October 15, 2020

National Institute of Standards and Technology
Attn: Information Access Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, MD 20899-2000

Via email to: explainable-AI@nist.gov

Re: ITI Comments on Draft NISTIR 8312: Four Principles of Explainable Artificial Intelligence

The Information Technology Industry Council (ITI) appreciates the opportunity to submit comments in response to the draft publication released by the National Institute of Standards and Technology, *NISTIR 8312: Four Principles of Explainable Artificial Intelligence.*

ITI represents the world's leading information and communications technology (ICT) companies. We promote innovation worldwide, serving as the ICT industry's premier advocate and thought leader in the United States and around the globe. ITI's membership comprises leading innovative companies from all corners of the technology sector, including hardware, software, digital services, semiconductor, network equipment, and other internet and technology-enabled companies that rely on ICT to evolve their businesses.

Artificial Intelligence (AI) is a priority technology area for many of our members, who develop and use AI systems to improve technology, facilitate business, and solve problems big and small. ITI and its member companies believe that effective government approaches to AI will clear barriers to innovation, provide predictable and sustainable environments for business, protect public safety, and build public trust in the technology. We support the responsible development of AI in line with the preceding objectives and view the release of these four principles as a positive step toward our shared goal of advancing the development of trustworthy AI.

In general, we support and advocate for risk-based, outcome-based approaches to policy and governance in the use of AI technology. It is our view that this will be the most effective means of addressing concerns that may be associated with AI, while simultaneously allowing for innovation and agility in development of AI applications. The AI development process is fast-evolving, highly varied between organizations, and geographically and technologically diffuse, while AI models themselves have the potential to be complex and highly commercially sensitive. These factors combine to suggest that an extensive, prescriptive, 'one size fits all' approach to AI governance will face similar, if not greater, challenges than in other areas of technology policy. This challenge is also manifest in the very diverse application of AI and machine learning (ML) solutions which vary in sensitivity, risk and benefit.

We acknowledge that algorithmic transparency is an important aspect of and necessary to developing accountable and trustworthy AI and avoiding biased outcomes. It is increasingly important to provide a level of accountability for decisions an AI system makes and explainability is a key aspect of this. We support the development of meaningfully explainable AI systems and increasing the availability of tools to achieve this. In that vein, we very much appreciate NIST's

Global Headquarters
700 K Street NW, Suite 600
Washington, D.C. 20001, USA
+1 202-737-8888

Europe Office
Rue de la Loi 227
Brussels - 1040, Belgium
+32 (0)2-321-10-90

info@itic.org
www.itic.org
@iti_techtweets

initial efforts to begin a conversation on explainability. To the extent NIST is endeavoring to create a common language or lexicon to better equip stakeholders to engage in more informed and meaningful discussions around these topics, that is a laudable goal worth pursuing.

While we welcome NIST's effort to tackle the hard problem of AI explainability and recognize this paper is not intended to create regulation, we do have questions regarding the scope of this effort, as indicated by what appears to be an underlying premise of the paper: that every single AI system must be explainable. As we elaborate on below, it is our view that it is not necessary for every AI system to be explainable, and in fact, in some cases requiring explainability may not only be technologically infeasible but could also hamstring continued innovation in AI applications by industry.

In this context, the intended audience of this paper would also greatly benefit from a discussion about the limitations of explainability and when explainability is and is not needed. Similarly, we encourage a discussion about the potential pitfalls of requiring explainability in all cases, particularly when in some instances there may be other approaches -- such as system reliability or engineering controls -- that can better address goals like ensuring the safety and trustworthiness of the AI system.

If we can start from a baseline, informed by common terminology, stakeholders will be better equipped to recognize where explainability makes sense in the AI space, and where it might not. Some low-risk applications of AI, for example, may not necessitate the same type of explanations as higher-risk applications; and in more benign applications that carry non-significant impacts on individuals, explanations may not be necessary at all. NIST's effort would be well-served by clarifying that these principles are not intended to be viewed by policymakers with a sense of rigid finality, to minimize the potential they could be translated into a requirement that every AI system be explainable, an unfortunate and unintended result that would hinder innovation and subvert the benefits of AI.

Equally important is recognizing that the attributes of explainability identified by NIST should not be prematurely used as the basis for metrics for testing or certification of AI algorithms, products, or systems. This initial set of attributes, while providing some insights into conveying why a particular system might have arrived at a particular outcome, are not quantifiable and have not been rigorously vetted as being adequately reliable for the purposes of testing and/or certification of AI systems. NIST should clarify that any such use - which significantly raises the risk of creating a misplaced sense of confidence in a product where such confidence might be important – is not intended at this time.

