Anthony Mader
Vice President, Public Policy
Anthem, Inc.
1121 L Street
Sacramento, CA 95814
(916) 403-0522

Submitted via email: explainable-AI@nist.gov

October 15, 2020

National Institute of Standards and Technology
44 100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

Re: Draft National Institute of Standards and Technology (NIST) "Four Principles of Explainable Artificial Intelligence" guidance (NISTIR 8312)

Dear NIST:

Anthem, Inc. (Anthem) appreciates the opportunity to provide comments on the Draft National Institute of Standards and Technology (NIST) "Four Principles of Explainable Artificial Intelligence" guidance (NISTIR 8312). Anthem's response draws on our significant experience in working to improve healthcare outcomes and lower total costs for the more than 106 million people we serve, including more than 42 million within our family of health plans.

We commend NIST for drafting this informative source of information to help create greater understanding of the current explainable Artificial Intelligence (AI) landscape. We echo and align with the comments submitted by the U.S. Chamber of Commerce (the Chamber) and support NIST's general approach taken in the draft publication. Anthem also wishes to add recommendations and clarifications.

General Comments on Explainable AI

Anthem supports NIST's work in outlining a set of foundational principles and guidelines for AI technology. In alignment with the Chamber, we have a few recommendations and considerations for NIST to consider across the principles:

First, we appreciate and acknowledge NIST's statement in lines 133-134 that explainable AI is one of several properties that build trust in AI systems. In connection to the additional principles mentioned (resiliency, reliability, bias, and accountability), Anthem asks NIST to also consider the relationship of transparency to explainability, as well as how transparency connects into the other properties of AI trustworthiness. Situation pending, transparency may include recipient awareness that AI was used to determine an outcome, as well as clarity on how to request information and corrections. We ask NIST to consider existing principles that discuss the intersection of AI principles, such as the Organisation for Economic Co-operation and Development (OECD) Principles on Transparency and Explainability. We also recommend NIST expand their guidance to demonstrate alignment with existing laws and regulations.

Additionally, in lines 159-163, NIST first defines the output of an AI system as the result of a query of that AI system. We support the Chamber's request, and seek further clarity from NIST on how outputs are defined in relationship to the principles, and how NIST views outputs and the explainable AI principles in regards to integrated systems that combine to form a decision, or set of decisions. Technological advances frequently involve combining existing technology, such as AI systems, in new ways. Requiring extensive explainability during all stages and for all potential queries could stifle emerging applications and inhibit smaller developers.

Four Principles of Explainable AI

Anthem broadly supports the four principles of explainable AI proposed, and aligns with the Chamber in our appreciation of NISTs literature review and analysis. The principles are useful for early discussions on AI standards development, and Anthem looks forward to continued and ongoing collaborations between public and private sector stakeholders to further develop related framework concepts. Overall, we emphasize that open discourse and shared standards established between government and industry are fundamental to ensuring the adoption and successful implementation of technologies and related services.

- **Explanation**
    - While we support this principle in a larger context of explainable AI, we ask NIST to consider the vast range of types of AI risk-based approaches when determining applicability of this principle. At a general level, AI tools and services range in complexity, and requiring explanations across all AI outputs may be onerous or even unnecessary, where the risk of stifling innovation could exceed the benefits of explainability. In addition, the ability of AI to offer useful explanations also relies on data and developer inputs, and NIST should further consider these interactions.
- **Meaningful**
    - Under section 2.2 of the guidance document, NIST encourages explanations tailored to the user groups, which would require that a recipient understands the explanation provided. We agree with NIST's statement that the recipients receiving explanations can vary greatly, and "no one size fits all." We do ask NIST to consider applicability and variance across AI use cases. Anthem recommends providing more details, perhaps by first considering specific recipients and use cases to better understand how to implement this principle.
- **Explanation Accuracy**
    - In alignment with the Chamber, Anthem also agrees with NIST's guidance that the explanation accuracy is distinct and separate from decision accuracy. In lines 216-217, NIST states that researchers are developing performance metrics for explanation accuracy. We ask NIST to provide further clarity on potential metrics under consideration. Anthem also encourages continued stakeholder collaboration on performance metrics, as AI and Machine Learning (ML) used in different industries will have various nuances, existing laws, and data impacts to consider.

- **Knowledge Limits**
    - We support including this principle, and agree with NIST that systems are not designed, approved to operate, or able to provide reliable answers for all cases. AI is trained on specific data sets and to answer specific questions, and acknowledging limits is important for transparency.

Types of Explanations

Anthem agrees that explanations will vary, and appreciates NIST including a non-exhaustive list gathered from the literature review. We encourage NIST to continue work with stakeholders on use cases, and applying risk-based approaches when determining applicability of the explainability principles and types of explanations necessary.

Humans as a Benchmark Comparison Group

NIST's research and inclusion of human decision explainability is valued and appreciated as a benchmark. A tendency exists for people to expect human error, but expect perfection from technology. It is important to consider that AI is trained from human input and human-collected data with its own historical errors, including population misrepresentation. It's also important to consider that AI is often used a tool to aid human decisions. Thus, further discussions are needed on the explainability principles and how and when they apply to decisions made from AI-human interactions.

Additional Considerations

We see a future where models grow increasingly complex in support of their performance, accuracy, and scalability. The AI and ML community may also face the trade-offs of accuracy and performance over the ability to provide explanations, and continued conversations are necessary to determine trade-off thresholds.

Anthem agrees with NIST and the larger community that standards for understanding AI are necessary, especially for decisions impacting human outcomes. We encourage NIST to continue their research and focus on guidelines that inform AI actors about the models themselves, including how the models are developed and tested in retrospective studies. This could include minimum information guidelines and documentation standards that can serve those using AI tools[1], which has shown downstream improvements in some current use cases. Similar to NIST's current workshops and guidelines documents, we look forward to collaborative stakeholder work to explore minimum information guideline opportunities.

[1] Norgeot, B., Quer, G., Beaulieu-Jones, B.K. et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 26, 1320–1324 (2020). https://doi.org/10.1038/s41591-020-1041-y.

Furthermore, we ask NIST to consider explainability impacts on information security. More specifically:

- How should the AI system provide auditable artifacts to explain the results of the model?
- How do we protect the artifacts or logs that explain the AI model? For example, what are the explanation retention periods? In the case of sensitive data, how do we ensure the explanation is only distributed to authorized personnel?
- AI models should include traceability and auditability to understand how the model changes over time. How does NIST apply the explainability principles to drift, and what's the oversight or expectation for continually learning models?

***

Anthem, Inc. is appreciative of the NIST draft guidelines, and for the opportunity to provide feedback. We welcome the opportunity to discuss our recommendations and the clarifications requested. Should you have any questions or wish to discuss our comments further, please contact Stephanie Fiore at (667) 209-1355, or Stephanie.Fiore@anthem.com.

Sincerely,

Anthony Mader
Vice President, Public Policy

Anthem is a leading health benefits company dedicated to improving lives and communities, and making healthcare simpler. Through its affiliated companies, Anthem serves more than 106 million people, including more than 42 million within its family of health plans. We aim to be the most innovative, valuable and inclusive partner. For more information, please visit www.antheminc.com or follow @AnthemInc on Twitter.