HITACHI
Inspire the Next

October 15, 2020

Dr. Walter G. Copan
Under Secretary of Commerce for Standards and Technology
Director, National Institute of Standards and Technology
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

RE: *Four Principles of Explainable Artificial Intelligence*

Dear Under Secretary Copan:

The following comments are submitted by Hitachi Group companies (Hitachi) doing business in the United States in connection with the National Institute of Standards and Technology (NIST) Draft *Four Principles of Explainable Artificial Intelligence* published on August 17, 2020, and released for public comment.

Background

Founded in 1910 and headquartered in Tokyo, Japan, Hitachi, Ltd. is a global technology corporation answering society's most pressing challenges through cutting-edge operational technology (OT), information technology (IT), and products/systems. A Social Innovation leader, Hitachi delivers advanced technology solutions in the mobility, human life, industry, energy, and IT sectors. The company's consolidated revenues for FY2019 (ended March 31, 2020) totaled $80.4 billion and 814 companies employ over 300,000 employees worldwide.

Since establishing a regional subsidiary in the United States in 1959, Hitachi has been a committed American partner. For over thirty years, it has invested heavily in research and development (R&D) in the U.S., and this continued reinvestment has resulted in 16 major R&D centers that support high-skilled jobs in manufacturing and technology. Dedicated to delivering the technologies of tomorrow, Hitachi opened a Center for Innovation in Santa Clara, California to explore applications in machine learning, artificial intelligence, Internet of Things (IoT) devices, data analytics, and autonomous vehicles among other advanced technologies. Hitachi is also proud of its human capital investment with over 20,000 employees across 75 companies in the U.S. At 13% of total revenue, North America is Hitachi, Ltd.'s second largest market, following only the Japanese market, with $10.1 billion in revenue in FY2019.

Hitachi commends NIST for its comprehensive federal engagement plan, and welcomes the opportunity to engage with the U.S. government as it sets standards that will determine America's standing as a global leader in artificial intelligence (AI).

Responses to NIST Draft Report

The 2019 NIST *U.S. Leadership In AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools* correctly cited "explainability" as one of a catalogue of characteristics related to trustworthy AI technologies. At that time, Hitachi called upon NIST to spearhead the creation of common, international definitions and standards around these characteristics. With the publication of this Draft, Hitachi once again thanks NIST for following the recommendations of the industry in working collaboratively in setting the foundation for the growth of AI as a technology and tool for economic growth and innovation.

From a basic logic standpoint, the "Four Principles" approach is acceptable. In particular, the use of humans as a comparison group for evaluating how well the Four Principles are met, including the perspectives from psychology and neuroscience, is beneficial. It is important to note that decisions by humans and by AI systems are different. One usually does not think of human experts as black boxes who need to prove their trustworthiness. For example, a doctor is trusted because of medical training and experience requirements for licensure. An AI system will have different explanation needs than a human person.

The Draft is written to allow understanding by a wide audience. While there are clear benefits to this approach, it can water down the overall standard and miss explaining that the practical applicability must be more rigid in its approach to reliably contribute to trustworthiness. It is clear NIST took much time and deliberation in creating the document, using the Four Principles as pieces to create a step-by-step definitional approach, spending time providing literature sources for varying methods, and breaking down some explainability models in existence. All of this is laudable, but NIST could help further instruct industry how this can be applied through examples or case studies.

With this attempt, the paper seems to emphasize explainability of the deployed AI system to its end users. Again, a valid pathway, but it could leave out the use of explainability as a means to identify errors or make improvements to the program or address concept drift. One use of explainability is for model development and debugging. If an AI system is thought of in a "lifecycle" context, comprising model development, implementation, and deployment, each stage has a need for an explanation to the user. The first users may include the data scientists who develop the model; for them, explainability is important for identifying errors, making improvements, etc. While the Draft appears to partly cover this in the "System development" example in Section 3, it would enhance the Draft to focus the needs at each stage of AI and further emphasizes explainability throughout the lifecycle of the AI system. It could be further enhanced by discussing how concept drift is addressed not only as part of the lifecycle evaluation, but also with regards to the original explainability standard's application to the AI system when the model adapts and changes over time.

