Artificial Intelligence Measurement and Evaluation at the National Institute of Standards and Technology

AIME Planning Team

June 14, 2021

Preamble: This document is being distributed as read-ahead material for the 2021 NIST AI Measurement and Evaluation Workshop, to be held on June 15–17, 2021. It is in an abbreviated and draft form and will be expanded and updated based on feedback received during and after the workshop.

Executive Summary

As part of its mission, NIST seeks to support the advancement & deployment of *measured* Artificial Intelligence (AI) technologies. Toward that end, NIST's AI Measurement and Evaluation (AIME) focus area conducts research and development of metrics and measurement methods in emerging and existing areas of AI, contributes to the development of standards, and promotes the adoption of standards, guides, and best practices for the measurement and evaluation of AI technologies as they mature and find new applications. This document provides a brief, high-level summary of NIST's AIME research and standards activities. Note, this document is living—periodic updates will be published as progress is made and plans evolve.

1 Introduction

AI technologies have been and currently are being deployed in applications of consequence to individuals and people groups, often without effectively measuring crucial system properties. In several cases, AI systems have been deployed at some effort/expense, only to be abandoned (or "put on hold" for an extended period) when indicators suggest that the system is problematic with respect to one or more of these properties. It is clear that evaluation of and standards for the measurement and evaluation of such properties are critically needed as AI technologies move from the lab into society.

The National Institute of Standards and Technology (NIST) has a long history of measuring and evaluating Artificial Intelligence (AI) technologies in diverse areas, such as information retrieval [26], speech [23] and language processing [8], computer vision [21],

biometrics [3], and robotics [7]. These measurement and evaluation activities have typically focused on measures of performance accuracy[1] and the robustness of performance[2]. As the impacts of AI on society have grown and become more evident, several properties beyond performance accuracy measurement that were historically viewed as outside the purview of AI have been recognized as essential to measure and evaluate in order to ensure justified confidence in AI systems before and during deployment.[3] Measuring these properties will require either a) formulating them in terms of mathematical objectives/constraints (which would double as useful for machine learning systems to "learn to be un-biased", for example), or b) having some approach to "litmus test" or otherwise explicitly specify the degree to which AI systems possess these properties.

Objectives: NIST seeks to support the advancement & deployment of *measured* AI technologies, and has been assigned responsibility by statute to advance underlying research for measuring and assessing AI technologies, including the development of AI data standards and best practices, as well as AI evaluation and testing methodologies and standards. Toward that end, NIST will:

1. define, characterize, and theoretically & empirically analyze metrics and measurement methods for specified properties of AI technologies, to include the curation and characterization of appropriate data sets;

2. facilitate the measurement and evaluation of AI technologies with respect to these properties through the publication and presentation of technical reports, as well as activities such as: software tools publication, hosting evaluations of externally developed systems, and publication of best practices and technical guidance documents;

3. contribute to the development of voluntary consensus-based standards for evaluation of AI technologies, through the leadership of and/or participation in broad standardization efforts in support of the deployment of measured AI technologies;

4. build a strong and active community around the measurement and evaluation of AI technologies.

This document details NIST's current efforts and future plans to accomplish the above objectives.

1.1 Limitations/Disclaimer

There are diverse and many AI applications, each, potentially, being applied in a wide variety of settings. Further, there are several properties of an AI system that one may wish

[1] such as F-measures, area under the ROC curve, detection cost functions, etc.

[2] to noise, domain shift, data imbalance/scarcity/dimensionality/..., scale, learning setting, and others.

[3] for example, properties specific to social considerations–e.g. bias, data privacy–or out-of-domain practical considerations—e.g., model security.

to measure. For these reasons and more, we do not believe it is possible to develop a single and meaningful evaluation/metric/measurement-method to be used to measure all of AI as we currently know it. Further, even with great labor, NIST will not be able to address each potential application and setting individually. Instead, our goal is to pursue fundamental metrology research and to develop evaluations for the applications and settings where the need for measurement is greatest.

2 AIME Research and Standards Activities at NIST

NIST has a long history of AI measurement and evaluation activities, starting in the late 1960s with the measurement and evaluation of automated fingerprint identification systems [3]. Since then, NIST has designed and conducted hundreds of evaluations of thousands of AI systems; a list of many of NIST's AI technology evaluations can be found in the Appendix (§A.1). These activities have typically focused on measures of performance accuracy and robustness[4], and NIST will continue to engage in them. NIST also plans to expand its efforts in order to meet the objectives described in Section §1 that are not currently being met by existing AIME activities. In general, NIST will seek to specify the property being measured, detail the chosen metric(s) and measurement method(s) and their properties, and to apply the measurement method(s) to real systems, analyzing and publishing the results alongside curated data sets capable of supporting these measurements.

The remainder of this section is divided into subsections that parallel the objectives listed in Section §1. Each sub-section restates the corresponding objective, lists questions to be addressed in order to meet said objective, and describes NIST's current and planned related work.

In addition, a glossary defining basic terminology used throughout this document can be found in Section §3.

What is out of scope? There are certain topics that, from some perspective, involve the measurement and evaluation of AI, but will be considered out-of-scope for this effort. For example, testing G/TPUs or FPGAs or AI-hardware-Xs for speed are out of scope for this document.

