USCIB UNITED STATES COUNCIL FOR INTERNATIONAL BUSINESS

October 15, 2020

Dr. Walter Copan
Director and Undersecretary of Commerce for Standards and Technology
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

Dear Dr. Copan:

The U.S. Council for International Business (USCIB) is pleased to respond to NIST's call for public comments on its proposed Four Principles of Explainable Artificial Intelligence. USCIB is a trade association composed of more than 300 multinational companies, law firms, and business associations, which includes a broad cross-section of the leading global companies in the information and communications technology (ICT) sector.

Overall, USCIB members are encouraged by NIST's effort, which represents a positive first step in terms of grappling with the issue of AI explainability. The draft document constitutes a thoughtful and accurate survey of the current XAI landscape. The draft, and NIST's broader program to develop approaches to AI trustworthiness, 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. We believe that, to be most effective, humans and machines should collaborate, combining their respective strengths to provide sustainable value for consumers, businesses, governments, and society.

USCIB offers comments below that draw from industry experience as well as from USCIB's direct input to the development of the OECD's AI Principles through USCIB's affiliation with Business at OECD (BIAC).

## Comments on NIST Four Principles of Explainable Artificial Intelligence
## October 15, 2020

Explainable-AI@nist.gov

Beginning of a Consultative Process – USCIB preferred not to utilize the Excel comment form provided in the Call for Comments notice. Rather, USCIB welcomes this opportunity to begin what we hope will be a continuing dialogue with NIST as part of an iterative process on Explainable AI. Ongoing discussions, in turn, will enable all parties to account for the dynamic evolution of AI technologies and the related accumulation of business and government expertise in understanding the potential of AI and appropriate policy and regulatory strategies. In our view, such experience will wisely inform the eventual development of NIST's AI Explainability Principles.

More Focus on an Enabling Environment for Innovation – In particular, we would like to see the NIST process include discussion of critical elements of an enabling environment for investment in AI innovation perhaps in the introduction to the Principles. Components of this environment would include inherent flexibility. In the current early stages of development and deployment of AI systems, the elements of the NIST draft should not be adopted as a one-size-fits-all solution for the very diverse range of AI systems and applications.

Other important elements include intellectual property (IP) protections and attention to the privacy and security implications of AI applications. Additionally, it is critical that the "Principles" as articulated in the NIST draft are not used for metrics for testing and certification of AI systems, but rather used as a starting point for an iterative dialogue with industry, academia, civil society, and other impacted groups.

Importance of OECD AI Principles as Foundation – As stated above, USCIB members from industry leaders such as Facebook, Google, IBM, and Microsoft actively participated in an Experts Group that developed the OECD's AI Principles. The U.S. government also contributed actively to the development of the OECD principles, under the leadership of the State Department. On May 22, 2019, the OECD's 36 member countries, along with Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania, endorsed the OECD Council Recommendation on Artificial Intelligence. The principles contained in the recommendation were subsequently endorsed by the G20.

The OECD Recommendation effectively packaged five complementary values-based principles[1] for the responsible stewardship of trustworthy AI as well as provided five recommendations[2] for governments consistent with the values-based principles. While they are not legally binding, the OECD AI Principles are highly influential as evidenced by the fact that several OECD non-members joined in endorsing them.

USCIB regards the OECD Principles as establishing a solid foundation and an approach to AI that is practical and flexible enough to stand the test of time in a rapidly evolving field. Work on the principles continues at the OECD: its network of experts has formed working groups, one of which will develop guidance for AI actors and decision-makers seeking to implement policies for trustworthy AI, including AI that is transparent and explainable. USCIB encourages NIST to engage in a dialogue with the State Department in order to both monitor and contribute to developments taking place at the OECD. By ensuring that NIST AI Explainability Principles and guidance

[1]

1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
2. AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
4. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
5. Organizations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

[2]

1. Facilitate public and private investment in research & development to spur innovation in trustworthy AI.
2. Foster accessible AI ecosystems with digital infrastructure and technologies and mechanisms to share data and knowledge.
3. Ensure a policy environment that will open the way to deployment of trustworthy AI systems.
4. Empower people with the skills for AI and support workers for a fair transition.
5. Cooperate across borders and sectors to progress on responsible stewardship of trustworthy AI.

regarding AI align with international standards, NIST will further the goal of interoperable and harmonized standards for AI, so that AI systems may be developed once and deployed in multiple jurisdictions.

Attributes vs. Principles – We were pleased to see that NIST's proposed four principles of explainable AI – explanation, meaningful, explanation accuracy, and knowledge limits –touch upon many of the same concepts addressed in the OECD's principles. Taken together, the proposed four NIST Explainability Principles address the need for transparency, accountability and responsible disclosure, which are cornerstones of the OECD Principles.

