

Department of National Institute of Standards and Technology (NIST)

Request for Information (RFI) Response

RFI Docket No: 19031229-9229-01

May 31, 2019

Submitted to:

AI-Standards
National Institute of Standards and Technology,
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

e-mail: ai.standards@nist.gov

Submitted by:

SAIC
12110 Sunset Hills Road
Reston, VA 20190-5916



This Request for Information response includes data that shall not be disclosed outside the government and shall not be duplicated, used, or disclosed, in whole or in part, for any purpose other than to evaluate this Request for Information. If, however, a contract is ultimately awarded to this responder [or submitter] as a result of or in connection with the submission of these data, the government shall have the right to duplicate, use, or disclose the data to the extent provided in the resulting contract. This restriction does not limit the government's right to use information contained in this data if it is obtained from another source without restriction. The data subject to these restrictions are contained in all sheets. In addition, the information contained herein may include technical data, the export of which is restricted by the U.S. Arms Export Control Act (AECA)(Title 22, U.S.C. Sec 2751, et seq.) or the Export Administration Act of 1979, as amended (Title 50, U.S.C., App. 3502, et. Seq.).

This material is not intended by Science Applications International Corporation to become a "record", within the meaning of 5 USC Sec. 552, and is entrusted to the government with the understanding that it will be returned if the government is unwilling or unable to maintain it as non-record material.

May 31, 2019

National Institute of Standards and Technology (NIST) Artificial Intelligence Standards

Attention: ai-standards@nist.gov

Subject: SAIC RFI

Reference(s): (a) Request for Information (RFI) Docket Number: 190312229-9229-01 dated 05/25/2019

Science Applications International Corporation (SAIC) is pleased to submit the subject Request for Information (RFI) in response to the Reference (a).

In accordance with the instructions set forth in the Reference (a) RFI, SAIC has provided the below enclosure for your review and consideration.

Enclosure (1) SAIC Responses to RFI Docket Number: 190312229-9229-01

The vendor name, number, address, and contact information for the purpose of this RFI is:

Science Applications International Corporation

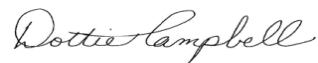
12010 Sunset Hills Road

Reston, VA 20190

Contact: Jerry Tipton, Sr. Program Manager, (301) 377-5658

SAIC appreciates the opportunity to submit the request for information. Please let us know if you have any additional questions or comments.

Very Respectfully,
SCIENCE APPLICATIONS INTERNATIONAL CORPORATION



Dottie Campbell
Contracts Analyst

About SAIC

SAIC is a premier technology integrator solving our nation's most complex modernization and readiness challenges across the defense, space, federal civilian, and intelligence markets. Our robust portfolio of offerings includes high-end solutions in systems engineering and integration; enterprise IT, including cloud services; cyber; software; advanced analytics and simulation; and training.



23,000+
EMPLOYEES



69%
CLEARED
HOLD A SECURITY
CLEARANCE



6,000+
ARE VETERANS



3,000+
CONTRACT
VEHICLES

With an intimate understanding of our customers' challenges and deep expertise in existing and emerging technologies, we integrate the best components from our own portfolio and our partner ecosystem to rapidly deliver innovative, effective, and efficient solutions.

SAIC is headquartered in Reston, VA and has a global presence with over 100 locations worldwide. SAIC is a user of artificial intelligence (AI) and machine learning (ML) technology and regularly partners with major AI technology developers, using advanced AI/ML tools to solve customer challenges. In addition to using and applying these tools, SAIC also performs evaluation and standards development for AI/ML technologies at SAIC's Identity and Data Sciences Laboratory.

About the Identity and Data Sciences Laboratory (IDSL)

SAIC's Identity and Data Sciences Laboratory (IDSL) is comprised of scientists, engineers, IT specialists, and program managers with strong expertise in testing and evaluating AI/ML systems for biometric identification. Members of the IDSL conduct applied research in operational use of biometric identity systems and regularly present results at industry conferences and by publishing peer reviewed scientific research. The IDSL provides classified and unclassified applied research and subject matter expertise in biometrics to several government agencies.

The IDSL operates the Maryland Test Facility (MdTF) for the Department of Homeland Security Science and Technology Directorate (DHS S&T). The IDSL provides technical services including the following: systems engineering, rapid prototyping, laboratory and scenario testing, human subject recruitment, Institutional Review Board protocol development, design and support of field trials, development and demonstration of functional models, biometric system assessments, human factors assessments, technical performance assessments, and identification of emerging technologies.