2.1 Characterizing AI Metrics, Measurement Methods, and Data Sets

Objective: define, characterize, and theoretically & empirically analyze metrics and measurement methods for specified properties of AI technologies, to include the curation and characterization of appropriate data sets.

[4]There are some exceptions, e.g., [6] measures bias in face recognition and [1] measures bias in information retrieval.

Questions To Address:

1. What properties of an AI system can/should be measured? Which of these properties have/lack effective metrics and measurement methods?

2. What are the different measurement methods that are used to measure AI system properties? What are the strengths/limitations of each method and in what circumstances is one method preferred over another?

3. What are the different types and uses of metrics? What are the various properties of a metric, and under what circumstances is it important for a chosen metric to possess a given property?

4. What impacts do the chosen metrics and measurement methods have on an evaluation? When does the design/approach taken by an AI system impact the chosen metrics/measurement methods?

5. What data sets are needed in order to measure each property? Which of these currently exist and are readily available?

6. What are the important attributes of a data set to characterize and how should they be characterized?

2.1.1 Current/Future Work

NIST has been engaged in focused efforts[5] to establish common terminologies, definitions, and taxonomies of concepts pertaining to properties of AI systems in order to form the necessary underpinnings for (1) developing AI systems that possess these properties, and (2) developing metrics and measurement methods to assess the degree to which they do. A property of primary interest is AI system *trustworthiness*, which is understood to decompose into several component properties, including accuracy, bias mitigation, explainability and interpretability, privacy, resilience and security, reliability, robustness, and safety, among perhaps others.

For each of these properties, NIST has and/or will: 1) document the definitions, applications, tasks, as well as the metrics/measurement methods that are currently in use or being proposed within the research community to measure said property, the strengths/limitations of each metric/method, in what circumstances one metric/method might be preferred over another, and the overall relative maturity of the art and practice/need for new/improved metrics/methods for said property, 2) prepare and curate data sets, and characterize these data sets with respect to various attributes of interest, and 3) apply chosen metrics and measurement method(s) to real systems.

[5] including, e.g., in bias, explainability, privacy, security, transparency, and human-trust in AI

Here we list several of these properties and some of the foundational work being done at NIST for each (see also Sec §2.2 for descriptions of additional efforts in several on these properties):

1. Trustworthiness: Trustworthiness as a property of an AI system is complex, in the sense that it is made up of component parts, and contextual, in the sense that the relative amount each component matters (and, potentially, how a given component is measured) can change based on the context in which the AI system is operating. It is worth distinguishing between cases where trust in a system is justified (i.e., the system is trustworthy) versus when trust is given (i.e., users trust the system), referring to the later as "user trust" [24]. The exact component parts and their nature have been described in different ways by different efforts, and NIST is currently engaged in an effort to characterize and unify the different perspectives on AI trust. NIST recently hosted a workshop together with the National Academies of Science, Engineering, and Medicine on Trustworthy AI [10]. One of the recommendations from this workshop was for NIST to begin hosting a conference on Trustworthy AI (see Section (§2.2.1)). NIST also has plans to develop a Trust/Risk framework for accessing AI technologies, utilizing the metrics and measurement methods developed and/or proposed through its AIME activities.

2. Explainability/Interpretability/Transparency: NIST hosted a workshop on Explainable AI [19] and published a draft NISTIR "Four Principles of Explainable Artificial Intelligence" [22]. These provided discussion on principles guiding the development and measurement of explainable AI systems. The NISTIR [22] mentions two principles that encourage the development of metrics to determine if and how well they satisfy these two principles: "Meaningful: Systems provide explanations that are understandable to individual users", and "Explanation Accuracy: The explanation correctly reflects the system's process for generating the output" ([22], pg. 2). Recently a NISTIR published on "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" [2] which draws from literature in computer science, systems engineering, and experimental psychology to define the concepts of interpretability and explainability for AI enabled systems.

3. (Mitigated) Bias: NIST hosted a workshop on bias in AI [20], a virtual event to develop a shared understanding of bias in AI, what it is, and how to measure it. The measurement of bias in AI systems is difficult for several reasons, including the broad meaning/use of the term and the numerous ways through which bias can arise. NIST is currently researching the measurement of AI system bias and will host a summer workshop on developing a risk framework for AI bias.

4. Security: Since AI systems can be vulnerable to a wide range of attack types with various potential mitigation, defining and collecting standardized metrics that provide

insight into the nature of each attack as well as the effectiveness, usability, and applicability of protections is crucial. NIST has published a Draft NISTIR on "A Taxonomy and Terminology of Adversarial Machine Learning" [25] to develop a taxonomy of concepts to inform future standards and key practices for assessing and managing the security of machine learning components through establishing a common language and understanding of the rapidly developing adversarial machine learning landscape.

5. Privacy: NIST has been engaging in research in various privacy topics highly-relevant to AI systems, such as de-identification[16] and privacy risk assessment[17].

6. Safety: Measuring the safety of AI systems is challenging since, like bias, safety has many different meanings and uses, and, sometimes, like trustworthiness, safety is considered a complex property, e.g., consisting of specification, robustness and assurance. NIST is pursuing research in AI safety, including considering new formal methods, as well as existing methods from fields such as healthcare, insurance, and others.

The following list describes a few related recent and current AI metrology research projects conducted at NIST:

- *Advanced Statistical Metrology for Information Retrieval (IR):* This project seeks to (1) further the development and understanding of the analysis of variance method when appled to several of NIST's current IR test collections and, including exploring its strengths and weaknesses, (2) develop a method for comparing information retrieval systems across multiple test collections, and (3) explore methods for computing effect sizes in information retrieval experiments.