As a general matter, we also regularly stress the importance of maintaining global interoperability and alignment of various AI frameworks to the greatest extent possible in the spirit of ensuring that technologies of all kinds can be used safely and reliably across the globe. In this era of global digital commerce, political and regulatory divergence poses real risks to the socio-economic benefits and opportunities of data-driven technologies such as AI, where fair, accurate, fit-for-purpose models depend on access to large, diverse data sets that can flow across borders. Thus, when considering how to approach AI explainability, we encourage NIST to also understand and take into account how countries around the world are thinking about this subject as well – with the goal of building a "common language" discussed above to inform and facilitate global conversations and harmonized policy development on this topic.

We also have a series of specific comments, which we believe will help to strengthen the draft document. We outline these in more depth below.

Specific Recommendations

It may be helpful for NIST to:

1) Recognize that garnering acceptance of AI systems is a shared responsibility.

The document places the onus to achieve societal acceptance and trust in AI on the developers of AI systems. However, developers are not the only – or even primary – individuals that bear that responsibility. We believe that multiple stakeholders, including business and policymakers, need to recognize the importance of gaining societal acceptance and trust. This will require these stakeholders and others (salespeople, product managers, executives, consumers, and end users) to work collectively with developers to achieve these goals. In effect, societal acceptance and trust in AI systems should be part of strategic governance and strategy discussions. NIST should reflect this shared responsibility in their introductory section.

2) Reconsider the use of the term "principles," while also adding more rigor to other key terms used in the document.

As a foundational matter, we encourage NIST to exercise caution in the use of the phrase "principles." While we think that a common lexicon for discussing AI explainability will ultimately be helpful, in many instances, particularly when considered by global policymakers, principles are thought of as something foundational, a finalized endpoint that they can look to in order to provide some level of certainty or direction. However, given the fact that our understanding of explainability is still evolving and it is likely that these attributes will similarly evolve, it may be premature to call these "principles." We instead recommend using the term "attributes," which perhaps assigns a less irrevocable quality to the information contained in NISTIR 8312. Furthermore, this approach would provide NIST the flexibility to update these attributes, including adding new attributes as NIST and the stakeholder community's experience with the issue of explainability grows.

The document also uses the terms "humans," "users," and "consumers" interchangeably, which adds a degree of confusion. Additionally, "users" is defined in multiple ways throughout the paper, adding to the confusion. We therefore recommend that NIST choose one term and use it throughout or otherwise clarify whether each term used means something different in the context of the paper and if so, how the reader should appropriately differentiate between the usage of these terms.

We also recommend that NIST define the phrase "high-stakes," used in line 120, or replace the term with something more common (i.e., "significant, or "high-risk"). Defining "high-stakes" or replacing with a commonly used term like "significant" will help readers understand what is meant by this term. Generally, in our view, an AI decision is "high-stakes" when (1) we cannot entirely test for safety, (2) when the notion of fairness is too abstract to be encoded in a machine learning explainable component, and (3) when a negative outcome could have a significant impact on people—especially as it pertains to health, safety, discrimination, or human rights.

3) **Consider the limitations of explainability principles.**

As global conversations on AI progress, particularly on complex topics like explainability and bias, we are grateful that NIST is thinking about how to best approach mitigating risks while encouraging the responsible growth of this technology. Indeed, we recognize that facilitating trust in AI is imperative to fostering widespread adoption of the technology and agree that one of the ways to do so is to improve understanding of and transparency associated with AI systems. We are therefore generally aligned with the principles laid out in the document. At the same time, as mentioned above we have questions about the baseline assumption that is made in the paper: that explanations are necessary or desired by users for *all* AI applications, or that *all* outcomes require an explanation. We think it important that NIST recognize and reflect in the document that there is a diversity of AI applications and some uses likely will not require the level of transparency suggested within the document, or require explainability at all. We encourage NIST to consider that the level of need for and desire from various stakeholders for explanations may significantly differ based on the AI application or intended use, and factor in risk and impact to determine when explainability should be required and to what degree. We suggest NIST avoid using the word "obligates" on line 173, which may be too strong.

Beyond that, it is important to consider whether the approach recommended in the paper could present any downsides and encourage NIST to include a more thorough discussion to provide a balanced perspective. For example, NIST should consider how explainability information will be presented to consumers and end users. If too granular of a level of detail is ultimately required, or the scope of explainable use cases becomes too large, requirements around explainability could ultimately have the opposite of NIST's intended effect. In addition to harming the consumer experience in some circumstances, consumers may ultimately become inured to voluminous or excessively detailed explanations and in turn tune them out.

