CHAMBER OF COMMERCE
OF THE
UNITED STATES OF AMERICA

NEIL L. BRADLEY
EXECUTIVE VICE PRESIDENT &
CHIEF POLICY OFFICER

1615 H STREET, NW
WASHINGTON, DC 20062
(202) 463-5310

October 15, 2020

National Institute of Standards and Technology
U.S. Department of Commerce
100 Bureau Drive, Stop 2000
Gaithersburg, MD 20899

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

To Whom It May Concern:

The U.S. Chamber of Commerce ("the Chamber") appreciates the opportunity to submit feedback to the National Institute of Standards and Technology's ("NIST") draft publication on "Four Principles of Explainable Artificial Intelligence ("AI")." The Chamber appreciates NIST's effort to advance standards for AI applications and convene diverse sets of stakeholders to ensure U.S. leadership in trustworthy AI.

The draft publication constitutes a thoughtful and accurate survey of the current explainable AI landscape. We believe, however, some areas could benefit from further discussion and clarification. The draft, and NIST's broader program to develop a much-needed comprehensive AI framework, should significantly contribute to the private and public sector's understanding of the many considerations necessary to implement AI, while also ultimately enabling broader, faster, and more responsible use of AI.

General Comments on Explainable AI

The Chamber agrees with the draft publication that explainable AI – in addition to other properties such as resiliency, robustness, and accountability – is an important property of trustworthy AI. As a general matter, we support the approach taken in this draft publication but would like to make several recommendations and seek clarification on several issues.

First, NIST's introduction places the onus to achieve societal acceptance and trust on the developers of AI systems. Yet, developers are not the only, or even primary, actors that bear that responsibility. The introduction should be a call to action to all actors involved in the AI ecosystem to recognize the importance of trustworthy AI, and that societal acceptance and trust in AI systems should be part of broader strategic governance and strategy discussions.

Second, on line 120, the draft uses the phrase "high stakes decision processes" to describe certain types of AI applications. The Chamber believes that this should be clarified in the introduction as it may present both novel and significant regulatory implications. Moreover, some jurisdictions have proposed to differentiate AI applications based on "high risk" and "low risk", an overly simplistic approach to governance that may create challenges to more nuanced risk-based approaches. Furthermore, it is critical that NIST understand and communicate that explainability, like other concepts in the AI space, be evaluated and determined based on context and level of risk associated with certain AI systems.

Third, the Chamber agrees the four principles outlined in the draft should not be viewed in isolation. There are already a robust set of existing laws and regulations that may intersect with these four principles, which the draft correctly acknowledges through references to the European Union's General Data Protection Regulation and the Fair Credit Reporting Act. Some laws already require explanations for decisions; those laws should still govern explainability in their respective jurisdictions, regardless of whether AI is used. The draft should continue to recognize that AI has numerous applications and crosses a wide range of sectors. Consequently, the draft should highlight additional specific use cases and note that AI applications have varying degrees of risk depending on the context in which they are deployed and the stakeholders involved. As such, the application of these principles in practice should be risk-based, use-case specific, performance-based, and consider existing laws. Furthermore, explainable AI is also not a self-contained concept; it overlaps with other pillars of trustworthy AI noted in the draft, in particular transparency and accountability.

Fourth, the draft briefly discusses AI outputs. The Chamber believes that further clarification on some of the following questions would be helpful. For example, how are AI outputs defined in relation to the principles, and how does the draft apply to integrated systems that combine to form a decision, or set of decisions? Technological advances frequently involve combining existing technology in new ways, and requiring extensive explainability during all stages could stifle emerging applications and inhibit smaller developers. Additionally, certain outputs may be intelligible and relevant to some explainees (i.e., developers) but not to others, such as users or regulators. Clarifications regarding degree, context, and relevance of outputs would be helpful in to this discussion.

Fifth, the Chamber welcomes additional information on the next steps after the comment period of this draft publication. Specifically, will there be greater focus on each principle, the types of explanations, or a research priority plan derived from this process?

Sixth and finally, it is essential to recognize that explainable AI is a nascent field of research. The Chamber recommends that NIST recognize this issue and retain flexibility moving forward to accommodate advances in this field.

Four Principles of Explainable AI

NIST proposes four principles of explainable AI systems: Explanation, Meaningful, Explanation Accuracy, and Knowledge Limits. The Chamber broadly supports these four principles and appreciates NIST's detailed literature review and thoughtful analysis. Our comments below outline our feedback on each principle.

Explanation

The Chamber generally supports the Explanation principle and believes that it forms the basis of establishing explainable AI. However, we believe that the word "obligates" on line 173 is too strong, and is not applicable in all circumstances. As mentioned above, not all AI systems necessitate an explanation and some AI systems cannot always deliver an explanation. Also, in many circumstances, market mechanisms can determine whether an explanation is required, such

as if it would help differentiate a product from a competitor or is a key part of a developer's branding. Of course, in circumstances where an AI system would significantly impact an individual life circumstance, such as employment decision, "obligates" may be appropriate. Moreover, some existing legal requirements also require an explanation. The Chamber recommends that this principle be modified to add more context in recognition that an explanation would not be needed for a wide range of AI applications.

Meaningful

The Meaningful principle correctly outlines that explanations need to be understandable to the user. Nonetheless, the principle does raise many questions. For example, who determines what is meaningful, what exactly is plain language, and how does the principle intersect with existing regulatory requirements? NIST should seek to further clarify these questions considering a wide range of users and applications would be implicated by this principle.