In their current form, however, we feel the proposed NIST principles represent aspirational attributes rather than concrete principles. While the OECD Principles serve as a conceptual foundation, we believe that NIST is uniquely positioned to take forward in concrete ways what specifically could be done in the area of AI explainability to "spur innovation in trustworthy AI" and "foster accessible AI ecosystems" and, in a technical sense, zero in on how this should be reflected in an AI system.

USCIB members feel that more discussion between business and government is warranted to determine how the attributes apply to "real" AI systems. Moreover, if NIST expects to further revise and update this report over the years as it gains additional experience, there might be additional attributes that may need to be added to help with explainability. We hope that NIST will be committed to an evolving and iterative process.

The sources referenced in the NIST principles do a good job of explaining the term "trustworthy." For the purposes of this document, since a common understanding of trustworthiness is foundational to the principles, we would recommend including some elaboration on the definition of this term in this document. Additionally, we would suggest considering broadening the examples referenced throughout the document. For examples, some examples used to explain the term "output" seem to focus on responses to queries. It might be useful to provide more complex examples of output, such as guidance given for guided surgery or decisions made by autonomous driving systems.

Additionally, "outputs" may be too simplistic a concept for assessing explanations provided by many AI-enabled and autonomous systems. It may be necessary to include outputs, decisions, and actions in explanations. These additional terms of the explanation would be especially useful when assessing multiple systems acting in concert as opposed to one system or subsystem in isolation.

Alignment with OECD Principles -- The OECD Principles speak to transparency and explainability in the following terms:

> AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
> i. to foster a general understanding of AI systems,
>
> ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,
>
> iii. to enable those affected by an AI system to understand the outcome, and,

> iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

The OECD Principles reiterate the importance of context and consistency with the state of the art on several occasions when recommending actions for AI actors to take. This flexible, context-dependent approach to explainability appears absent from the draft NIST report. USCIB would encourage NIST to consider whether the principles regarding explainability may apply in different ways to different AI systems depending on the types of decisions they produce.

For example, the NIST report proposes that at least five categories of explanations might be given and that explanations may need to be tailored to the level of the individual. USCIB submits that it is not necessary, feasible, or cost-effective to build this capacity for explanation into every AI system.

On line 173 of NIST's document, in particular, the word "obligates" may be too strong. AI systems cannot always deliver explanations, and some explanations are useful only under certain circumstances. In relatively trivial AI decisions, such as a song recommendation on a streaming app, an explanation may be unnecessary. In this example, a general explanation for how the algorithm works could be beneficial for the broader public but explanations about specific song recommendations to an individual should not be mandatory because the recommendation has an insignificant impact on individuals and does not affect them financially, politically, or in terms of their personal well-being.

In some cases, business leaders will have some incentive to use more robust explanations as a differentiator against their competitors. In areas with significant impacts on human life, like medical diagnoses or mortgage applications, it becomes increasingly vital that entities can explain why AI systems reached a specific decision. NIST should add more context around when AI systems may or may not need to be explainable, and when explanations are critical. A more flexible approach would help stakeholders to better understand the types of explanations that different AI systems are expected to produce.

The OECD Principles also put forth a simple and straightforward concept of "explainability" which is defined by its objectives. This is relevant for the draft NIST report, because the report does not contain a clear definition of "explainability." The statement that "different users may take different meanings from identical AI explanations" also creates uncertainty as to what makes an explanation meaningful. The NIST report properly considers that there are different target groups for an explanation (e.g. users and regulators), but its concept of explainability would be clearer if some illustrative examples were provided.

In general, the principles of meaningfulness, explanation accuracy, and knowledge limits need refinement. For example, meaningfulness seems, in the definition, to be restricted to individual users, but is also discussed in terms of explanations being tailored to user groups. Explanation accuracy is audience dependent, but it may also be time dependent. Additionally, although training domain and confidence levels are important facets of knowledge limits, system states may play a role as well. For example, the level of knowledge a system has may be sufficient to make a good decision, but time may not allow the system to take the best action. More use cases with examples of dynamic AI behavior are likely needed to fully flesh out what might be another aspect of explanation accuracy.

Acknowledgment of Risks – Related to consideration of IP, privacy, and security concerns, the NIST process should consider attendant risks in realizing "explainable AI" in addition to, as it already does, articulating the desirable attributes of explainable AI. USCIB members share the goal of increasing transparency and responsible disclosure around AI systems to build public trust and understanding of AI-based outcomes and allow them to be challenged. At the same time, these interests must be balanced with risks related to IP protections and the privacy and security implications of AI applications.

Businesses should not be expected to disclose sensitive, proprietary information about their algorithms or systems in all instances, particularly if these systems do not have a significant impact on individuals and their well-being. As mentioned above, it would be disproportionate to create an expectation that every content recommendation engine will disclose its unique formula. Similarly, a standard that encourages businesses to disclose information about security features that make systems more vulnerable to cyberattack in the name of enhancing safety would be inappropriate. In these cases, the balancing of interests may require a simpler explanation or one that is targeted to a limited number of experts and/or is triggered in certain situations.