- *Evaluation Metrics for Automatically Generated Video Description:* The goal of this project is to design metrics for measuring systems that can describe a short video using natural language. Video-to-text technology has a number of applications including video summarization, automatic labeling and searching of videos, describing videos to people who suffer from visual impairment, cross-cultural understanding of videos with support of machine translation, etc.

- *Novel Measurement Methods for Information Retrieval Evaluation:* This effort seeks to improve the documentation and analysis methods of TextREtreival Conference (TREC) datasets as well as increase their impact in the research community, boosting the longevity of TREC datasets as well as increasing their impact in the research community through improved documentation and analysis methods. It also seeks to implement a testbed-based IR evaluation using category theory principles.

- *Active Evaluation:* The goal in active evaluation is to be able to minimize the amount of labeling necessary to obtain precise measurements of multiple algorithms performing

a single task. In particular, there are many tasks for which 1) there are large amounts of unlabeled data available for performance measurement, 2) the algorithms to be measured can process a large amount of data, and 3) labels can be obtained from a non-noisy oracle at a cost. For such tasks, one is able to use a large unlabeled dataset as the test set, and then interactively select which labels to obtain so as to jointly minimize labeling effort and uncertainty in the measurement. To perform this joint minimization, one can borrow methods from active learning. This approach to measuring algorithm performance is sometimes called "active evaluation".

- *Generative Models for Evaluation:* The goal of this project is to be able to use a limited amount of labeled data to more effectively measure system performance and robustness, where effectiveness is measured by the ability to produce confidence intervals that match the confidence intervals produced using large-scale evaluation datasets.

- *Neural Calculator:* Neural network calculator (NN Calculator) is an interactive visualization of neural networks that operates on datasets and NN coefficients as opposed to simple numbers. `https://pages.nist.gov/nn-calculator/`

- *Deep Video Understanding:* The High-level Video Understanding Project has established the first benchmark in High-level Video Understanding through our Workshop and Grand Challenge. Current work is underway to extend and expand this benchmark, both in terms of experimental data and ground truth,and to extend its scope, enbaling the evaluation and assessment of the current state of the art on a more focused and fine-grained scale.

- *Uncertainty in Clustering:* Clustering is a fundamental unsupervised learning task. Traditional clustering algorithms and many modern ones provide a point estimate solution. The focus of this work is to study exact and approximate inference methods and representations of uncertainty in flat and hierarchical clustering.

2.2 Conducting and Facilitating AI Measurement and Evaluation

Objective: facilitate the measurement and evaluation of AI technologies with respect to (specified) properties through the publication and presentation of technical reports, as well as activities such as: software tools publication, hosting evaluations of externally developed systems, and publication of best practices and technical guidance documents.

Questions To Address:

1. What are the different software needed to measure and evaluate AI? Which software could be general enough to be used across multiple and varied measurement and evaluation activities? Which software would be helpful for NIST to publish?

2. Which specific users and applications are most in need of evaluation?

3. What are the existing best practices for AI measurement and evaluation? Are they documented and, if so, where? What aspects of AI measurement and evaluation could benefit from additional guides and best practices?

2.2.1 Current/Future Work

NIST is currently running, organizing, or supporting evaluations of several AI technologies. As part of these evaluation processes, NIST publishes software, guidance and key practices documents, and technical reports. The rest of this section describes ongoing or upcoming NIST efforts to facilitate the measurement and evaluation of AI technologies.

Hosting AI System Evaluations: NIST will select topic areas for AI measurement and evaluation by sampling the spectrum of measurement type, task, domain, and modality. NIST seeks to focus its initial efforts on evaluations that address real-world use cases that correspond to existing/emerging applications that are of high-value and where the need for measurement and the likely impact is greatest.

Trustworthiness Inspired by the TextREtreival Conference (TREC), NIST will begin hosting an **AI TRUstworthiness Conference (TRUC)** that seeks to: (1) support research in trustworthy AI systems based on the availability and use of common data, metrics, and measurement methods, to increase communication among industry, academia, and Government on topics in AI trustworthiness by creating an open forum for the exchange of research ideas, and to improve the state of the art in the measurement and evaluation of AI system trustworthiness. TRUC will be overseen by a program committee consisting of representatives from government, industry, and academia. The TRUC cycle ends with a workshop that is a forum for participants to share their experiences.

(Mitigated) Bias NIST has been measuring and evaluating bias in information retrieval and face recognition for the last few years. **TREC 2021 Fair Ranking Track** [1]: The 2021 TREC Fair Ranking track evaluates systems according to how well they fairly rank documents. The 2021 tracks and focuses on fairly prioritising Wikimedia articles for editing to provide a fair exposure to articles from different groups. **FRVT Demographics Effects Analysis** [6]: As part of the Face Recognition Vendor Test (FRVT), NIST has conducted tests to quantify demographic differences in contemporary face recognition algorithms. Grother et al. [6] provides details about the recognition process, notes where demographic effects could occur, details specific performance metrics and analyses, gives empirical results, and recommends research into the mitigation of performance deficiencies.

