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

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government nor those of the National AI Research Resource Task Force., and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality, or content of all external links included in this document.

CrowdAI

RFI Response: National AI Research Resource (September 2021)

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.]**

D. Capabilities required to create and maintain a shared computing infrastructure to facilitate access to advanced computing resources for researchers across the country, including provision of curated data sets, compute resources, educational tools and services, a user-interface portal, secure access control, resident expertise, and scalability of such infrastructure;

We recommend incorporating end-to-end (i.e., from data ingestion and labeling to model production), open, modular, code-free AI enablement tools in order to facilitate access to the proposed NAIRR beyond just AI/ML engineers (of which there is short supply).

Such tools provide numerous benefits at the enterprise-level, including but not limited to:

- Scale the adoption of AI by enabling a wider workforce to engage in AI research and AI-enabled research without needing to turn vasts numbers of people into AI/ML engineers or data scientists
- Help standardize data ontologies and curate data sets for wider, more efficient use
- Provide a single user-interface
- Streamline secure access controls

E. An assessment of, and recommended solutions to, barriers to the dissemination and use of high-quality government data sets as part of the National Artificial Intelligence Research Resource;

AI models require case-specific training data, and cannot be tuned without representative data that is appropriately labeled. However, some USG programs expect to evaluate and purchase models without providing AI vendors consistent, relevant, and diverse training data.

AI model accuracy directly results from robust, well-curated training data. Without consistent, relevant, and diverse training datasets, AI vendors are more likely to over-fit or under-fit to the data, decreasing the quality of models.

Part of the problem has to do with the USG operating under a misperception that if it provides all available training data to AI vendors, they will be "giving too much away," and the USG will lose the ability to effectively test and evaluate delivered models. While withholding some data to support testing and evaluation generally is considered a best practice for AI development, we often have experienced the USG not providing adequate training data to build truly scalable, domain-specific models.

Another part of the problem stems from the security classifications of data.

We recommended the following solutions to these barriers to the dissemination and use of high-quality government data sets:

- NAIRR should commit to providing AI researchers without security clearances access to unclassified training data.
- NAIRR should examine the feasibility of holding clearances for cleared researchers who are affiliated with uncleared organizations and would like to work with classified data.
- NAIRR should coordinate a USG effort to put forth a standardized ontology for creating training data.

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

We recommend incorporating end-to-end (i.e., from data ingestion and labeling to model production), open, modular, code-free AI enablement tools in order to facilitate access to the proposed NAIRR beyond just AI/ML engineers (of which there is short supply).

Such tools provide numerous benefits at the enterprise-level, including but not limited to:

- Scale the adoption of AI by enabling a wider workforce to engage in AI research and AI-enabled research without needing to turn vasts numbers of people into AI/ML engineers or data scientists
- Help standardize data ontologies and curate data sets for wider, more efficient use
- Provide a single user-interface
- Streamline secure access controls

2. 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?

We realized early on the importance of building AI that responsibly captures our users' workflow throughout the entire AI development process. Keeping the researchers' workflow in mind is the best way to build AI that provides the most value, and reinforces principles of ethical and responsible AI R&D.

Tools that engage users throughout the process from the beginning to the end, and back again (that is, for model sustainment), serve as guardrails for making sure people are building AI models in a manner that is responsible, equitable, traceable, reliable, and governable.

Capturing researchers' workflow can best be done through end-to-end (i.e., from data ingestion and labeling to model production), modular, code-free AI enablement tools.

There are multiple pieces to the AI process that need to be taken into account when building a responsible AI ecosystem. These include considerations around -- and accompanying enablement tools for --

- Compute
- User Management
- Data (ingestion, management, labeling)
- Model Training
- Model Testing & Evaluation
- Model Deployment & Sustainment

Below are considerations around each of these pieces of the AI process:

- Compute - AI, especially computer vision (or CV), requires significant compute resources.
    - Deciding where to allocate compute resources should be considered from a mission perspective. For example, within the GEOINT world, identifying Indications & Warnings, or doing mapping wildfire extent, requires real-time analysis. To reduce latency of model outputs, the reliable option here from a mission perspective would be to allocate compute resources as close to the sensor(s) as possible. For other missions, real-time analysis might not be as pressing, so compute resources can be allocated closer to end users.
    - That said, cost also factors here when considering scalability.