AI Technical Standards and Related Tools Development:

AI Standards for Biometric Systems

The IDSL's input into standards is based on technology evaluations performed at the Maryland Test Facility (MdTF), including the 2018 and 2019 Biometric Technology Rallies. The IDSL has performed technology evaluations at the MdTF since 2014, testing dozens of commercial biometric technologies with diverse users and various use-cases. This experience gives us a strong applied understanding of the difference in technology performance in engineer-oriented benchmarking versus actual realized performance with an untrained user. We currently leverage this experience to inform existing standard development activities. Below we describe our current activities and a roadmap for future work.

Current Standards Work

The IDSL is working to develop and inform industry AI standards targeted toward the biometric technology sector and is actively participating in the development of ISO 19794 – Biometric performance testing and reporting, ISO 21472 – Scenario evaluation methodology for user interaction influence in biometric system performance, and ISO 22116 – Identifying and mitigating the differential impact of demographic factors in biometric systems.

Prior research from our group developed a method for evaluating autonomous biometric system performance in a high throughput environment, such as in a travel use-case. This method included modifications to standard metrics in order incorporate operational time constraints [4]. We successfully applied this method to testing an array of commercial face recognition systems as part of the 2018 Biometric Technology Rally showing their utility for identifying suitability for deployment under different time constraints [3]. This work from our group can be broadly translated to evaluation of other AI systems used for automation in time-constrained use-cases, such as a store checkout or in a medical office. We are now working to include these metrics into ISO 19794 and ISO 21472.

Evaluation Methods/Tools for Biometric Systems

The IDSL understands that AI technologies must be evaluated within specific contexts of use because the output of these systems depends critically on the data on which they operate. In the context of biometrics, the IDSL has shown that face recognition algorithm performance depends critically on when and how the face photo is acquired. For this reason, the IDSL tests biometric technologies embedded in full scenarios which allows us to gather information not only on the performance of specific algorithms, but also on the scenario conditions that alter algorithm performance that cannot be ascertained from the input data alone [2, 3].

The IDSL maintains a comprehensive dataset of finger, face, and iris images and system use video clips for ~2,000 diverse individuals gathered responsibly under Institutional Review Board (IRB) oversight, with appropriate human subject protections including informed consent by all participants. Select portions of this data have been previously shared with NIST, as approved under the IRB, and have informed several evaluations including [5] and [6]. Each individual within the IDSL's dataset has numerous images gathered

since 2014. All images within IDSL's dataset include strong ground-truth information including comprehensive demographics and psychographics of the subjects and the scenario conditions of acquisition.

Roadmap for Future Work

In the future, the IDSL plans to continue developing relevant biometric technology metrics and setting performance goals as part of its support for DHS S&T's Biometric Technology Rallies. The IDSL also plans to contribute to standards on appropriate methods for measuring demographic effects in biometric systems, including techniques for identifying the presence and weighing the importance of a demographic variable on system performance.

Recommendations for the Development of AI Technical Standards

1. Model the Development of new AI Standards on the Existing Biometric System Standards Framework

Responding to:

- Topic Area 5. Any supporting roadmaps or similar documents about plans for developing AI technical standards and tools.
- Topic Area 10. Where the U.S. is currently effective or leads in AI technical standards development.

Biometric systems are a specific type of information technology system used to establish the identity of individuals using their distinct physiological or behavioral features through the application of ML and AI principles. Biometric AI systems are increasingly publically facing with a strong operational need to ensure public trust and acceptance of these systems. Biometrics are a longstanding use case for applied AI and standards regarding the use, testing, and vocabulary surrounded biometric systems have been in development since the early 2000s. For these reasons, we believe the existing biometric standards are a good starting point for the development of new AI technical standards.

AI technical standardization efforts should start with the development of a comprehensive standard describing the vocabulary and terminology associated with AI systems and their testing. This would be similar to the existing ISO 2382-37 document that outlines over 100 standard biometric terms in nine functional categories. Many of these categories could be directly transferrable to a new AI technical standard, such as general, system-level, data-centric, functional, personnel, application, and performance terms.

