core concept of model tuning, where the hyperparameters used to create a model and the underlying data given to a model are both required to recreate/replicate that model. A model that uses different hyperparameters given the same data will yield different results.

Similarly, a model that is created based on a specific underlying data set with a certain model specification will be different than a model created using different data but the same model specification, even if that data only differs by one data point.

IMPORTANT MODELING CHARACTERISTICS

Building models with explainability, auditability, robustness, and governance ensures that organizations can maximize the utility of their models while reducing risk and exposure to issues with implementation (e.g., lack of transparency into how the model reached a conclusion, biased decision-making or recommendations, inaccurate results due to poisoned data or model tampering from adversarial attacks, etc.). Additionally, from a technical introspection lens, model versioning should be considered within the context of building and deploying machine learning models.

Explainability and Auditability

AI systems must be as transparent and explainable as possible. This should be tackled from the bottom-up so that auditability is a feature built into AI systems at the model level. This elevates and informs the process for creating explainable and trustworthy AI. Realistically the exact level of explainability can and should be tailored to the intended use case or scenario. The opacity of AI systems is one of the greatest barriers to trustworthy AI, and despite heavy research investment in AI over the last decade, models still remain black box in many cases.

The first step to ensuring that models are built into systems is to ensure that AI models are designed to be auditable. This means it is possible to have insight into the training data, model information, and other system information which inform the end decision, recommendation, or output. Auditability is key to understanding how and why AI systems arrive at conclusions or recommendations and can then inform the appropriate explanation for the recipient of the information. Building auditability into systems as a part of the fundamental design ensures that there is transparency into how systems make decisions. Auditable AI systems should contain key metrics and information around system build and performance, which can be sourced by engaging with the appropriate agency and industry groups to identify the most salient features for agency/industry verticals.

Like version history in software releases, AI audit trails should at a minimum contain information such as the model version (e.g., source of the original model, the technique used to train it, performance metrics around accuracy, when it was last tuned) and the data's provenance (e.g., source of the training data). At an even more granular level, systems should be built so that this information can be accessed and tracked in real-time, so that it is possible to parse this information at the time of inference. Although many of these standards with respect to a model's versioning provenance and audit trails are likely already part of organizations' internal software development practices, and not necessarily shared publicly, they should be required model characteristics to improve AI R&D and testing. This is especially important when applied to the use of public sector information or decision-making systems. Auditability is very similar to the concept of reproducibility, which ensures that models are built in a way to allow them to be independently reproduced.

Robustness

Due to the lack of design standards and architecture governance, AI systems are susceptible to adversarial attacks that can seriously threaten the integrity of the data and models. If AI systems lack robust security controls to prevent adversarial attacks, they cannot be trustworthy despite features enabling auditability or explainability. Adversarial threats can occur during the training phase of building an AI model, as well as the operational data ingestion phase, both of which impact the performance and output of the solution.

In the training phase, data can be compromised (i.e., 'data poisoning') by introducing intentionally engineered bad data into the training data set. This will impact the model's development and ultimately, performance, if it is mistakenly deployed. In the ingestion phase, white box or black box attacks (i.e., 'model tampering') occur when attackers manipulate ingested data to fool a trained model or use the model output to reverse-engineer and produce adversarial input data that impacts model performance.

Although the risks are relatively low today due to challenges with deploying and operationalizing AI to the enterprise, the challenges with respect to preventing and safeguarding these systems against attacks should not be taken lightly. It is important to build the foundation now to proactively address adversarial attacks by understanding best practices to safeguard against data poisoning and model tampering. Many of these approaches are still nascent as research is still ongoing to create methods to reliably identify tampering and establish proactive safeguards and measures.

Governance

The current process for developing and deploying AI models is largely piecemeal, with bespoke models developed and deployed for specific data or problem sets. Rather than considering how to best develop and deploy AI systems to be operationalized at enterprise scale, groups operate independently of one another, creating AI models and tools as applied to their individual problems and applications. This approach, defined by a lack of coordination and oversight, exposes organizations to serious risk because there is no way to establish common governance controls and procedures.

Common standards around governance must be developed to ensure that the appropriate level of oversight is employed commensurate to the deployment, thereby reducing risk.

Model Versioning

Machine learning versioning approaches are emerging to capture the breadth of concerns related to AI model development contexts and applications. There are two main areas for expansion of scope for version control. The first area focuses on traditional techniques to the versioning of machine learning models and data, in which iterations of the code are saved as named instances, similar to multiple drafts of writing an essay (e.g., MemoVersion1, MemoVersion2, etc.) The second area focuses on the versioning and reproducibility of exploratory data analysis. This focuses more on the meta aspects of the creation process, tracking not just changes in code, but also evolving motivations for those changes. This is analogous to writing a paragraph about the reasons for updates in a specific version of the essay.

