IBM Response to NIST RFI on Artificial Intelligence (AI)

Contacts: John R. Smith (jsmith@us.ibm.com), IBM T. J. Watson Research Center,
Mark C. O'Riley Esq. (mcoriley@us.ibm.com), IBM Government and Regulatory Affairs

Executive Summary

IBM appreciates the opportunity to respond to the Request for Information (RFI) on Artificial Intelligence (AI) Standards issued by National Institute of Standards and Technology (NIST) on May 1, 2019. As a worldwide leader in Information Technology (IT), which includes leadership in research, development and application of AI across a broad set of industries and enterprise domains, IBM is pleased to share its deep expertise and insight.

Critically, we hope NIST will recognize AI as an emerging general-purpose technology that will transform industry. The rapid pace at which the field is developing is a strong indication that AI is still in an early stage. As such, NIST's work should be directed foremost at accelerating AI technology advancements by contributing to the development of open frameworks, shared definitions, and related tools – including evaluations, data sets, and metrics, — rather than creating technical standards at this stage, so as to not introduce premature barriers to innovation.

Like NIST's work on risk-based cybersecurity and privacy frameworks, NIST should convene stakeholders to create an overall AI accountability framework that provides a shared conceptual foundation and guidance around important aspects of trustworthy AI – including fairness, explainability, robustness, and transparency — and fosters development of trust-related evaluations, data sets, and metrics. Such a framework will accelerate tooling and benchmarking for the safe testing and deployment of AI and support the development of trustworthy AI.

Further, we encourage NIST to continue to engage in and build from existing efforts, including those within traditional standard setting organizations like International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) and Institute of Electrical and Electronics Engineers (IEEE), as well as with global fora such as the European Commission, which recently released the *Ethics Guidelines for Trustworthy AI*, and with internal industry-government coordination activities, such as the *NIST Privacy framework*.

Doing so, NIST will help America to continue to define and achieve US leadership in AI.

Detailed Response

Historically, IBM has joined with the majority of industries that support and espouse the open, private-sector-led, consensus-based, voluntary standard setting approaches of the United

States. With the charter of ANSI and the government's light hand in standard setting activities, current industry and government practices have worked to foster innovation and keep US companies competitive around the world.

NIST, with its mission to promote innovation and industrial competitiveness, has played a vital role in supplying industry, academia, government, and other organizations with support around standards, guidelines, and measurement and evaluation methodologies. As AI is emerging as an important foundational technology that will transform industry, NIST should foster technical innovation and industrial competitiveness in AI by:

- Engaging with stakeholders to create an overall AI accountability framework that provides a conceptual foundation of shared definitions for trustworthy AI that includes fairness, explainability, robustness, and transparency;
- Organizing open evaluations using increasingly large and diverse data sets for training and testing of fair, accurate, explainable, and robust AI technologies and systems at scale; and
- Developing a strong technical foundation of evaluation protocols and measures for trustworthy AI, which includes both cross-cutting and industry sector-specific metrics.

The right balance needs to be achieved with standardization. AI technologies hold great promise to transform US industry in a tremendously positive way [1]. However, the fast pace at which the AI field is still developing is a strong indication that many AI technologies are still in a nascent state. As such, any development related to standards and tools should be principally directed toward accelerating technical innovation, with care taken to not introduce unnecessary bottlenecks to progress in the field.

Trustworthy AI

Fairness, explainability, robustness, and transparency underscore *IBM's Principles for Trust & Transparency*[1]. These principles are essential throughout the entire AI lifecycle to effectively ensure trust. This includes all stages of AI applications spanning specification of requirements, collection of data, building of models, deployment, and operation. Although these principles apply generally, different industry sectors, application domains, and use cases have specialized requirements. These can result from the unique needs of specific users or because of regulatory or compliance obligations. Ultimately, trustworthy AI may require some form of standardization, either explicit or de facto, for some or all of the principles. However, creating a standard prematurely, before AI technology matures over the coming decades, can drive industry to adhere to substandard practices and dampen innovation.

Recommendation: NIST should help establish an overall AI accountability framework that provides a shared conceptual foundation with a consistent set of definitions for *trustworthy AI* and fosters development of trust-related tools including evaluations, data sets, and metrics.

[1] https://www.ibm.com/blogs/policy/trust-principles/

This foundation will help advance technical innovation and industry competitiveness in the development of trustworthy AI. More specific recommendations follow for individual dimensions of trust, including fairness, explainability, robustness, and transparency.