Security As part of the IARPA **TrojAI** team, NIST has been evaluating AI security. Using machine learning, an artificial intelligence (AI) is trained on data, learns relationships in that data, and then is deployed to the world to operate on new data. The problem is that an adversary that can disrupt the training pipeline can insert Trojan behaviors into the AI. TrojAI's goal is to foster research into how to detect Trojans. NIST is building reference evaluation data and administering the challenge leaderboard where research teams submit trojan detection solutions for evaluation on sequestered datasets to evaluate their accuracy. `https://pages.nist.gov/trojai/`

Accuracy NIST has been involved in several evaluations, where different evaluation measures have been used for accuracy and error analysis, please see Appendix A.1.

Privacy The **Differential Privacy Synthetic Data Challenge** tasked participants with creating new methods, or improving existing methods of data de-identification, while preserving the dataset's utility for analysis [11].

Explainability NIST has nascent efforts to measure the explainability, interpretability, & transparency of AI systems. The principles discussed in the NIST-hosted workshop on Explainable AI [19] and the Draft NISTIR "Four Principles of Explainabile Artificial Intelligence" [22] provide a groundwork for different properties on which to evaluate explainable AI systems. As part of the User-Inspired AI project "Trusted AI framework for a new class of Standard Reference Materials and Data (SRM/SRD) with exquisitely characterized uncertainty for billions of properties", NIST will be hosting a small-scale explanation accuracy evaluation of the explainations given by the AI systems used to create new NIST SRM/SRD certified values.

Technical Reports NIST generally publishes NIST technical reports after each evaluation (see Appendix A.1 for a list of evaluations, many of which have such publications). NIST has also published three Interagency/Internal reports (NIST-IR's): "Four Principles of Explainable Artificial Intelligence" [22], "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence" [2] on Explainability and Interpretability, and "Trust and Artificial Intelligence" [24] on human-trust in AI, and has written a NIST-IR titled "A Proposal for Identifying and Managing Bias in Artificial Intelligence", with publication imminent. NIST is working on a paper titled "AI Measurement and Evaluation", that gives a high-level overview of the concepts, considerations, trade-offs, and complications of designing AI systems evaluations.

Software Tool Publication NIST has designed and is developing tools to advance efforts for the measurement and evaluation of AI systems, including software of several kinds:

Measurement Software NIST has written several software tools to support the different evaluations it hosts (see Appendix A.1 for several such evaluations). This includes software implementing the evaluation metrics (e.g., Classification Accuracy, Detection Error Trade-off (DET) Curves, Precision and Recall (PR), F-measures, Area Under ROC curve (AUC), Probability of Missed Detection at a Fixed False Alarm Rate (Pmiss@FA), Time-based False Alarm (TFA), Logarithmic Loss, Mean Squared Error, and many more), performance factor analysis software, and software to estimate the uncertainty in the obtained measurements.

Baseline, Naive, No Information, and State-of-the-Art Systems Baseline, naive, no-information, or current/previous state-of-the-art systems are often useful to be able to place obtained measurements in context, to track progress over time, and to validate evaluation data sets. NIST often develops and sometimes publishes reference implementations for these types of systems for the various evaluations it supports and will continue to do so as appropriate.

General Frameworks for AI Measurement and Evaluation NIST has developed software to support the AI system evaluation and analysis workflow for several of its evaluations, including TrojAI[18], ActEV[15], among many others. Future plans involve increasing the generality of these frameworks to increase code sharing and reuse. Toward that end, the National Cybersecurity Center of Excelence (NCCoE) is building the **NCCoE AI Software-Testbed**, a modular test bed for organizing and running machine learning experiments, currently focused on security concerns, but has been built with a goal to support additional test and evaluation use cases. This project is released as an open source project at: `https://github.com/usnistgov/dioptra`.

Best Practices and Technical Guides Rigorous measurement of AI systems requires certain practices, as does the ability to place the measured values in their appropriate context. The machine learning research community, and related industry (e.g., [5, 9]), have started emphasizing the importance of these practices. Much has been/is currently being done to promote best practices in ML generally (e.g., [4, 27]) and within specific domains, but there is still more to do. AIME will publish a survey document as well as one or more guidance documents in AI measurement and evaluation, leveraging the existing metrology expertise of NIST and the burgeoning output on this topic of various AI/ML communities. These will serve as both a unification effort (collecting and integrating best practices from many existing sources) as well as to identify the current limits of the practice of AI measurement. NIST will attempt to engage relevant AI research/practice communities in hopes of improving the content of the document(s), and promoting the broad adoption of this guidance, as appropriate.

AIME Handbook. One planned guidance document is the AIME Handbook. The AIME Handbook will be a user guide driven by use cases from AI application areas across NIST and the larger community. The handbook will consist of reproducible case studies and lessons-learned from a variety of domains, illustrating the core AIME challenges and community techniques when AI is applied to solving problems. The AIME Handbook will be community-driven, version-controlled living set of documents, developed in the well-accepted "feature-branch" workflow.

The AIME handbook will provided guidance/best practices for AI Measurement and evaluation within the following context. AI measurement and evaluation, when used to solve problems within a larger system, must be conceptualized in context with that system. Proper development, validation, and use of AIME must take into account not only AI methods, but data sources, application domain information, and end users' high level requirements as context. AI evaluation in practice must recognize the human aspect of it's use and development. Metrics should consider the means and value of feedback in a manner useful and intuitive to users for the domain of application. Where feasible, metrics should be flexible to include minor user specific interests about both the effect on the larger system and the impact and interface with end users. Where possible, evaluation methods and results should be designed to be accessible by business and engineering stakeholders that may not have strong backgrounds in AI.