In addition to versioning, the machine learning field is starting to converge on general paradigms for model representation, and agencies have an opportunity to drive consistencies into approaches by development teams.[2]

IMPORTANCE OF AI IN COMBINING DATA SETS

When using models or analysis, it is important to note that one of the greatest value-adds from AI comes from cross-analysis of multiple data sets often from disparate domains in order to readily identify value, trends, and patterns.

For example, combining education records with data about student participation in after school activities could help build a case for the effectiveness of after school programs. Without combining these two data sets from disparate domains, this ability to determine if the school programs had any impact would not be possible. There are often issues, however, in combining data sets; we discuss these below. Addressing these issues will be important for improvements to AI.

2 – QUALITY AND ACCESS IMPROVEMENTS FOR AI

Regarding quality improvements to accessible data and models to improve AI R&D and testing, the RFI asks:

- As agencies review their data and models, what are the most important characteristics they should consider?
- What characteristics should the Federal Government consider to increase a data set or model's utility for AI R&D (e.g., documentation, provenance, metadata)?

CHARACTERISTICS TO FACILITATE CROSS-ANALYSIS USING AI

The OPEN Government Data Act (P.L. 115-435) provides new authorities and statutory requirements for making federal data sets available in an ideal form. The definitions found in Section 202(a)(18-21)) state:

> (18) the term ‘machine-readable’, when used with respect to data, means data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost;
>
> (19) the term ‘metadata’ means structural or descriptive information about data such as content, format, source, rights, accuracy, provenance, frequency, periodicity, granularity, publisher or responsible party, contact information, method of collection, and other descriptions;

[2] See Booz Allen Hamilton’s response to a previous RFI: https://www.nist.gov/sites/default/files/documents/2019/07/17/nist-ai-rfi-boozallen-001.pdf.

> (20) the term ‘open Government data asset’ means a public data asset that is—
>
> (A) machine-readable;
>
> (B) available (or could be made available) in an open format;
>
> (C) not encumbered by restrictions, other than intellectual property rights, including under titles 17 and 35, that would impede the use or reuse of such asset; and
>
> (D) based on an underlying open standard that is maintained by a standards organization;
>
> (21) the term ‘open license’ means a legal guarantee that a data asset is made available—
>
> (A) at no cost to the public; and
>
> (B) with no restrictions on copying, publishing, distributing, transmitting, citing, or adapting such asset;

The OPEN Government Data Act (P.L. 115-435) also contains provisions to ensure a common repository (i.e., “Federal Data Catalogue” understood to be Data.gov). The newly-codified Chief Data Officers (CDO) Council provides the ideal governance structure for government agencies to ensure consistent standards for data integrity and composition.

The Data Coalition encourages that agency-specific data governance policies be reviewed and updated to better link data and data sets together. Additionally, the Administration, and the CDO Council specifically, can help strengthen certain characteristics of data to facilitate analysis, including storing the data sets in an accessible location in a machine-readable format with standardized metadata as discussed below. It should be noted that agency data governance leadership would benefit from formal mechanisms to understand the developmental intent and R&D hypothesis motivating the AI applications. This will enable continuous improvement in the provisioning of the federal data and models. We encourage OMB to also explore how AI application feedback mechanisms can be appropriately built into the release processes.

The Data Coalition acknowledges that many government data sets are unavailable due to privacy or national security concerns, as well as legal restrictions and to fulfill pledges of confidentiality. We urge agencies to closely examine their processes for determining whether a data set can be safely and legally released, including in de-identified forms when data sets can be made available in a manner that facilitates some analysis without compromising sensitive or confidential data. This aligns with the OPEN Government Data Act’s expectation that government-wide are open by default. The Executive Branch must further consider how tiered access applications and other approaches can help foster this open data mandate in the most responsible and reasonable manner possible.

MACHINE-READABILITY AND METADATA STANDARDS

Machine-readability means that the end user does not have to perform initial format conversions to begin using the data for machine-based analysis. For example, many PDFs are not easily read by a machine, so an analyst would have to manually copy and paste the relevant information—a task that is exceedingly costly and inefficient. Even the most advanced modern optical character recognition cannot fulfill this task without introducing errors and inefficiencies to large scale data sets. This then causes analysts to spend more time determining where the errors have occurred.

When combining data sets and creating new data sets, the resulting data sets should be created and stored in a machine-readable format. In particular, data should not lose semantic meaning when processed by a machine. Beyond machine readability there are many forms of recorded information (e.g., video, audio, text, image, etc.) which will be valuable for AI development and R&D which should be made available in their natural formats.

However, as OMB's RFI assumes, data should also be stored in an easily accessible and well-documented location with the appropriate metadata. This is important when linking data sets together. As the amount of data held by the Federal Government increases, the majority of the time creating models and doing analysis using AI should not be focused on locating the data sets. It is important that the metadata documentation itself be recorded in machine readable and open formats with mechanisms to allow for bulk download. This will facilitate the discovery of the most appropriate data sets.