AI standards for testing should be modeled after the ISO 19795 approach to biometric system testing which recognizes the need for scenario evaluation in specific use cases (ISO 19795-2). The performance of biometric systems has the potential to impact people's lives. This impact can range from simple inconvenience, such as being locked out of a phone, to the serious, such as being accused of a crime. Recognizing both the diversity of applications and the variety of impacts that can occur from AI errors, the biometric

testing community has long maintained an array of standards designed around specific biometric system use cases. Biometric test standards, therefore, are sectioned into different levels and categories of testing, including:

- ISO 19795-1 - Basic testing principals,
- ISO 19795-2 - Testing methodologies for scenario evaluations
- ISO 19795-3 - Modality specific biometric testing
- ISO 19795-4 - Interoperability performance testing
- ISO 19795-5 – Access control scenario testing
- ISO 19795-6 – Operational system testing

Similarly, AI systems are used in a variety of different environments and scenarios for which the costs of AI errors can vary in severity and prevalence based on the particulars of each application. For example, an AI system for detecting early signs of diabetic blindness made few errors when using sequestered datasets, but had operational performance challenges when using images gathered in the field¹. Underestimating the number and types of errors made by AI can lead to inappropriate planning for technology deployments. We therefore encourage developing AI testing standards that include strong provisions for testing in applied scenarios using a standard vocabulary for quantitative metrics.

2. Include Frameworks and Nomenclatures that support the Evaluation of AI Equitability

Responding to:

- Topic Area 8. Technical standards and guidance that are needed to establish and advance trustworthy aspects of AI technology
- Topic Area 11. Specific opportunities for, and challenges to, U.S. effectiveness and leadership in standardization related to AI technologies.

Because of its demographic diversity and history, the US currently leads on tackling the issues of performance differences across demographics in biometric and other AI systems. By addressing and acknowledging the need to assess the performance of AI systems for diverse users, the US will foster development of better AI systems that are more robust to these factors versus fragile systems that have large performance gaps that may go unnoticed and unaddressed without such analysis.

Separating the quantitative technical aspects of technology performance from qualitative social factors is critically important when discussing these topics. For example, recently, the term “bias” has become the catchall for discussing demographic effects in biometric systems like facial recognition despite the fact that the term has no specific technical definition. Frequently, saying a system is biased is conflated with saying that is “does not work” for specific demographic groups. This can lead to decisions by policy makers that may

¹ <https://www.wsj.com/articles/googles-effort-to-prevent-blindness-hits-roadblock-11548504004>

lack a firm statistical foundation. We believe that strong performance testing standards will help make appropriate data available to understand how well the technology works for different demographic groups using standard metrics and benchmarks that depend on the outcomes of technology use.

Consequently, the IDSL has developed a framework for assessing the “equitability”² of a biometric system in the context of specific biometric task being performed, which we believe is a better alternative to the term bias. We introduced the following set of terms for describing different demographic effects for a biometric system and believe they are readily applicable to AI systems as a whole:

- **Differential Performance.** We define differential performance as a difference in the genuine or imposter distributions for specific demographic groups independent of any decision threshold. This is closely related to the concept of “biometric menagerie”, a phenomena in which subject-specific genuine and imposter distributions are statistically different. Differential performance is this same effect, not for specific subjects, but for specific demographic groups.
- **Differential Outcome.** We define differential outcome as a difference in FM or FNM rates for different demographic groups relative to a decision threshold. Similarity scores in and of themselves are not the outcome of an identity decision. They must be re-cast to match/no-match decisions using a decision threshold. These match decisions can then be used to calculate FM and FNM error rates.
- **False Negative Differential.** We use the term False Negative Differential as a tendency, for a specific demographic group, to experience a false negative error. That is, a failure of the group member to be identified as themselves.
- **False Positive Differential.** We use the term False Positive Differential as a tendency, for a specific demographic group, to experience a false positive error. That is, the tendency to mistake the group member for somebody else.

Considering differential performance and outcome separately helps acknowledge that just because a system shows demographic variation in some internal variable, such as training sample composition, network weights, unit activation patterns, or similarity scores, it does not necessarily manifest in outcomes for users of the system. The IDSL believes that teasing apart internal performance measures from actual operational outcomes helps make sober decisions about whether the technology is suitable for specific use-cases.

Furthermore, separating the concepts of False Negative and False Positive Differentials to better estimate the cost of differential outcomes to affected individuals is vitally important. In AI systems, false positive and negative can carry drastically different costs, which must be considered separately when tuning a system toward optimal equitability in each use-case. Additionally, the presence of one kind of error for a specific group does not independently suggest the other kind of error exists. This concept is frequently overlooked.