2.3 Supporting AI Measurement and Evaluation Standardization

Objective: contribute to the development of voluntary consensus-based standards for evaluation of AI technologies, through the leadership of and/or participation in broad standardization efforts in support of the deployment of measured AI technologies.

Questions To Address:

1. What aspects of AI measurement and evaluation can/should be standardized?
2. What AIME standards efforts already exist and how should NIST engage with them?
3. Are there AIME standards needed that are not already being addressed? Which of these are ready for standards development and how should those not yet ready be driven toward standardization?

Once AIME metrics/measurement measurement methods for a given AI system property mature and become established, they will be considered as candidates for standardization. NIST is currently engaged in several AIME standards efforts, and seeks to expand its engagement with AIME standards over time. Here is a list of several AIME standards efforts that NIST is currently engaged with:

- *ASME V&V 50 Verification and Validation of Computational Modeling for Advanced Manufacturing:* The ASME V&V 50 Subcommittee provides procedures for verification, validation, and uncertainty quantification (VVUQ) in modeling and computational simulation for advanced manufacturing.

- *IEEE P7001 Transparency of Autonomous Systems:* IEEE P7001 Subcommittee is developing standards to (1) articulate measurable and testable levels of transparency for autonomous systems, both physical (e.g., autonomous vehicle or assisted living robot) and non-physical (e.g., medical diagnosis system), that accommodate a range of stakeholders and their needs, (2) validate and certify that autonomous systems conform to transparency standards, and (3) guide system developers self-assessment of transparency during development and suggest mechanisms for improving transparency.

- *IEEE P7003 Algorithmic Bias Considerations:* IEEE P7003 Subcommittee is developing standards to (1) provide individuals or organizations creating algorithmic systems (i.e., autonomous or intelligent systems) with certification-oriented methodologies that provide clearly articulated accountability and clarity around the validation data sets and how the algorithms are targeting, assessing and influencing the users and stakeholders of said algorithmic system, and (2) help developers of algorithmic systems and those responsible for their deployment to identify and mitigate unintended, unjustified and/or inappropriate biases in the outcomes of the algorithmic system.

- *ISO/IEC JTC 1/SC 42 Artificial Intelligence* ISO/IEC Artificial Intelligence (AI) standards development will be carried out by the JTC 1/SC 42 subcommittee, which was established in 2017. The JTC 1/SC 42 subcommittee's scope extends well beyond measurement and evaluation, however among its activities it will develop concepts and methods for testing of AI components in simulated and live environments.

2.4 Building AI Measurement and Evaluation Community

Objective: build a strong and active community around the measurement and evaluation of AI technologies. AIME is a necessarily expansive focus area that requires NIST-wide collaborations, alongside broad collaborations with academia, industry, and Government.

Questions To Address:

1. What are useful activities/infrastructure to organize/provide in order to support the AIME community?

2. Who are the stakeholders and how should they be engaged?

2.4.1 AI Measurement and Evaluation Presentation Series

NIST will host a bi-weekly AI metrology presentation series where leading researchers share current and recent work in AI measurement and evaluation. Among the objectives of this presentation series is to provide a dedicated venue for the presentation of AI metrology research and to spur collaboration among AI metrology researchers. The presentations will be open to the public and the presentation formats will be flexible, though will generally consist of 50-minute talks with 10 minutes of questions and discussion.

2.4.2 Periodic Workshops

NIST has held a series of AI workshops [14] that started with a workshop on trustworthy AI [13], then continued with a bias workshop [12], explainability [19], with the next one being this AIME workshop. NIST is planning to hold a document writing workshop on Risk of Bias Management this summer. Additionally, many evaluations hold their own events periodically. See Section 2.2.1 for more information and a description of some of the AI evaluations hosted at NIST. Additionally, NIST will begin hosting an AI TRUstworthiness Conference (TRUC), inspired by the TextREtreival Conference (TREC). See Section 2.2.1 for more information on TRUC.

3 Glossary

Please note, that these definitions are given w.r.t. the context they are used in this document and are not intended to be general/universal/comprehensive. We intend to expand and formalize this glossary and, when useful, we will also note different uses of the same terms/terms w/the same meaning sometimes used in different research communities.

(AI) System: A (fully trained) system that can be applied to a specific AI task. Sometimes shortened to "system".

Best Practices: Community-wide guidelines that are generally the accepted approach for how to do something.

Evaluation: A formally defined, fully contained set of experiments including, documentation (including defining the task and otherwise describing the experiments), data, metrics, measurement methods, system descriptions, measurement results, and analysis.

Guides: Documents describing NIST's proposed approach for how to do something.

Measure: (v) take/obtain one or more measurements.

Measurement: (n) the application of a given metric to a given AI system using a given measurement method.

Measurement Method: a formally described process through which a measurement is taken.

Metric: Loosely speaking, a mathematical function mapping from the set of possible system outputs to the set of real-valued numbers (alternatively, to a set of categories). (cf.,

the typical definition of metric used in mathematics.)

References

[1] A. J. Biega, F. Diaz, M. D. Ekstrand, and S. Kohlmeier. Overview of the TREC 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*, 2019. URL `https://fair-trec.github.io/`.

