October 15, 2020

Via Electronic Mail

The Honorable Walter G. Copan
NIST Director and Undersecretary of Commerce for Standards and Technology
National Institute of Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, MD 20899-2000

Re: Comments on the Four Principles of Explainable Artificial Intelligence (Draft NISTIR 8312)

Dear Dr. Copan:

The Bank Policy Institute[1], through its technology policy division known as "BITS", appreciates the opportunity to comment on the National Institute of Standards and Technology's (NIST) draft Four Principles of Explainable Artificial Intelligence (the "Principles").[2]

The financial services sector is strongly committed to promoting the responsible use of artificial intelligence (AI) and transparency in decision making, given the potential long-term benefits that AI may provide to consumers and the future of financial products. Banks currently utilize AI and machine learning in a wide variety of bank operations, including marketing, customer service, fraud detection and prevention, anti-money laundering, credit underwriting, back-office processing, pricing, and more. The application of AI systems within banks continues to evolve as we learn more about AI and the interactions between machines and humans within organizations and as well as externally with customers. We support NIST's efforts to help develop trustworthy AI systems and appreciate NIST's recognition of the challenges in defining principles over evolving technologies and applications. In any next steps, NIST should ensure that principles are flexible enough to encourage innovation and the continued adoption of AI across all industries in a responsible manner.

---

[1] The Bank Policy Institute is a nonpartisan public policy, research and advocacy group, representing the nation's leading banks and their customers. Our members include universal banks, regional banks and the major foreign banks doing business in the United States. Collectively, they employ almost 2 million Americans, make nearly half of the nation's small business loans, and are an engine for financial innovation and economic growth.

[2] National Institute of Standards and Technology, "Four Principles of Explainable Artificial Intelligence", Draft NISTIR 8312, (August 2020) *available at* https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8312-draft.pdf.

In Part I of this letter, BPI/BITS proposes some high-level comments about how risk management and governance processes are critical to enabling trust in AI technologies. Part II of this letter provides specific comments and considerations on each of the draft four principles of explainability, providing context from the financial sector. We look forward to continued opportunities to comment on principles developed by NIST on AI, especially considering the technology is rapidly evolving and there is much to learn about how AI affects society.

I. Risk Management and Governance is Key to Building Trust in AI Systems

The financial services sector is unique and stands out from other industries in that banks implement comprehensive risk management and corporate governance processes over AI systems, as required by Supervisory Guidance on Model Risk Management ("Model Risk Management Guidance" or the "Guidance").[3] The Guidance, issued jointly by the Office of the Comptroller of the Currency (OCC) and the Federal Reserve Board (FRB) and subsequently adopted by the Federal Deposit Insurance Corporation (FDIC), requires banks to develop effective model risk management frameworks, including robust model development, implementation, and use; effective validation; and sound governance, policies and controls.[4] Banks approach explainability of AI systems within the context of risk management frameworks by embedding explainability expectations and requirements for AI systems throughout the process. As a result, a bank's decisions about the requirements of any AI system are necessarily tied to the potential risks, outcomes, and impacts on both the bank and its customers.

We generally agree with the draft Principles that 'explainability' is a key aspect to enabling trust, understanding, and adoption of AI technologies. However, the Principles focus specifically on the output of an AI system rather than the overall risk management and governance of the process, which is equally as important towards building trust in AI solutions. Banks dedicate significant attention to assessing risk and building a governance framework around the entire AI lifecycle, from development of AI models, to implementation and use of models, to continued oversight of outputs. Explainability is one of several components within the model risk management framework and overall governance process to determine whether to use (or continue to use) an AI system. Additionally, it is important to understand that AI algorithms are constantly changing, creating challenging notions of explainability. For example, some AI models and algorithms adapt and retrain automatically over time, so it may be necessary to reevaluate explainability of the AI system. The environment in which the AI system operates is also constantly changing, requiring retraining of the same algorithm on new data to ensure the AI model is working the way it was originally intended. Within the model risk management framework, banks regularly reassess models over time as a part of ongoing monitoring to ensure models, including aspects of explainability, are working as intended.