Fairness

Since many AI systems use machine learning to train AI models, the resulting systems can directly reflect biases. This can have negative consequences for fairness when unwanted bias influences model outputs. Researchers in machine learning and social justice have shown that there are many ways to measure this bias [2]. However, there is no single metric or mitigation technique that can form a standard for fairness. To help accelerate research on this topic, IBM has released an open source toolbox called *AI Fairness 360 (AIF360)* [3]. AIF360 implements more than ten bias mitigation algorithms and seventy state-of-the-art metrics related to fairness in a common software framework[2]. The AIF360 toolbox, which includes extensive educational material, enables researchers and practitioners to experiment with different metrics and mitigation methods to determine what is most appropriate for a specific situation. AIF360 is actively being used and developed by the open source community and has been incorporated into coursework at universities. The AIF360 toolbox is industry sector neutral, and thus, can be applied to a wide range of problem domains.

Recommendation: NIST should take actions to accelerate technical progress related to *fairness*:

- NIST should encourage the development and integration of new fairness metrics and methods into a common framework, like AIF360, so that comparisons among bias mitigation methods and metrics can be made by researchers, social scientists, policy makers, and machine learning experts.
- NIST should encourage the creation of diverse data sets to facilitate the development of fair and accurate AI systems across different modalities such as vision, speech, language and other forms of structured and unstructured data.
- NIST should administer open challenges to accelerate research and development on fairness and bring focus to different requirements across industry sectors and domains.

This should be done in an open, inclusive, and independent manner to allow full participation by researchers and practitioners.

Explainability

To establish trust and confidence in AI systems, it is important for the outputs of AI systems to be easily understood, as well as the process by which the outputs are produced. In practice, there are many possible stakeholders that need this understanding: data scientists (seeking to improve their model or the underlying training data); developers (working to debug their AI application); end users (wanting to understand why the system made a specific recommendation e.g. a doctor evaluating a medical treatment plan); affected users (e.g. a

[2] https://aif360.mybluemix.net

patient trying to understand a diagnoses); and regulators (ensuring a system is fair). These users have varying domain expertise and technical sophistication. Thus, we need a general way to measure the usefulness of explanations. Metrics are needed that adequately capture explainability and allow researchers and practitioners to compare and improve techniques. A good example is the recent *FICO Explainable Machine Learning Challenge*[3], which asks participants to create a model that simultaneously predicts credit risk and provides explanations for predictions. The results are assessed subjectively by human judges, since there are no single objective correct answers. The FICO challenge is a good initial step toward developing evaluations for explainability by providing ground-truth for both answers and explanations as well as metrics for judging explanations. However, it also illustrates the early state of technology for explainability and the technical gaps that remain.

Recommendation: NIST should facilitate activity for advancing the technical evaluation and instrumentation for *explainability* to:

- Attain a more complete understanding of the various needs for explainability from users and stakeholders;
- Make data sets with scored explanations openly available to further drive research and technology development for explainability; and
- Define metrics for measuring the usefulness of explanations.

One way to do this would be for NIST to create challenges that focus on specific problems within different industry sectors, where in each case a stakeholder provides a clear picture of what constitutes a meaningful explanation. Organizing a series of such challenges could facilitate the creation of valuable data sets with explanations and help gain important insight regarding the needs of users and metrics for explainability.

Robustness

Robustness is essential for trustworthy AI and aims to ensure the resiliency of AI systems. This includes hardening of models to make them robust to adversarial attacks and validating them using metrics and benchmarks. An important aspect of robustness is the careful tracking of data provenance and models to protect against attacks such as poisoning. The security of the runtime for AI models is also important for guarding against unwanted access, manipulation or compromise. To advance the study of robustness including evaluating and hardening AI models, IBM has released the open source *Adversarial Robustness Toolbox (ART)*[4]. ART implements state-of-the-art attacks and defenses, including adversarial training and data poisoning detection, as well as multiple metrics for robustness. Other related work is developing more general measures for robustness, such as CLEVER [4], CNN-Cert [5], and Randomized Smoothing. Current best practices involve subjecting an AI model to a known set of attacks

[3] https://community.fico.com/s/explainable-machine-learning-challenge

[4] https://github.com/IBM/adversarial-robustness-toolbox

using specific strength metrics, typically based on L_p norms, in both white- and black-box settings. However, current attack measures do not correlate well with human perception and are available only in some data domains like images. The AI community needs a stronger foundation of these measures across data modalities to address the requirements for achieving robustness across a broader, more diverse set of AI models.

Recommendation: NIST should develop technical metrics and benchmarks for evaluating the *robustness* of AI models against attacks. These attacks should include methods based on adversarial samples, poisoning, model inversion, and others, and span a wide range of data modalities and model types.