Explaining the model base of an AI system on a few dominant features can result in explainability making sense. However, the Draft uses decision trees, linear and logistic regression as examples of simple and self-explain models. This is not necessarily the case as these models could involve hundreds of features and interdependencies. As an example, a linear model uses each coefficient as the change in the predicted value per unit change in the corresponding feature if all else remains equal. This may be an accurate description of how the model works, but may not be meaningful if the feature in question cannot be changed without also changing the values of other features. Data scientist are looking at the best accuracy possible from the model. That means in the end it might not be explainable and in fact the model could be solving a problem that is too complex for a meaningful explanation. NIST should consider exploring the possibility of allowing a model to be described after its accuracy is established in detail. The explanation could work backward from the outcome to make it explainable to the average person. This may be acceptable because there is a rough idea of how it works, and being realistic on explainability is important.

Explainability is perhaps one of the harder characteristics for standards development. As the Draft correctly recognizes in the "Meaningful" principle, different groups of users will require differing levels of explanation for the end output to fulfill each of the Four Principles laid out by NIST. This is further complicated by proprietary requirements for intellectual property (IP) protection. Providing the ability for a meaningful explanation to any group of users could demand the disclosure of proprietary algorithms, undercutting business competitiveness if it is revealed, or losing explainability certification from the NIST standard if the information is not provided. This is not clearly addressed in the Draft and should be clarified.

NIST's role in creating standards for use by the industry is vitally important. Equally important are the practical implications of the standards. Under the "Meaningful" principle, the Draft discusses broad groups of people whose differing perspectives may make what is "meaningful" to one group not "meaningful" to another. Using just one example, that of judges and jurors, NIST should think of explainability in regard to how court proceedings are handled. When jurors are selected for a court case, there is no burden on either defense or prosecution to conduct aptitude tests or other requirements to demonstrate expertise in the issue being heard. Instead, courts use expert witnesses who have established credentials, who offer testimony and are cross-examined by both. This provides judges and jurors information in an explainable manner that can then be assessed by the jurors/judges as part of their deliberations.

In this context, consider how an expert trained in machine learning pipeline design, development, deployment, maintenance and enhancements could offer an informed opinion on an AI system being evaluated. The expert can be examined, and the explainability need rise only to the level of the expert. Once evaluated, the AI system can meet the explainability standard based on the expert evaluation provided, removing the burden of selecting a panel of candidates with advanced statistics knowledge and aptitude who could evaluate the model. By contrast, if the model must be explainable and meaningful to the common man, then we risk significant detrimental impacts and burdens on innovation & model creation under this regulation.

The Draft, as one focused on explainability, does not speak to bias in the dataset. This is a severe problem for the overall trustworthiness debate of AI systems. If a standard for datasets is not established, explainability standards will not help further the confidence of AI tools. In Hitachi's 2019 comments, we urged NIST to ensure that there are strong standards developed around data as the critical component of AI systems. We called upon NIST to function as the lead source for testing methods that establish dataset accuracy, privacy, data leakage, reproducibility, provenance, and identify bias. We again urge NIST to start the trustworthiness standard-setting process by working to overcome bias and dataset accuracy prior to finalizing the explainability standards or to develop testing standards for datasets.

NIST, in addition to addressing bias in datasets first, might want to consider working on the "trustworthiness" characteristic before explainability. As we have noted, explainability can require different details for different audiences and might well result in IP disclosures that fundamentally hurt innovation and research into the topic by industry. A trustworthiness standard first, coupled with methods to test datasets and address bias, may result in a better understanding of AI systems.

As a final recommendation to NIST, Hitachi strongly asks NIST to provide additional drafts of this "Four Principles" document before making anything final. There are numerous topics not able to be covered in detail during the prescribed time period given to respond to this Draft. Meetings with Hitachi and others in the industry can lead to further dialogue and understanding of the numerous complex issues involved with explainability. Those discussions can strengthen the standards, and more importantly make them meaningful. When standards are meaningful and well-crafted, industry has more confidence in relying on them for design. More confidence will result in higher trustworthiness of the end result, thus promoting the adoption of AI across sectors. Rushing to finish this standard, especially without meaningful work on datasets and trustworthiness, could undermine NIST's ability to craft AI standards.

Conclusion

Hitachi appreciates NIST's vigorous effort to implement President Trump's February 11, 2019 Executive Order (EO 13859) on securing the country's leadership in AI. This Draft is a helpful step in that implementation, furthering the U.S. advancement in AI and working with industry to set standards for the

future innovation in this area. We look forward to our continued collaboration to assist the federal government as is works to develop internationally agreed-upon, consensus-based standards that promote trustworthiness and widespread AI adoption.

Sincerely,

Koji Takaichi
President & CEO
Hitachi America, Ltd

Hitachi America, Ltd.
2535 Augustine Dr. 2nd Floor, Santa Clara, CA 95054 Tell: 408.986.6300
www.hitachi.us