---

[3] FRB, SR 11-7, Supervisory Guidance on Model Risk Management (Apr. 4, 2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf; OCC, Bulletin 2011-12, Supervisory Guidance on Model Risk Management (Apr. 4, 2011), https://occ.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf ; FDIC, FIL-22-2017, Adoption of Supervisory Guidance on Model Risk Management (June 7, 2017), https://www.fdic.gov/news/financial-institution-letters/2017/fil17022.pdf.

[4] The current Model Risk Management Guidance does not mention AI-based models at all. However, at least one member of the FRB has stated that the Guidance should apply to banks' use of AI systems. See Governor Lael Brainard, What Are We Learning about Artificial Intelligence in Financial Services?, Speech at Fintech and the New Financial Landscape, Philadelphia, Pennsylvania (Nov. 13, 2018), https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm.

The overall risk management process implemented by banks is not only central to complying with regulatory requirements, but also to developing confidence internally within the organization to ensure AI systems are used responsibly. In finalizing any principles, we recommend that NIST consider the extent to which AI risk management and governance can and should be used to build trust across all industries, especially those providing financial services and other highly regulated services.

II. Comments on the Four Principles of Explainable AI

The four principles of explainable AI – explanation, meaningful, explanation accuracy, and knowledge limits – provide a valuable starting point for how we think about explainability of the outputs of AI systems. We encourage NIST to consider how the draft principles are structured so that each principle may be meaningful on its own, while together the principles form a solution for AI explainability. As currently outlined, having *Explanation* as the first principle of 'explainability' is somewhat confusing. It is also important to highlight that the concept of AI explainability will always accompany AI interpretation. In other words, any framework of explainable AI is aimed towards achieving interpretations of the system's behavior from the perspective of users, developers, and owners. However, AI systems maintain some level of uncertainty. To that end, it is essential to understand how each of these principles may work in practice otherwise they may be misleading. Further, in evaluating these principles of explainability, we must bear in mind that humans are involved in all aspects of decisions and provide the final determinations of whether to use an AI system or the outputs of an AI system. We offer the following comments and recommendations on each of the four principles.

A. Explanation

According to the draft, the *Explanation* principle obligates AI systems to supply evidence, support, or reasoning for each output. We believe that an explanation should be commensurate with the level of risk in the AI system; the Principles should not necessarily obligate an explanation of all AI systems. The notion behind explainability and AI is often misunderstood – just because an AI system is unable to explain itself does not necessarily mean the system is bad or should not be used. As FRB Governor Lael Brainard noted in November 2018, "There are likely to be circumstances when using an AI tool is beneficial, even though it may be unexplainable or opaque" and in these instances, "the AI tool should be subject to appropriate controls, as with any other tool or process, including how the AI tool is used in practice and not just how it is built."[5] There should be different standards of explanation based on the context in which the AI system is being used and the recipient of the explanation, which banks evaluate and determine within the risk management framework. For example, banks may hold AI systems that are customer facing, such as those used for consumer lending, to a higher standard of explanation than AI systems that are used to identify potential fraud or for authentication purposes,

---

[5] See Governor Lael Brainard, What Are We Learning about Artificial Intelligence in Financial Services?, Speech at Fintech and the New Financial Landscape, Philadelphia, Pennsylvania (Nov. 13, 2018), https://www.federalreserve.gov/newsevents/speech/brainard20181113a.htm. See also Randal Quarles, 2018 Financial Markets Conference, A Conversation on Machine Learning in Financial Regulation ("The SEC after the flash crash, for example, and after earlier crashes, established systems of circuit breakers. Instead of saying, 'We need to understand exactly how these trading tools will operate in every circumstance and why, and be able to predict how they're going to operate in every circumstance,' if they move outside of certain ranges, then we're going to stop for a moment so we can figure out what's going on.")

based on an evaluation of risk. In these determinations, a human is always involved in determining whether and how to use the output from an AI system, and if overlays or adjustments are necessary.