Transparency

Transparency of AI refers to the need of users to understand the intended purpose of an AI system, including how it was developed, to ensure appropriate usage. This need is not specific to AI and is seen within many established industries from finance to children's toys to packaged food. The AI community is exploring different mechanisms for transparency including approaches similar to the "nutritional label" for food products. For example, the EU recently released a checklist that includes AI factsheet components aimed at providing transparency. So far, efforts have been sector-neutral, but it is likely that the specific contents of an AI factsheet will depend on the specific needs of each industry sector.

Recommendation: NIST should engage with stakeholders in the context of an overall AI accountability framework to establish definitions related to *transparency*. AI factsheets are an important potential future direction for AI transparency. At this point, factsheets should be voluntary since consensus on the content and format still needs to be developed with input from different stakeholders including suppliers and consumers of AI services. Care needs to be taken to balance the requirements for transparency with the needs for AI service providers to have flexibility for continuous innovation.

Ethics and Responsibility

IEEE has initiated 14 projects related to AI ethics in the context of the P7000 series of standards[5]. They were developed by the *IEEE Global Initiative on Ethical Considerations on Autonomous and Intelligent Systems*[6] as accompanying instruments to the IEEE book: *Ethically Aligned Design* [6]. The High-Level Expert group on AI, nominated by the European Commission, recently published the *Ethical Guidelines for Trustworthy AI*[7]. These guidelines define seven requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, fairness, societal and environmental wellbeing, and

[5] https://ethicsinaction.ieee.org

[6] https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[7] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

accountability. The document also provides a checklist that organizations can use to assess these seven properties of their AI systems. Use of this checklist will be piloted by stakeholders between June and November 2019, with final version planned for December 2019 release. The checklist may evolve toward a global standard if there is consensus around its content and value in practice. Additionally, ISO/IEC JTC 1 has recently established SC 42[8] to be the focal point for standardization on AI. SC 42 has launched multiple working groups, including WG 3 on *Trustworthiness*, which is drafting a technical report on trustworthiness in AI as well as on robustness of neural networks and bias in AI systems.

Recommendation: NIST should maintain awareness of global efforts related to *AI ethics*, including establishment of best practices, guidelines and standards, and follow the related activities from IEEE, European Commission, and ISO/IEC JTC 1 SC 42.

Privacy

There has been significant discussion on AI's impact on individual privacy including but not limited to knowledge and consent for the collection and use of individuals' personal information or the creation of personal information through inferences as well as individual data access, correction and protection. While some of these topics have been addressed, partly by IEEE (see above), they should also be addressed, at a high-level, by the *NIST Privacy Framework*[9] currently in development.

Recommendation: Any additional standards or tools touching on *privacy* and useful for the development of AI should supplement the *NIST Privacy Framework* and not be independent efforts. This "future proofed" Framework is intended to be principle-based, technology-neutral, and applicable across various sectors and businesses. The Framework should therefore by definition apply to the development of AI technologies.

Evaluations

NIST has long partnered with the Linguistic Data Consortium (LDC) to host evaluations related to human language technology, which provide training, development, and test data for research areas that include speech recognition, language recognition, machine translation, cross-language retrieval, and multimedia retrieval[10]. NIST has also co-sponsored the TREC (Text Retrieval Conference)[11] since 1992, which has resulted in significant improvement to information retrieval technology and substantial economic benefit to US industry [7]. NIST TREC efforts today extend to more sophisticated AI tasks including complex question answering,

[8] https://www.iso.org/committee/6794475.html

[9] https://www.nist.gov/privacy-framework

[10] https://www.ldc.upenn.edu/collaborations/evaluations/nist

[11] https://trec.nist.gov

incident management, and news summarization, as well as to industry specific challenges such as building systems in healthcare that use data to link oncology patients to clinical trials for new treatments and evidence-based literature to identify the most effective existing treatments[12]. NIST TREC has also expanded into modalities beyond text, such as with the NIST TRECVID evaluations for tasks related to digital video[13], NIST Multimedia Event Detection (MED), and NIST Multimedia Event Recounting (MER)[14]. These evaluations are important for driving fundamental advancements in the accuracy of AI technologies on a growing field of tasks using data modalities such as images, video, speech, and text.

Recommendation: NIST should organize *evaluations* specifically related to trustworthy AI and look for ways to expand its ongoing evaluations to incorporate essential aspects of fairness, explainability, robustness, and transparency, in addition to accuracy, for a growing set of cross-cutting and industry-specific AI tasks and data modalities.