Similarly, data sets should be well documented. For example, a data set for a specific year should be labeled as such while a continuously rolling data set should also be labeled as such.

Standardized metadata facilitates cross-agency analysis by reducing the amount of time an analyst who creates the AI models must invest to become familiar with the data. Rather than learn the differences between how two agencies store the same type of information and manually create code to harmonize these data sets into a standard format, an analyst using data with standardized metadata would only need to learn one metadata taxonomy.

For example, appropriate and complete keys are an important aspect of metadata for combining data sets. Standardized and complete keys will immensely reduce the time required to link data sets. Not only a time-saver, this also decreases errors in analysis that result from misinterpretation and human error when combining the data sets. Implementing standardized metadata will be easier for some agencies than for others due to resource and legal constraints. The Data Coalition recognizes that agencies must prioritize how to develop metadata for critical mission data, and that phased-in approaches may be necessary for releasing specific agency or mission data sets.

Although not explicitly mentioned in the definition of metadata from the OPEN Government Data Act (P.L. 115-435), standardized metadata for AI use must consider missing data. Reasons why data are missing should be explained in the metadata to the extent possible.[3] Because AI relies so heavily on existing data being representative of the population, missing data can skew results. A variety of tools and techniques exist to *impute*, or fill in missing data values, but AI users must have some knowledge of the extent to which missing data may affect imputations or introduce bias in conclusions from AI analyses.

Even with standardized and machine-readable metadata, the data sets themselves may still have errors. Responsible parties should ensure that data sets have been appropriately transformed and cleaned. In addition to standards for using the data, future consideration should be given to standards for managing, analyzing, and disseminating AI products.

Setting standards is one of the key facilitators for using data to inform decision-making. The DATA Act (P.L. 113-101), for instance, resulted in efforts to bring together federal financial data by applying consistent data standards. By leveraging existing frameworks and initiatives, the Administration can establish benchmarks and parameters to help meet management reform goals and tie disparate data sets together into a government-wide framework of valuable operational data (i.e., mission agnostic support data representing resources, decisions, transactions, and outputs of administrative functions). The resulting overarching framework will allow AI systems to derive insights about the structure and financial function of the federal government itself.

As the Administration makes plans to improve management practices we hope leadership will consider leveraging existing standards efforts, such as the Financial Accounting Standard Board's (FASB) US GAAP Taxonomy for Security and Exchange Commission (SEC) financial regulatory reporting, the DATA Act Information Model Schema (DAIMS) for government-wide agency spending transparency, and the National Information Exchange Model (NIEM) for homeland security, justice, and public health Federal-State-Local programmatic data exchange standards (e.g., mission-related data representing persons, places, and things which are created by or collected for specific program or regulatory functions). These standardization efforts represent the high value work of using government data to drive oversight, reform, and accountability through accurate, consistent, and controlled, high quality data.

3 - DATA SET ACCESS

The open data assets indexed in the Federal Data Catalogue (Data.gov) represent immediate candidates for identifying related, but 'private' or otherwise sensitive, data sets which could be released in a controlled manner for limited access. These more granular and transactional data sets are more valuable for AI algorithm testing and R&D applications. Many of the data sets that are available today are aggregations in lower granularity and suffer from unclear or absent metadata and business definitions for them to be utilized consistently and accurately.

In addition to more granular, structured data assets, agencies should also consider raw, unstructured information and data sets for public or private/restricted access. For example,

[3] Datasheets for Data sets, currently in development by researchers, could be helpful: https://arxiv.org/abs/1803.09010

agency call center logs, consumer inquiries and complaints, regulatory inspection and investigative reports represent high value resource. With the proper redaction and safeguarding of entity and individual PII attributes, such data can be very valuable for AI applications.

As with the operational data standardization work discussed above (e.g., DATA Act), the Data Coalition also encourages OMB to consider the primary role and value of the information contained in various federal text-based documents such as legal (e.g., US Code, Statutes), regulatory (e.g., FAR, Regulations.gov, notice and comment process), management guidance (e.g., OMB memorandum, Circulars, Executive Orders), and various internal budgetary, financial, and performance reporting requirements. These documents and reports should be treated as valuable data assets presenting opportunities for AI applications to develop insights into the federal government's structure, performance, legal compliance, and financial management.

Due to the reliability of the reports and documents (i.e., regularly recurrent, centrally controlled, government-wide in nature) they also serve as a critical link to drive exploration into the more unstructured, transactional, or programmatic data sets (e.g., mission-related data representing persons, places, and things which are created by or collected for specific program or regulatory functions). We encourage OMB to apply modern "data-first" technologies for authoring documents and begin the process of defining a data format and strategy specifically related to the drafting and interoperable dissemination/issuance of management and reporting documentation. That is, the Administration should seek to publish this documentation in integrated, machine-readable data formats instead of documents.