The draft Principles note that, "By itself, this principle [Explanation] does not require that the evidence be correct, informative, or intelligible; it merely states that a system is capable of providing an explanation." We understand that this principle is meant to be taken in context with the *Meaningful* and *Explanation Accuracy* principles, but as written may be interpreted that it is acceptable to provide an incorrect explanation. We suggest revising this language, as the acceptance of incorrect explanations may negatively affect end users and society's willingness to trust AI systems.

We agree that one-size-fits-all explanations do not exist, and different users will require different types of explanations. However, we interpreted the types of explanations outlined in the draft Principles as drivers of explanations, or reasons for why you would need an explanation, rather than an explanation itself that differs based on the audience. Additional detail on how an explanation differs based on the audience or end user may help clarify the need for different types of explanations.

B. Meaningful

Regarding the *Meaningful* principle, we agree with the assessment that two individuals viewing the same AI system's output may not interpret the meaning in the same way, but the draft implies that the principle is fulfilled if the user can understand the explanation. This principle poses practical challenges from an oversight perspective, in terms of who determines what is understandable or meaningful. For example, the principle recognizes the potential differences in meaningfulness by user (i.e., developer versus end user), but does not go the next step to determine how this works in practice if the AI system is meaningful to one user (developer with AI subject matter expertise) and not the other (end user), or how to evaluate changes in meaningfulness over time.

Within the financial services sector, we believe meaningfulness of explanations is best assessed across groups of users, rather than at the individual level. Tailoring the meaningfulness of explanations at the individual level is difficult, if not impossible, to accomplish for many AI systems. While we agree that individuals interpret what is meaningful differently, any such requirement or expectation to demonstrate that an AI system produces a meaningful explanation at the individual level would be challenging to accomplish.

C. Explanation Accuracy

Similar to the *Meaningful* principle, the *Explanation Accuracy* principle presents challenges in practice. From an oversight perspective, it is difficult to determine what level of accuracy of the explanation is sufficient until there is a common way to measure explanation accuracy, which the draft notes is still under development. Further, NIST should consider whether striving for 100% accuracy of the explanation is a worthwhile goal within the realm of explainability. As referenced in the comparison to human decision-making section of the draft, humans often do not meet the definition of the *Explanation Accuracy* principle, leading one to question whether it is more appropriate for machines to be held to a different standard that is equal to or higher than humans, rather than a standard of 100% accuracy. As with other elements of explainability, banks approach the *Explanation Accuracy* principle within the context of the risk management framework. In determining whether to use an AI model for a specific use case, the accuracy of the explanation is one of many factors that influence the decision.

There may be instances where the accuracy of an explanation is insignificant to the output of the AI system, or where one may be comfortable with less accuracy of the explanation in certain use cases.

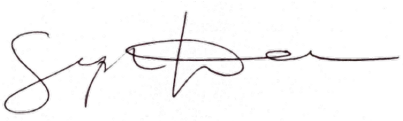
D. Knowledge Limits

We agree with the intent of the *Knowledge Limits* principle and the importance of preventing misleading, dangerous, or unjust decisions or outputs. Within the financial services sector, assessing whether models are operating within the appropriate real-world environment is a key aspect of model risk management. For example, banks ensure that a specific material interest rate model is not used within a negative rate environment if it has not been modeled in that environment. The relationship between humans and AI systems is particularly important within the *Knowledge Limits* principle as AI systems continue to evolve, human oversight can help to ensure that AI systems and explanations are operating within the appropriate environment.

* * * * *

We appreciate NIST's efforts to develop the draft Principles of Explainable AI and recognize the inherent challenges in defining acceptable standards of AI. Thank you for the opportunity to comment on the draft Principles. If you have any questions, please contact the undersigned by phone at 202-589-2432 or by email at *Stephanie.Wake@bpi.com.*

Respectfully submitted,



Stephanie Wake
Vice President, BITS
Bank Policy Institute