Additionally, the draft publication recommends that the principle "allows for tailored explanations at the level of the individual." While the Chamber agrees with NIST that a one-size-fits-all approach is inappropriate, tailoring to this extent may not always be necessary for an explanation to be meaningful to a user. Also, it is difficult for developers and operators of AI systems to predict and measure an individual's particular experiences and prior knowledge to tailor an explanation to the degree sought by this principle. Tailoring based on user groups (e.g. customer, regulator, owner, etc.) may be more appropriate in certain circumstances. The Chamber also recommends that NIST conduct additional work to enable entities to characterize and measure meaningfulness as well as considering variance in specific cases to prevent excessive burdens on actors in the AI ecosystem.

Explanation Accuracy

The Chamber agrees that the concept of Explanation Accuracy should relate to the accuracy or robustness of the explanation rather than the accuracy of the decision, and that those explanation accuracy metrics would likely differ among groups and users. The Chamber has two suggestions to improve this subsection. First, simply because an AI system may generate more than one type of explanation does not necessarily increase the explainability of that system. Not all AI systems require multiple types of explanations, or in some cases any explanation at all; the degree of explainability will depend on the accuracy of the explanation. Second, we recommend that NIST consider mechanisms, including performance standards, for assessing what types of metrics may be appropriate to help determine the accuracy of a given explanation in a given context in this subsection and engage with stakeholders regarding mechanisms related to this principle.

Knowledge Limits

The Chamber generally supports including this principle as it recognizes that AI systems are naturally limited in their capabilities to respond to questions based on the specific datasets upon which they are trained. The Chamber suggests one modification. On line 235, we suggest adding a third reason for low confidence related to the quality and size of the trained model:

when the algorithm is not provided a large enough sample and/or with necessary features to be able to learn to differentiate between classes from the training set.

Types of Explanations

The Chamber broadly agrees that there are many different types of explanations, and that a one-size-fits-all explanation would be inappropriate given the diverse number of AI applications and user types. While we generally support the five categories outlined in the draft, it may be more accurate to characterize these categories as reasons for an explanation rather than a type of explanation. Consequently, The Chamber views the "societal acceptance" explanation as duplicative of the other four explanations considering explainability is intended to build trust and acceptance of AI systems across society. This is particularly true for the "user benefit" and "regulatory and compliance" categories. Furthermore, the "societal acceptance" category seems ill-defined and without further clarification it could encompass explanations that objectives unrelated to trustworthiness or existing legal and regulatory requirements. The Chamber recommends that future publications strike this category entirely or further refine the category to address the concerns raised above.

In addition, the Chamber recommends that this section include a reference to counterfactual explanations, which can often be more actionable and beneficial than other explanations. Counterfactual explanations make human-machine collaboration possible even if the AI system was not designed to explain its decision-making process. This would enable a user to understand what changes would be required to adjust the output of the system.

Finally, NIST should clarify that not every AI application can or should provide all the different categories of explanations, to prevent imposing a significant burden on innovators. The applicable category or categories should depend on the particular AI application and other relevant factors such as existing legal requirements.

Humans as a Comparison Group for Explainable AI

In the draft publication, NIST examines the explainability of human decision-making to serve as benchmark metrics for explainable AI systems. The Chamber believes it is generally appropriate to use humans as a comparison group considering AI systems can often replace or supplement human decision-making in certain contexts. When examining human decision-making, NIST concludes that humans insufficiently meet the principles outlined in the draft. The implication of this conclusion is that if humans cannot meet the principles, then an AI system that is designed and operated by humans may also have similar challenges in meeting the principles. The Chamber is concerned that this would set an excessively high threshold for AI systems to measure performance and raises the question as to the ultimate objective of benchmark metrics established under this framework. Clarity from NIST would be helpful to ascertain if future benchmark metrics are intended to establish heightened scrutiny for all AI systems or for alternative reasons. Also, clarity is welcome on any systems or processes that do meet the draft's explainability principles.

Further Considerations and Recommendations

The Chamber is appreciative of the opportunity to provide feedback to NIST on the draft publication and for consideration of our questions and concerns. In addition to the feedback provided above, the Chamber would like to provide two additional considerations for NIST as work continues to advance AI explainability and trustworthiness.

First, the draft publication discusses other properties that characterize trust in AI systems – bias, resiliency, accountability, and reliability. However, transparency is not included in that list, but is a necessary component to build trust in AI. For instance, in order for a consumer to be aware of the option to ask for a meaningful explanation, an important first step is disclosing that AI was used to determine a decision impacting that consumer. NIST should further consider the relationship of transparency to explainability, the context of transparency as it relates to AI, and how transparency fits into the other properties of trustworthiness.

Second, the Chamber recommends that further discussions are necessary to consider information security questions. Some of these questions include:

1. How should the AI system provide auditable artifacts to explain the results of the model?

2. How do we protect the artifacts/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?

3. How does NIST apply the explainability principles to drift, and what is the oversight on continually learning models?

Third, the draft publication should engage in more of a balancing of interests, which juxtaposes the interest in an explanation with the need to protect other interests such as security or intellectual property rights. For example, information disclosed to consumers affected by an AI application should not increase the application's vulnerability to cyberattacks, nor should it require companies to reveal proprietary information. Instead, NIST should consider how explainability may be satisfied in certain use cases, such as disclosing to a limited number of experts or regulators rather than to a user or the public at large.

Conclusion

NIST has a critical role in convening stakeholders to lay the foundation for AI-related standards, and the draft publication on explainable AI is a step in the right direction. The Chamber strongly supports NIST's efforts in this regard and again appreciates the opportunity to submit comments on this draft publication. We look forward to collaborating with NIST on the next steps for this publication and on AI-related activities in the future.

Sincerely,

Neil L. Bradley