[2] D. A. Broniatowski et al. Psychological foundations of explainability and interpretability in artificial intelligence. 2021. URL `https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf`.

[3] M. D. Garris and C. L. Wilson. Nist biometric evaluations and developments. In *Photonics for Port and Harbor Security*, volume 5780, pages 26–38. International Society for Optics and Photonics, 2005.

[4] Google. Best practices for ml engineering. `https://developers.google.com/machine-learning/guides/rules-of-ml/`, 2021. [Online; accessed 9-June-2021].

[5] Google. Responsible ai practices. `https://ai.google/responsibilities/responsible-ai-practices/`, 2021. [Online; accessed 9-June-2021].

[6] P. J. Grother, M. L. Ngan, and K. K. Hanaoka. Face Recognition Vendor Test Part 3: Demographic Effects. NISTIR 8280, National Institute of Standards and Technology, Dec. 2019. URL `https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf`.

[7] A. Jacoff, E. Messina, and J. Evans. Performance evaluation of autonomous mobile robots. *Industrial Robot: An International Journal*, 2002.

[8] A. F. Martin and M. A. Przybocki. Nist 2003 language recognition evaluation. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[9] MLCommons. Ml commons. `https://mlcommons.org/en/`, 2021. [Online; accessed 9-June-2021].

[10] National Academies. Opportunities for measurement science research at nist to advance the trustworthiness of ai systems: A workshop. `https://www.nationalacademies.org/our-work/opportunities-for-measurement-science-research-at-nist-to-advance-the-trustworthiness-of-ai-systems-a-workshop/`, 2021. [Online; accessed 9-June-2021].

[11] NIST. 2018 differential privacy synthetic data challenge. `https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic`, 2018. [Online; accessed 9-June-2021].

[12] NIST. Bias in AI Workshop, June 2020. URL https://www.nist.gov/news-events/events/2020/08/bias-ai-workshop. Last Modified: 2020-08-21T10:52-04:00.

[13] NIST. Exploring AI Trustworthiness: Workshop Series Kickoff Webinar, June 2020. URL https://www.nist.gov/news-events/events/2020/08/exploring-ai-trustworthiness-workshop-series-kickoff-webinar. Last Modified: 2020-08-28T14:53-04:00.

[14] NIST. NIST AI Workshop Series, June 2020. URL https://www.nist.gov/artificial-intelligence/nist-ai-workshop-series. Last Modified: 2021-05-25T09:15-04:00.

[15] NIST. Activities in extended video. https://actev.nist.gov/, 2021. [Online; accessed 9-June-2021].

[16] NIST. Privacy engineering program. https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id, 2021. [Online; accessed 9-June-2021].

[17] NIST. Privacy engineering program. https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/risk-assessment, 2021. [Online; accessed 9-June-2021].

[18] NIST. Nist trojai. https://pages.nist.gov/trojai/, 2021. [Online; accessed 9-June-2021].

[19] NIST. Explainable AI Workshop, Jan. 2021. URL https://www.nist.gov/news-events/events/2021/01/explainable-ai-workshop. Last Modified: 2021-01-22T13:35-05:00.

[20] NIST. Bias in ai workshop. https://www.nist.gov/news-events/events/2020/08/bias-ai-workshop/, 2021. [Online; accessed 9-June-2021].

[21] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quéot. Trecvid 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. 2013.

[22] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki. Four Principles of Explainable Artificial Intelligence. Draft NISTIR 8312, NIST, Aug. 2020. URL https://doi.org/10.6028/NIST.IR.8312-draft.

[23] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero. The 2016 nist speaker recognition evaluation. In *Interspeech*, pages 1353–1357, 2017.

[24] B. Stanton and T. Jensen. Trust and artificial intelligence. 2021. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087.

[25] E. Tabassi, K. Burns, M. Hadjimichael, A. Molina-Markham, and J. Sexton. A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269, NIST, Oct. 2019. URL https://doi.org/10.6028/NIST.IR.8269-draft.

[26] E. M. Voorhees, D. K. Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, MA, 2005.

[27] B. Wujek, P. Hall, and F. Günes. Best practices for machine learning applications. *SAS Institute Inc*, 2016.

A Appendix

A.1 List of NIST AI Evaluations

A.1.1 Computer Vision

The NIST computer vision program includes several activities contributing to the development of technologies that extract information from image and video streams through systematic, targeted annual evaluations and metrology advances. https://www.nist.gov/programs-projects/video-analytics

Media Forensics Challenge The Media Forensics Challenge (MFC) Evaluation is the annual evaluation to support research and help advance the state of the art for image and video forensics technologies – technologies. https://mfc.nist.gov/

TrojAI Using machine learning, an artificial intelligence (AI) is trained on data, learns relationships in that data, and then is deployed to the world to operate on new data. The problem is that an adversary that can disrupt the training pipeline can insert Trojan behaviors into the AI. TrojAI's goal is to foster research into how to detect Trojans. NIST is building reference evaluation data and administering the challenge leaderboard where research teams submit trojan detection solutions for evaluation on sequestered datasets to evaluate their accuracy. https://pages.nist.gov/trojai/

ActEV The Activities in Extended Video (ActEV) series of evaluations is designed to accelerate development of robust, multi-camera, automatic activity detection systems in known/unknown facilities for forensic and real-time alerting applications. https://actev.nist.gov/sdl