Regulatory and Compliance - It would be instructive to consider how complex systems are currently certified for safety, with the focus being more on an explanation of intent in the context of system performance. This approach might involve explanations related to avoiding risks rather than achieving particular goals. This approach is helpful when assessing system outputs, decisions, and actions that do not trace directly to a single, or even a small number of inputs.

Assumptions –NIST should consider whether explanations should include assumptions made by methods or models under certain circumstances. Examples of assumptions include:

- Distribution of the data, e.g. the model assumes data are normal.
- Is the model deterministic?
- If it is not, it means different runs can generate different results even when input is the same. The result could be hard to reproduce too.
- Is the model adaptive and robust?
- If the input has a slight perturbation, the result would be drastically different. This is crucial in order to determine how often we would need to update the model as input data evolve over time.
- How is the model tested?
- If it is only tested on simulated data, how that is generated and how well that represent real data should be further studied/

Examples of Applications – While the discussion in the draft lays out the concepts that can help with explainability, in order for these to be better understood and adopted by AI system designers and other practitioners, it would be very helpful to include illustrative examples of how these principles can be applied in real world systems. For example, AI systems are often used in identifying and selecting a marketing audience for targeted marketing communications. In most circumstances, this use case does not have a significant impact on individuals and presents a low risk of harm. Consequently, it may be sufficient to provide a simple, high-level explanation of the factors that are considered in making targeting decisions.

We further note that the terms "user benefit" and "owner benefit" may need clarification as the system operator and user may well be the same person, but the system operator might not be the system owner.

Additionally, development of a framework for balancing meaningfulness and explanation accuracy with function criticality may be needed. So, for example, a system may misidentify a Scarlet-Banded Barbet (a bird native to South America), as a Yellow-Bellied Sapsucker (native to North America), but satisfy explainability requirements by noting that it has only been fully trained on North American birds.

<u>Individual Level Versus System Level</u> – The draft should also consider including a discussion about the balance between explainability at the individual level versus explainability at the system level. This concept is particularly relevant when a system is composed of a number of individual AI components. In such a circumstance, it is worth considering whether or not explainability is needed for each individual AI component if that explainability can be achieved at the system level. And, conversely, explainability at the system level may not be needed if the subcomponents of a larger system are explainable. The figure that shows the four use cases currently being discussed in the document is useful but would benefit from the inclusion of more complex use cases as suggested above.

<u>Explainability and System Learning</u> – Another element of particular value would be a discussion about how explainability might change (or might evolve) with system operation over time or with integration over time as the system learns over time. This can help address the question of explaining aggregate effects as opposed to explainability as a snapshot of a moment in time. For this reason, the ability of a system to assess and report its own competency would be useful to understanding its trustworthiness.

In summary, it might be useful to clarify what facets or categories of AI must be explainable. The document focuses on machine learning and particularly convolutional neural networks, but perhaps those are only examples of AI. The provision of more examples may help to resolve any confusion.

Finally, USCIB members have raised some basic questions regarding the draft principles which are reproduced in order to help inform the next version:

- How were the four cardinal principles decided and were they distilled and vetted from the community of XAI? It might be useful to include a brief history in the document.
- Is there a need to reframe the original thesis of this paper, which focuses mostly on "output," and expand this to "output, decisions and action?"
- How statistically/conceptually independent are the four principles?
- Why was there no reference made to Artificial Generalized Intelligence (AGI) as possibly relating to the roadmap of XAI?
- Is there a need to develop a "Language" to standardize the accuracy and semantics of Explainable AI, perhaps based on the matrix included in the paper?
- In terms of mapping the task of XAI, is it worth considering that there is a cost/criticality trade-off which might lead to a standardization of "Levels" of Explainability that takes cost and criticality into account?
- Would it be useful to consider merging "meaningfulness" and "explanation accuracy" into one attribute to remove confusion over the intended audience?

- Should we refer to output or actions from AI systems as being “biased” or would other terms such as “in error” or “mis-prioritized” in most cases be more descriptive?
- Should there be some more discussion around “counterfactual explanations?” While counterfactual explanations may enable companies to protect IP and be intuitively satisfying, they could also be irrelevant or at odds with overall system intent.
- Regarding human explanation accuracy and decision accuracy, should the distinction between the two in human reasoning be applied to AI systems? If the desire is to create a trustworthy AI, then it may be of interest to investigate the perceived trustworthiness of systems that can or cannot be presented with and properly react to a fact countering its assumption model.

Thank you for the opportunity to provide input. We stand ready to assist NIST and the Administration in ongoing efforts to ensure U.S. leadership in AI. This includes advancing trust and explainability so the American people may fully realize the positive benefits of these technologies.

Sincerely yours,

Barbara P. Wanner

Barbara P. Wanner
Vice President, ICT Policy
U.S. Council for International Business
1400 K Street, NW, Suite 525
Washington, DC 20005