October 15, 2020

National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899-2000

Comments of Bank of America
Submitted to explainable-AI@nist.gov

Re: "Four Principles of Explainable Artificial Intelligence"

Thank you for the opportunity to offer our perspective on the National Institute of Standards and Technology draft document on Four Principles of Explainable Artificial Intelligence (NISTIR 8312). At Bank of America, we continue to be deeply committed to the topic of Artificial Intelligence (AI), and supportive of all efforts to promote responsible use of data, technology and AI. As part of that, Bank of America is the founding donor of the Harvard University Kennedy School of Government's Council on the Responsible Use of AI.

We agree that 'explainabilty' can be associated with trust in AI systems and their outcomes, and that laying out principles which help to further define that term are an important foundational element to support the responsible use of AI. We also appreciate your recognition of the inherent challenges in defining acceptable standards of Explainable AI, as each consumer of the AI output will come with differing backgrounds, objectives and expertise. Finally, as you raise in your conclusions, you recognize the potential for integration of AI and human capabilities to create better outcomes than either in isolation. We support each of those points and agree that further work can be done to deepen the dialogue in these areas.

We would recommend the following points be taken into consideration as you refine your work in this area:

(1) Reinforce the importance of end-to-end governance

In some ways, your four principles of Explainable AI (Explanation, Meaningful, Explanation Accuracy and Knowledge Limits) can be aligned to the objectives of model validation contained in SR 11-7, that is "Model validation … verif(ies) that models are performing as expected, in line with their design objectives and business uses. Effective validation helps ensure that models are sound. It also identifies potential limitations and assumptions, and assesses their possible impact." However, just as a sound model risk management framework, as per SR 11-7, requires not only "rigorous validation", but also "sound development, implementation and use of models" as well as "governance and control mechanisms", we continue to believe it is critical to focus attention on the need for a robust AI risk management framework that encompasses the AI lifecycle. [1]

Well-governed and transparent processes that control what gets built and for what use, as well as proactive identification, assessment and remediation of risk during the development stage, are equally contributory toward building trust in AI solutions as is the explainability of the output. Valid considerations around ethics and privacy, for example, may impede the acceptance of an AI solution; the governance steps described above are likely the most effective mechanism to mitigate those concerns and build trust.

(2) Clarify that "Explainable AI" does not imply "universal explanations"

There could be many ways to explain the relationship of inputs to outputs, not only one explanation.

Not only do all AI systems have knowledge limits that should constrain outputs (as described under "Four Principles of AI Knowledge and Limits"), it is also the case that there may be a large number of AI systems which could produce similar or better outcomes. This is important because Explainable AI is not just an explanation of an AI system itself - it is can be received as an explanation, though imperfect, of reality.

We recommend consideration be given to including this point, so that the end-users of the AI Explanation are clear that what is being described is how and why a specific AI system reached a certain output, given a set of inputs, but that this is not a universally-applicable explanation for the relationship between the inputs and output.

[1] Board of Governors of the Federal Reserve System and Office of the Controller of the Currency (2011). Supervisory Guidance on Model Risk Management (SR Letter 11-7).