TRECVID The TREC Video Retrieval Evaluation (TRECVID) is an ongoing series of evaluation to promote progress in content-based video analysis and retrieval via open, metrics-based evaluation. https://trecvid.nist.gov

Multimedia Event Detection The goal of Multimedia Event Detection (MED) is to assemble core detection technologies into a system that can search multimedia recordings for user-defined events based on pre-computed metadata. The metadata stores developed by the systems are expected to be sufficiently general to permit re-use for subsequent user defined Ad-Hoc events. https://www.nist.gov/itl/iad/mig/multimedia-event-detection

Surveillance Event Detection Surveillance Event Detection (SED) The objective of the Surveillance Event Detection evaluation is to promote the development of technologies that detect activities that occur in the surveillance video domain. https://www.nist.gov/itl/iad/mig/trecvid-surveillance-event-detection-evaluation-track

Video Surveillance Technologies for Retail Security The Video Surveillance Technologies for Retail Security (VISITORS) project was to advance predictive analysis technologies and methodologies that are able to detect persons engaged in suspicious activities in surveillance video. VISITORS is being applied in the retail domain. https://www.nist.gov/itl/iad/mig/video-surveillance-technologies-retail-security-visitors

CLEAR Classification of Events, Activities and Relationships (CLEAR) was a multinational evaluation series that brought together the researchers from the US ARDA VACE Program and the European Union Computers in the Human Interaction Loop Program to focus research on detecting and tracking people, faces, vehicles, etc. and acoustic event detection.

VACE Video Analysis and Content Extraction (VACE) program was established to develop novel algorithms for automatic video content extraction, multi-modal fusion, and event understanding. During the program, progress was made in the automated detection and tracking of moving objects, including faces, hands, people, vehicles, and text.

Handwriting Recognition and Translation Evaluation The NIST Open Handwriting Recognition and Translation Evaluation (OpenHaRT) is an evaluation of transcription and translation technologies for document images. The evaluation seeks to break new ground in the areas of document image recognition and translation toward the goal of document understanding capabilities. The objective is to assess the current state-of-the art and to build the critical mass required to solve challenges posed in these areas so that technologies developed from OpenHaRT can be used to distill the vast amount of information available only in foreign language documents in a timely manner. https://www.nist.gov/itl/iad/mig/openhart

SD19: Special Database 19 contains NIST's entire corpus of training materials for handprinted document and character recognition.
https://www.nist.gov/srd/nist-special-database-19

A.1.2 Information Retrieval

The information retrieval research involving large, human generated text, speech, and video files by providing test collections and organizing the TREC, TAC, and TRECVID conferences and their proceedings. NIST continues to create new test collections, focusing mainly on collections to support specific information retrieval sub-tasks such as cross-language retrieval and multimedia retrieval. We also develop better evaluation methodology for information access, including improved evaluation measures for comparing systems using test collections and new evaluation measures for interactive searching and browsing operations.

Text REtrieval Conference The Text REtrieval Conference (TREC) is an ongoing series of evaluation workshops focusing on a list of different information retrieval (IR) research areas. `https://trec.nist.gov`

Text Analysis Conference The Text Analysis Conference (TAC) is a series of evaluation workshops organized to encourage research in Natural Language Processing and related applications, by providing a large test collection and common evaluation procedures. `https://tac.nist.gov`

TREC Video Retrieval Evaluation The TREC Video Retrieval Evaluation (TRECVID) is an ongoing series of evaluation to promote progress in content-based video analysis and retrieval via open, metrics-based evaluation. `https://trecvid.nist.gov`

Spoken Document Retrieval The spoken document retrieval (SDR) evaluation designs and implements evaluations of Spoken Document Retrieval (SDR) technology within a broadcast news domain. SDR involves the search and retrieval of excerpts from spoken audio recordings using a combination of automatic speech recognition and information retrieval technologies.

A.1.3 Speech Processing

NIST's Speech Processing program has a long history of activities supporting the development of technologies that extract content from recordings of spoken language and of metrology advancements, primarily through systematic and targeted annual evaluations. NIST's research in speech processing supports broad technology areas, including speech recognition, speaker recognition, diarization, speech activity dectection, language recognition, keyword spotting, rich transcription, and speech-to-speech translation.

Automatic Speech Recognition NIST has a long history of conducting evaluations in Automatic Speech Recognition (ASR). Recetnly, the focus of the NIST's ASR evaluations

is to assess the state of the art of ASR technologies for low-resource languages. https://www.nist.gov/itl/iad/mig/openasr-challenge

Speaker Recognition The NIST Speaker Recognition Evaluation (SRE) is an ongoing series of speaker recognition evaluations conducted by NIST since 1996. The objectives of the evaluation series are to measure system performance of the current state of technology. https://www.nist.gov/itl/iad/mig/speaker-recognition

Open Speech Analytic Technologies The Open Speech Analytic Technologies (OpenSAT) Evaluation Series focuses on the following tasks: Automatic Speech Recognition (ASR), Speech Activity Detection (SAD), and Keyword Search (KWS). https://www.nist.gov/itl/iad/mig/opensat

Language Recognition The NIST Language Recognition Evaluation (LRE) series is to evaluate the performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field. https://www.nist.gov/itl/iad/mig/language-recognition