Data sets

Data sets have been fundamental for AI with the advent of modern neural network-based machine learning, where data is essential for training and applying AI models. Prominent examples of AI data sets include MNIST[15], CIFAR[16], ImageNet[17], and PASCAL VOC[18], where each is responsible for driving thousands of projects and efforts across industry and academia. As the AI field evolves toward more complex problems at enterprise- and industry-scale, there are needs for larger and larger data sets. One US company recently reported training a state-of-art image recognition system using a proprietary data set of more than 3 billion consumer photos [8]. Another AI company has built its own proprietary training data set of 2 billion face images, which they use to develop their face recognition system[19]. Training data at this scale can be essential for advancing AI technologies across a wide range of data modalities and problem domains but is not uniformly or openly available for research and development in AI.

Recommendation: NIST should play an active role in making large *data sets* available to the AI community for training, development, and testing. Where possible, NIST should explore ways to grow access to larger, more diverse data sets by working with US agencies, industry partners

[12] https://trec.nist.gov/pubs/call2019.html

[13] https://trecvid.nist.gov

[14] https://www.nist.gov/itl/iad/mig/multimedia-event-detection

[15] http://yann.lecun.com/exdb/mnist

[16] https://www.cs.toronto.edu/~kriz/cifar.html

[17] http://www.image-net.org

[18] http://host.robots.ox.ac.uk/pascal/VOC

[19] https://www.forbes.com/sites/shuchingjeanchen/2018/03/07/the-faces-behind-chinas-omniscient-video-surveillance-technology

and other stakeholders. These data sets should enable further technology development specifically related to trustworthy AI.

Metrics

Metrics are essential for providing a quantitative foundation to compare AI systems and measure performance of AI tasks. In its TREC evaluations, NIST has helped to establish important metrics for the AI field – *precision vs. recall*, *mean average precision*, and *false alarm vs. miss rate*. Industry has also played a prominent role in the development of metrics, such as in the case of the *BLEU* metric created by IBM Research, which has achieved wide use for evaluating natural language-related AI tasks [9]. More recent industry- and academia-driven evaluations have used metrics such as *top-1* and *top-5 accuracy* for evaluating classification results and *intersection-over-union* to measure localization in object detection. While some metrics such as these have a cross-cutting relevance, different industries require specific metrics. For example, medical diagnosis is evaluated using *sensitivity vs. specificity*, which is unique to healthcare. Other industries and enterprise applications also have specific metrics, for example, *call deflection rate* is an important metric for AI-assisted customer care, which provides an application-level measure of the effectiveness of an AI assistant to use speech, language and dialog technologies to perform its tasks.

Recommendation: NIST should develop *metrics* for trustworthy AI that provide technical measures for fairness, explainability, robustness, transparency, accuracy and overall effectiveness of AI systems. This investigation should consider both cross-cutting and industry specific requirements as well as the needs of a growing field of AI tasks that use data modalities such as images, video, speech, and text.

Summary

NIST has an opportunity to play a vital role in helping America achieve technology leadership in Artificial Intelligence. IBM's response to the NIST RFI provides ten actionable recommendations that advance Trustworthy AI, Fairness, Explainability, Robustness, Transparency, Ethics and Responsibility, Privacy, Evaluations, Data sets, and Metrics. Given the rapid pace of technical progress being made in the AI field, NIST's work should be directed foremost at accelerating technical innovation by facilitating the development of a conceptual foundation for trustworthy AI along with tools, evaluations, data sets and metrics across these ten dimensions, which will be essential for establishing and furthering US leadership in AI.

References

1. "Summary of the 2018 White House Summit on Artificial Intelligence for American Industry", The White House, Office of Science and Technology Policy, May 10, 2018.
2. A. Narayanan, "21 Fairness Definitions and their Politics," *ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)*, February 2018, https://www.youtube.com/watch?v=jIXIuYdnyyk.

3. R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, "AI Fairness360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, arXiv:1810.01943, October 2018.

4. T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, L. Daniel, "Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach," *Intl. Conf. on Learning Representations (ICLR),* May 2018.

5. A. Boopathy, T.-W. Weng, P.-Y. Chen, S. Liu, L. Daniel, "CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks," *AAAI Conf. on Artificial Intelligence*, February 2019.

6. The IEEE Global Initiative, *IEEE Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition (EAD1e), March 2019.

7. B. R. Rowe, D. W. Wood, A. N. Link, D. A. Simoni, "Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program", *RTI Project Number 0211875*, July 2010. https://trec.nist.gov/pubs/2010.economic.impact.pdf

8. D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. van der Maaten, "Exploring the Limits of Weakly Supervised Pretraining," *European Conf. on Computer Vision (ECCV),* pp 185-201, October 2018.

9. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation," *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics.* pp. 311–318.