The document would also benefit from a discussion about the potential issues that may arise from attempting explanations where explainability does not address a clearly defined and identifiable outcome. Explainability does not necessarily equate to a higher confidence level, it simply means that actions/decisions can be explained (without indicating that they are the 'right' ones). This could lead to a misplaced sense of confidence or understanding of an AI system, when an enhanced understanding of the AI system's scope and limitations might be needed, or system confidence could be developed through the use of appropriate tools; for example, engineering controls and application-specific standards such as safety, security, reliability, and performance engineering controls.

4) **Consider that when the risks are low, it may not be necessary for every principle laid out in the paper to be met.**

We strongly support the acknowledgment that appropriate explanations will and should vary according to audience, context of use, as well as other risk-based factors. To this end, there may be instances in which meeting each of the four principles is unnecessary and explainability may be achieved in some other way. For instance, if a keyboard auto-correct function is considered to be an AI application, should the same level rigor of explainability be required as for an AI-application being used to identify potential cybersecurity threats or attacks in a data stream?

ITI   Promoting Innovation Worldwide   ⊕ itic.org

Further, the application of the principles may vary according to whether the party is the developer or the implementer of the AI, or another party entirely such as someone who is evaluating the system. Explanations and expectations appropriate for the developer of the AI system and the end users with whom it interacts are quite different. Additionally, the intended use-case of a given AI application can similarly influence the type of explanation required, or indeed whether any explanation is necessary. For example, a benign AI application that allows users to transform their profile picture into a Renaissance-era portraiture likely doesn't require an explanation of how the system achieved its given outputs, whereas a higher-risk or more materially impactful application, like a medical diagnosis or the result of a mortgage application, would. It would be helpful for NIST to add more context around when AI systems may or may not need to be explainable, and when explanations are critical. As relevant, it may be helpful for NIST to delineate the different roles or assert that flexibility must apply here. As AI technology continues to flourish and we develop new and innovative ways to implement explainable AI practices and mitigate bias, ITI encourages NIST to avoid prescriptive rules or frameworks that could stifle these dynamics.

In this context, we strongly encourage NIST to provide illustrative examples of how these principles could be applied to real world applications, and across different contexts involving different stakeholders. Doing so would provide users more practicable advice about how to consider, design and implement explainability in their systems. Currently, the discussions in the paper are rather abstract and leave a number of issues open to interpretation and question.

5) **Consider refining the four principles in the ways outlined below, including to make them more consistent with the concept of "attributes."**
- *Explanation*
    - NIST should consider replacing the term "Explanation" with "Explanatory" because it makes more sense within the context of "attributes" and as a principle.
    - NIST should include data explainability in the first principle, "Explanation." The data used to train AI systems and the variation of the labels between annotators is rarely measured. Given that this is a source of bias in most systems, measurement is critical.[1,2]
    - On lines 179 – 180, Explanation Accuracy does not impose any metric of quality. This conflicts with lines 216 – 217 of Section 2.3 Explanation Accuracy.
- *Meaningful*
    - Meaningfulness, as defined in the NISTIR, is difficult to measure. We recommend including the discussion referenced above about comprehensibility and actionability, to help further enable entities to characterize and measure the relevance of explanations.
    - It may help to replace the term "meaningful" with "relevant" as this could assist in clearing up some of the confusion and uncertainty around the word "meaningful." Using the term "relevant" may help address concerns and questions around just how much information and what type of information may need to be provide by an AI

[1] Nassar, J., Pavon-Harr, V., Bosch, M., McCulloh, I. (2019). Assessing Data Quality of Annotations with Krippendorff's Alpha for Applications in Computer Vision. In *Proc. AAAI 2019 Fall Symposium.* Arlington, VA: AAAI

[2] McCulloh, I., Burck, J., Behling, J., Burks, M., Parker, J. (2018) Leadership of Data Annotation Teams. In *Proceedings Social Sens 2018*. Orlando, FL: IEEE.

system. Some information may not be relevant to an explainee in certain contexts but certainly relevant in others.

- o We recommend asking several questions when considering the meaningfulness of explainable AI: How do you quantify meaningfulness? Meaningful to whom? Which protocols should be used? Which questions should we ask of human annotators? How do we sanity-check those questions? How do we sample the audience? How do you handle human disagreement?