Speech Activity Detection The purpose of a Speech Activity Detection (SAD) system is to find regions of speech in an audio file. The NIST Open Speech-Activity-Detection evaluation (OpenSAD) is intended to provide Speech-Activity-Detection system developers with an independent evaluation of performance on a variety of audio data. https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation

Rich Transcription The Rich Transcription evaluation series promotes and gauges advances in the state-of-the-art in several automatic speech recognition technologies. The goal of the evaluation series is to create recognition technologies that will produce transcriptions which are more readable by humans and more useful for machines. https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

Keyword Spotting An annual evaluation of technologies that perform keyword search in a new language each year. The evaluation is an outgrowth of the 2006 Spoken Term Detection evaluation. https://www.nist.gov/itl/iad/mig/open-keyword-search-evaluation

Spoken Document Retrieval The spoken document retrieval (SDR) evaluation designs and implements evaluations of Spoken Document Retrieval (SDR) technology within a broadcast news domain. SDR involves the search and retrieval of excerpts from spoken audio recordings using a combination of automatic speech recognition and information retrieval technologies.

A.1.4 Natural Language Processing

NIST's Natural Language Processing (NLP) research involves developing large corpora of human-generated text as well as common metrics and evaluation procedures. NIST's research in NLP supports broad technology areas, including information retrieval, machine translation, and low-resource language applications.

Text REtrieval Conference The Text REtrieval Conference (TREC) is an ongoing series of evaluation workshops focusing on a list of different information retrieval (IR) research areas. `https://trec.nist.gov`

Text Analysis Conference The Text Analysis Conference (TAC) is a series of evaluation workshops organized to encourage research in Natural Language Processing and related applications, by providing a large test collection and common evaluation procedures. `https://tac.nist.gov`

Machine Translation The Multimodal Information Group's machine translation (MT) program includes several activities contributing to machine translation technology and metrology advancements, primarily through systematic and targeted annual evaluations. `https://www.nist.gov/programs-projects/machine-translation`

LORELEI/LOREHLT Low Resource Languages for Emergent Incidents (LORELEI) was a DARPA-sponsored program. The goal of the program is to dramatically advance the state of computational linguistics and human language technology to enable rapid, low-cost development of capabilities for low-resource languages. The Low Resource HLT (LoReHLT) open evaluations serve to evaluate component technologies relevant to LORELEI. `https://www.nist.gov/itl/iad/mig/lorehlt-evaluations`

A.1.5 Biometrics

Biometrics are human measurements that can be used to identify a person for a variety of applications, e.g., to grant access to devices, systems or data. NIST have been testing and evaluating biometric recognition technologies and assisting in determining where and how biometric recognition technology can best be deployed.

Face Recognition Face Recognition research and Face Recognition Vendor Tests (FRVT) provide independent government evaluations of face recognition technologies and assist in determining where and how facial recognition technology can best be deployed. `https://www.nist.gov/programs-projects/face-projects`

Fingerprint ITL evaluates fingerprint matching technologies by developing datasets to support standards, measurement and evaluation methods, and technology capabilities. https://www.nist.gov/programs-projects/fingerprint

Biometric Quality Performance of biometric systems is dependent on the quality of the acquired input samples. If quality can be improved, either by sensor design, by user interface design, or by standards compliance, better performance can be realized. For those aspects of quality that cannot be designed-in, an ability to analyze the quality of a live sample is needed. https://www.nist.gov/programs-projects/biometric-quality-homepage

Iris ITL/IAD conducted and managed the Iris Challenge Evaluation (ICE) projects https://www.nist.gov/programs-projects/iris-projects

Tattoo Recognition The Tattoo Recognition Technology Program features a family of activities designed with goals to evaluate and measure image-based tattoo recognition technology. https://www.nist.gov/programs-projects/tattoo-recognition-technology

Speaker Recognition The NIST Speaker Recognition Evaluation (SRE) is an ongoing series of speaker recognition evaluations conducted by NIST since 1996. The objectives of the evaluation series are to measure system performance of the current state of technology. https://www.nist.gov/itl/iad/mig/speaker-recognition

A.1.6 Forensics

NIST is working to strengthen forensic practice through research and improved standards. Our efforts involve three key components: science, policy, and practice.

Media Forensics Challenge The Media Forensics Challenge (MFC) Evaluation is the annual evaluation to support research and help advance the state of the art for image and video forensics technologies – technologies. https://mfc.nist.gov/

TrojAI Using machine learning, an artificial intelligence (AI) is trained on data, learns relationships in that data, and then is deployed to the world to operate on new data. The problem is that an adversary that can disrupt the training pipeline can insert Trojan behaviors into the AI. TrojAI's goal is to foster research into how to detect Trojans. NIST is building reference evaluation data and administering the challenge leaderboard where research teams submit trojan detection solutions for evaluation on sequestered datasets to evaluate their accuracy. https://pages.nist.gov/trojai/

Fingerprint ITL evaluates fingerprint matching technologies by developing datasets to support standards, measurement and evaluation methods, and technology capabilities. `https://www.nist.gov/programs-projects/fingerprint`

A.1.7 Material Science

JARVIS-ML JARVIS-ML introduced Classical Force-field Inspired Descriptors (CFID) as a universal framework to represent a material's chemistry-structure-charge related data. With the help of CFID and JARVIS-DFT data, several high-accuracy classifications and regression ML models were developed `https://www.nist.gov/programs-projects/jarvis-ml`