ITI Promoting Innovation Worldwide

September 8, 2021

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: ai-bias@list.nist.gov

RE: ITI Comments on National Institute of Standards and Technology (NIST) Draft Special Publication 1270: A Framework for Identifying and Managing Bias in Artificial Intelligence

The Information Technology Industry Council (ITI) appreciates the opportunity to submit comments in response to the publication of *Draft NIST-SP 1270: A Framework for Identifying and Managing Bias in Artificial Intelligence* (the draft).

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 clear barriers to innovation, provide predictable and sustainable environments for business, protect public safety, and build public trust in the technology.

ITI is actively engaged on AI policy around the world and issued a set of *Global AI Policy Recommendations* earlier this year, aimed at helping governments facilitate an environment that supports AI while simultaneously recognizing that there are challenges that need to be addressed as the uptake of AI grows around the world.[1] We have also actively engaged with NIST as it has considered various aspects important to fostering trust in the technology, most recently on explainability.

We share the firm belief that building trust in the era of digital transformation is essential and agree that there are important questions that need to be addressed with regard to the

[1] Our complete *Global AI Policy Recommendations* are available here: https://www.itic.org/documents/artificial-intelligence/ITI_GlobalAIPrinciples_032321_v3.pdf

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

responsible development and use of AI technology. As AI technology evolves, the tech industry is aware of and is already taking steps to understand, identify and mitigate the potential for negative outcomes that may be associated with the use of AI systems, including biased outcomes.

We appreciate that NIST is exploring many areas that will be of importance to fostering trustworthy and responsible AI. We agree that determining an effective approach to addressing bias is vital, especially given the frequency with which it comes up as a very specific risk that policymakers around the world are concerned about. Addressing bias will require collaboration across the public and private sectors in order to foster a practical understanding of how AI tools are designed, developed, and deployed and create state-of-the-art approaches to address identified challenges. It is also necessary to develop data-driven techniques, metrics, and tools that industry can operationalize to properly measure and mitigate bias in concrete terms. On the whole, we agree with the way that NIST has broken down the AI lifecycle, as well as the stages in which bias can be introduced and also managed.

Below, we offer several recommendations and perspectives that we encourage NIST to consider as it revises its draft, which we believe will help to strengthen it.

Specific Considerations & Recommendations

1) **Provide additional language to indicate the preliminary nature of the draft, adding useful context for international policymakers so they do not view it as a definitive guide to approaching and managing bias.**

At the outset, it is necessary to note that while a comprehensive approach to detecting and mitigating bias is important, generally accepted approaches for doing so in *all* circumstances do not yet exist. We recognize that NIST's draft attempts to contribute to the effort to develop standards and a risk framework for building and using trustworthy AI by focusing on the challenge of bias in AI. However, to appropriately build a framework that does so, consensus methods for assessing, measuring, and comparing data and AI systems, as well as standards for reasonable mitigations, are needed. This will require the development of new frameworks, standards, and best practices.

That being said, we believe the draft is a solid first step in contributing to the conversation around addressing AI bias, though would exercise caution in describing it as a "framework" given it is still at a preliminary level. Given the success of NIST's Cybersecurity and Privacy Frameworks and the interest of global policymakers in adopting those frameworks, we are somewhat concerned that policymakers may misinterpret the draft as a definitive guide to addressing and managing bias without recognizing that additional standards and practices need to be developed to do so effectively. As such, we recommend NIST include language up front to provide additional context including to explain that this is a preliminary document and is not intended to solve for every bias-related challenge in all circumstances.

2) **More clearly state whether the paper is referencing unintentional or other types of bias and clarify definition of bias.**

We encourage NIST to more clearly identify up front the different types of bias (e.g., by using definitions and labels for the bullets and scenarios outlined on p. 4, line 335) and provide more detail on how the approach outlined in the proposed framework aims to specifically address each type. While there is some discussion of conscious or unintentional bias in the paper (e.g., line 453, p. 7), we recommend NIST include more discussion around these concepts earlier in the paper. It may also be helpful to provide the definitions of different types of bias as they appear in the paper, as opposed to listing them all in a glossary, as doing so will provide helpful context for the definitions. Finally, we encourage NIST to define bias more clearly, particularly given the divergent definitions of statistical, legal, and other types of bias.

3) **Consider that in order for the framework to be effective, more specific technical guidance is needed.**

We recognize that this is a preliminary document. However, in future iterations of the draft, it would be helpful for NIST to offer concrete technical guidance as to *how* to address bias in specific instances.

For example, NIST notes that pre-deployment testing can help address public distrust. However, for certain classes of AI technologies, particularly novel ones, large testing datasets do not exist, making it difficult to test them to the same extent as in areas where large test sets do already exist. In cases such as these, articulating standards, a framework or set of guidelines that reflect these differences in capabilities would be helpful. Additionally, for a certain important subset of bias testing, labelled data about protected or marginalized identities is often required[2]; however, such data is not always available, and it often is not easy to collect and store such information. In its efforts to develop a framework to identify and manage bias, it would be helpful for NIST to develop additional guidance about what reasonable expectations are in such circumstances.