- *Explanation Accuracy*
  - o NIST should consider replacing the term "Explanation Accuracy" with "Explanation Quality." Most current Explainable AI (XAI) research is focused on the quality of an explanation, not accuracy (accuracy is in fact a separate area of research and development). This section opens a broader discussion on benchmarks for XAI pipelines, evaluation protocols, and metrics. Some protocols are purely synthetic and do not involve humans, while others involve expert or non-expert human annotators. These emerging discussions should be more clearly reflected in the draft report.

- *Knowledge Limits*
  - o NIST should consider replacing the term "Knowledge Limits" with "Knowledge Limitation" or "System Awareness" as these terms may more clearly identify what the contents of the principle are trying to achieve. It appears that the Knowledge Limits section is suggesting that a system must understand if it is being weaponized. It is impractical to harden systems in this way.

6) Provide further clarity around the scope of AI as addressed in this paper.

As we have indicated in previous comments[3], there is not a universally accepted definition of AI. Beyond that, the scope of AI is incredibly broad and could theoretically capture many different types of systems and processes. There are currently a number of nuances that are not captured in the text and that unfortunately contributes to a vague scope for the sake of defining and discussing explainability.

Particularly when thinking about explainability, it would be useful to differentiate between the types of AI systems and make clear what it is that is trying to be explained and by which type of system. Certain types of AI systems may be more explainable than deep neural networks, for example. One way to narrow the scope may be to consider whether the system needs to be explainable, and if so, in what form and to whom? Once the scope has been narrowed, it will be easier to assess when it might be infeasible to explain individual outcomes. Furthermore, it would be helpful for the paper to address the issue of how the need for explainability and the explanation itself change as a self-learning system learns and gets better. Does NIST envision the need for explainability changing as the system evolves or does NIST view explainability as a static metric that is determined at a particular point of time?

7) Clarify the goal of the paper as well as the intended audience.

When reading the paper, it is not immediately apparent who the intended audience is. It is not clear whether the audience is a developer of an AI system or a government regulator who has to figure out whether the system poses a high level of risk. Indeed, while the paper does not seem

---

[3] https://www.iti.org/policy/1009ITIViewsonIIARequirementsforArtificialIntelligence.pdf

consumer-focused, without additional clarity it leaves that possibility open. This sort of distinction can have a significant impact on how the paper is perceived and we encourage NIST to consider clarifying the paper's audience at the outset. Perhaps one way to consider the question of who the target audience is is to think about the reason underlying the explanation. For example, is the explanation needed for experts to stress test a system, for regulators and/or policymakers evaluating a system for safety features or developing a framework for AI, or for consumers to better understand what they can expect from AI technology? We appreciate that NIST has thought about this to some extent, as evidenced by the section that explores the five types of explanations. However, we would encourage a deeper dive into this, including a section that discusses counterfactual explanations. These types of explanations can often be more actionable and beneficial than other explanations. Counterfactual explanations make human-machine collaboration possible even if the AI was not designed to explain its decision-making process. For example, a counterfactual explanation could tell a rejected loan applicant which inputs (income, assets, etc.) would have needed to change for the application to have been approved.[4] This provides the loan applicant with concrete actions they can take to alter the decision made by the AI system.

Beyond the issue of intended audience, the purpose of the paper is vague. Is it simply an academic exercise or are the principles intended to be practically implementable? If the latter, NIST should elaborate on how these principles could be implemented in practice (and, importantly, by whom) and provide examples of how one might use the principles, including in different contexts, in a future iteration of the document. In doing so, it is likely important to consider and explain whether and how to demonstrate that a system is meeting and/or exceeding the principles or attributes laid out.

8) **Include information about the process for updating this document.**

Finally, we noticed that there is no information included as to how NIST intends to update this draft guidance as NIST and the larger AI community gain experience with AI systems and develop a better understanding of how best to use the range of tools that help further confidence in AI systems and their outcomes. NIST should provide such information in a future iteration of the document.

<p align="center">***</p>

Once again, ITI appreciates the opportunity to provide feedback on this draft document. We appreciate NIST's desire to begin a conversation about this important subject and encourage NIST to continue to engage with stakeholders as it seeks to refine this document. We absolutely think that explainability plays an important role in facilitating trust in AI and support NIST's efforts to begin a conversation around it. At the same time, it is important to balance the benefits from continued innovation and flexibility in meeting the shared goal of trustworthy AI.

Sincerely,

---

[4] Costabello Luca, McGrath, Rory. "Interpreting AI 'Black Boxes' with counterfactual explanations." 31 July 2019.

John S. Miller
Senior Vice President of Policy
and Senior Counsel

Courtney Lang
Director of Policy