- User Management - While software is integral to AI, at the end of the day, it's people that are interacting with data.
    - It's important to have a user management tool to grant and restrict access to 1) see and/or edit data, and 2) train, test, and deploy models, in a manner that accords with data protection.

- Data - Data forms the backbone of AI. It's what it's built on, and it's what makes an AI model performant, useful, and responsible, or not.

  - Researchers need to know why they need plentiful, diverse, and appropriately labeled data in order to build robust models, which requires easy-to-use (i.e., code-free) tools to ingest data and label it.
  - Data standards for ontology and metadata would go a long way when it comes to an AI ecosystem's governability and reliability. NAIRR and its AI stakeholders should work together to set standards for ontology and metadata, would result in the kind of consistency and quality we would like to see and would benefit everyone. We've seen this being useful for FEMA disaster response categories.
  - Data management tools can help mitigate bias as well. E.g., by grouping data with similar data with respect to any given characteristic (e.g., data from a certain type of sensor), and then flagging for the user than if all the training data for a model comes from only one grouping (e.g., all of the data comes from just one sensor) it can't be expected to generalize, or perform on data beyond that type (in this case data from other sensors).

- <u>Model Training</u> - Model training needs to be transparent and traceable when it comes to how a model was trained, and on what data. to have the ability to log model training, especially if AI is to earn the long-term trust of users.

- <u>Model Testing & Evaluation</u> - Here it's important to have visual tools and education around test & eval issues to ensure models are reliable.
  - E.g., researchers need to understand that test set data must be separate from training data.
  - Researchers need to understand how their particular use-case informs the precision and recall they might be looking for (e.g., in the case of countering illicit trafficking there may be a tolerance for lots of false positives, but absolutely no false negatives)

- Model Deployment & Sustainment - In order for models to be reliable for mission use, there needs to be an easy way for users to deploy them and continuously sustain them through retraining with new data and new contexts.
    - Here, a tool to automate deployment through containerization is key.
    - For sustainment, a feedback loop wherein users rate model output, thereby creating more training data for the model to improve with.

In addition, intuitive training sessions, workshops and/or documentation on data literacy are extremely important. Researchers need to know why they need plentiful, diverse, and appropriately-labeled data in order to mitigate bias.

We recommend conducting workshops for users, in addition to the brief tutorials that may exist in the code-free enablement tools for data ingestion, data management, and data labeling.

A data literacy workshop or tutorial would typically highlight why data needs to not only be plentiful, but diverse as well in order to create a model that is better generalized to the task at hand, and not overfit on a small homogeneous set of data where the model might pick up on and be biased towards the wrong characteristics.

A simple example we might use to highlight this: a model to detect individuals who are armed with weapons is only trained on dark-skinned individuals with weapons and light-skinned individuals from different video feeds without weapons, the model might erroneously assume that the defining characteristic of having a weapon is having dark skin.

Another simple example: show two sets of pictures, 6 cats and 6 dogs. In the pictures, it is clear to humans that there are 6 cats and 6 dogs. The six cats are on sofas and the six dogs are on the grass, the animals take up less than 10% of the image. Then show a dog on a couch and a cat on grass. The model may get confused, because models take the easiest approach.

The data literacy workshop or tutorial would also cover how to label data appropriately for the task at hand.

Take for example an object detection bounding box model. It's important to show what's an appropriately sized bounding box to use in any given piece of data to turn that into good training data (the bounding box needs to capture all of the relevant pixels and nothing else).

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

Given 1) the relative scarcity of AI and machine learning expertise in the workforce, 2) the need for solutions that recognize the iterative nature of effective AI, and 3) the need to create and manage quality training data for AI models, we're seeing the emergence of no-code enablement tools and platforms. Lots of companies are offering components of the AI pipeline, like data labeling or data management tools.

At CrowdAI, we focus on offering every component of the end-to-end AI pipeline, specifically as it pertains to computer vision (user management, data labeling/management, model training, model testing and evaluation, and deployment)

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

Companies with end-to-end, modular, code-free AI enablement tools should offer use of their platforms at discount or no-cost, similar to academic partnerships they might already have.

5. 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?

There is a short supply of AI skills within the USG, and private sector. Obtaining the required skills to develop AI solutions is onerous. The skills gap often hinders the USG's ability to purchase, deploy, and/or operate AI solutions.

The only way to efficiently democratize access to AI R&D is to arm the workforce with code-free AI enablement tools through which they can build, deploy, and sustain their own AI models, thereby removing reliance on the short supply of AI engineers.