NIST also references "disparate impact" throughout the draft. Providing additional technical guidance around how a developer could test for disparate impact in machine learning contexts would be helpful, as the notion of disparate impact has a particular, established interpretation that can be challenging to map into novel contexts. The draft also references "valid" performance; guidance here would be similarly helpful as notions of validity in predictive tools are well-established yet still contested in certain domains such as employment (e.g., the American Psychological Association's Principles for the Validation

[2] https://arxiv.org/pdf/1912.06171.pdf

and Use of Personnel Selection Procedures[3] differ from those outlined in the Uniform Guidelines on Employee Selection Procedures).

Finally, we recognize that NIST is working to develop an AI Risk Management Framework. In revising and updating this draft, we encourage NIST to consider how this Framework and the AI RMF complement or otherwise interact with each other. We agree that a risk-based approach to identifying and managing bias is appropriate, but as the concept is used frequently in the abstract throughout the draft, we recommend including cross-references to the RMF or otherwise defining more concrete guidelines around what constitutes a risk-based approach.

4) **Provide additional guidance related to problem formulation in the pre-design stage.**

In the pre-design stage, NIST references problem formulation as one step of the AI lifecycle where bias can be introduced, but also mitigated. We agree that the problem formulation phase is a critical inflection point for understanding and working to mitigate potential biases in a system. However, this element of AI design is highly context-specific and may be more challenging for practitioners less familiar with the risks of AI bias to operationalize compared to technical standards. Therefore, providing more concrete guidance around how NIST imagines standardizing expectations regarding the project/problem formulation would be helpful. In this way, industry will be better equipped to ensure the introduction and enforcement of relevant guidance and be confident that such guidance is in line with reasonable best practices. Some examples of such guidance could include suggesting more diverse (e.g., in characteristics, professional backgrounds, or life experience) representation in product formation and strategy teams or the recommendation to include or consult members of the community in which the technology will be deployed to understand how it might impact them (e.g., smart cities).

As NIST notes, developers may prioritize using data most readily available, but it may often be the case that no other datasets exist and that the task is nevertheless critical to performing predictive tasks (e.g., a model to enforce policies to prevent harmful online harassment may only have data that has been previously labelled to be harassment, and not control data that is free of flagged words or hate speech). As such, it may also be helpful for NIST to spearhead a USG effort to consolidate and make readily available inclusive datasets for this kind of training, as data and datasets – aside from a handful of large public ones -- are inherently expensive to license or create.

[3] https.www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf

5) **Include education/awareness as an important action to take in managing and addressing bias in the pre-design phase.**

While NIST puts forth many considerations around where bias may emerge and how or where it might be managed in the pre-design phase, a recommendation to boost education and awareness of the different types of bias and their impacts is missing. Training sessions, in particular, are subject to inherent biases as humans who are not properly trained will take their biases with them to the development process. While bias cannot be removed completely, it is vital that bias is both disclosed and assessed. We thus believe that adding education and awareness of bias as a category would be helpful, as this is one way in which developers can begin to understand and manage bias. Indeed, awareness that bias exists is the first step to addressing it.

Specific ideas that NIST could consider to address the above are:

- Practices, standards, curricula, and modules around educating developers about identifying risks from bias and approaches for mitigating those risks.
- Practices and standards around how AI providers can educate downstream users about bias so that developers and system operators who incorporate AI systems or components into larger systems can address bias in *their* design phases, with a particular focus on prevention of deploying AI systems in unexpected ways or on unforeseen populations, which is a frequent source of bias.
- Practices and standards around educating end users around the intended use of AI systems, so that systems are used for their intended purposes.

6) **Reference and integrate ongoing standards efforts.**

NIST should consider how and if it can reference ongoing international standards development activities in the draft. To be sure, ISO/IEC JTC 1 SC 42 is exploring themes that may be relevant to reference and/or integrate in the draft. For example, that committee is developing a standard on bias in AI systems and AI aided decision-making. We encourage NIST to consider if or how this standard can be integrated into future guidance to facilitate interoperability.

7) **Articulate a clear plan for how work on measuring and mitigating AI bias will translate into adoption across federal agencies.**

Recently the FTC has announced its intent to begin investigating AI bias. At the same time, the DOJ, HUD, CFPB, and other government agencies are also increasingly concerned about AI. There is, however, no concrete consensus understanding regarding what the terms fairness and bias mean in the context of AI and what reasonable efforts companies should be expected to take to mitigate bias.

The proliferation across the federal government of varied definitions of fairness and bias, and differing expectations regarding what constitutes reasonable efforts to mitigate bias,

creates a real challenge for entities seeking guidance on how to develop and measure bias in their AI systems.

NIST played a pivotal role in driving adoption of cybersecurity standards across the federal government. We therefore encourage NIST to articulate a clearer plan for how its work on measuring and mitigating AI bias can translate into harmonized standards across federal agencies.

***

Once again, we appreciate the opportunity to provide feedback on *Draft NIST-SP 1270.* Developing a practical approach to identify and manage AI bias is a key component to ensuring trustworthy AI systems. While we believe the draft acts as a solid foundation, there are several areas we believe can be strengthened in future iterations of the document, including outlining specific ways in which bias can be actively managed. We look forward to continuing to engage with NIST as it refines its recommendations around addressing AI bias. Please contact John Miller (JMiller@itic.org) or Courtney Lang (CLang@itic.org) with specific questions.

Sincerely,

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

Courtney Lang
Senior Director of Policy