² i.e. the extent to which performance is constant for different user cohorts

3. Include Frameworks for Full System Testing, Including Signal Acquisition Mechanisms and Human Interaction Influence

Responding to:

- Topic Area 8. Technical standards and guidance that are needed to establish and advance trustworthy aspects of AI technology
- Topic Area 11. Specific opportunities for, and challenges to, U.S. effectiveness and leadership in standardization related to AI technologies.
- Topic Area 12. How the U.S. can achieve and maintain effective leadership in AI technical standards development.

The IDSL has been performing tests of full biometric technologies for DHS S&T since 2014. We strongly believe that similar in-system scenario testing is critical for accurately characterizing commercial AI technology performance for the following reasons:

- AI technology developers have blind spots in their understanding of how the technology is used operationally and are motivated to cast reasonable use-case facts on the ground as erroneous use outside the scope of testing that should go into system metrics.
- AI companies have many algorithms and are tweaking and changing the technology daily. Even when performance of a specific company's AI algorithm is available, the specific AI algorithm included in a particular commercial system may not be clearly understood by the customer or even by the vendor. This makes in-system testing the only way to accurately know the performance of a specific AI product.
- AI technologies perform differently based on the use-case and therefore must be tested in a way that incorporates the nuances of the use case. Similar to pharmaceuticals, AI systems are complex, with different risks and costs in different use-cases.

It is vital that these technology tests are representative of real world applications. For example, AI algorithms are often developed on training sets of data, collected in laboratory conditions with a single or small set of sensors. When deployed, the variety or exact type of sensor used to collect the same sensory input can change, causing errors that are impossible to predict from laboratory evaluations alone. Furthermore, human factors can diminish the performance of the most adept AI technologies. An operator can easily misinterpret a system response or mis-calibrate a system setting. These could have drastic consequence in AI deployed for critical infrastructure or healthcare purposes and are often not considered during AI development. Standards and frameworks need to be developed, similar to ISO 19795-2 and ISO 21472 that addresses these considerations.

4. Require Robust, Independent, Third-Party Certification for Specific AI Use Cases

Responding to:

- Topic Area 12. How the U.S. can achieve and maintain effective leadership in AI technical standards development.

- Topic Area 15. How the Federal government should prioritize its engagement in the development of AI technical standards and tools

The IDSL believes that AI systems should be certified for specific use-cases much like drugs are approved for treating specific diseases on the label. The specific performance of these systems “on-label” is tractable and can be certified as meeting requirements through third-party testing. Systems that perform well can be certified as suitable for the use-case. However, performance in “off-label” use should be treated with more caution and with greater regulation since the performance characteristics and impact cannot be ascertained.

Programs already exist via the National Institute of Standards and Technology National Voluntary Laboratory Accreditation Program (NVLAP) to evaluate public and private labs on their technical qualifications and competence to carry out specific tests. We believe a similar framework in conjunction with an ISO 19795 type testing model and the concepts of “approved use cases” as mentioned above creates a strong standardization model that will encourage responsible usage and continued development of AI technologies in the US.

References

[1] Howard, J., Sirotin, Y., and Vemury, A. “The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance” .IEEE 10th Intl. Conference on Biometric, Theory, Applications and Systems. September 2019 (pending).

[2] Cook, C., Howard, J., Sirotin, Y, et al. “Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems”. IEEE Transactions on Biometrics, Behavior, and Identity Science, Volume: 1, Issue: 1. January 2019.

[3] Howard, J., Blanchard A., Sirotin, Y., et al. “An Investigation of High-Throughput Biometric Systems: Results of the 2018 Biometric Technology Rally”. IEEE 9th Intl. Conference on Biometric, Theory, Applications and Systems. October 2018.

[4] Howard, J., Blanchard A., Sirotin, Y., et al. “On Efficiency Effectiveness Tradeoffs in High Throughput Facial Biometric Systems”. IEEE 9th Intl. Conference on Biometric, Theory, Applications and Systems. October 2018.

[5] Grother, P., Quinn, G., Ngan, M., et al. “Face in Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects”. National Institute of Standards and Technology Interagency Report 8173. March 2019.

[6] Grother, P., Ngan, M., et al. “Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification”. National Institute of Standards and Technology Interagency Report XXXX Draft. June 2018.