URBAN INSTITUTE · ELEVATE · THE · DEBATE

May 5, 2022

National Institute of Standards & Technology
100 Bureau Drive
Gaithersburg, MD 20899

RE: Comments on NIST Initial Draft of Artificial Intelligence Risk Management Framework

To Whom It May Concern:

**Introduction and Disclaimer**

We appreciate the opportunity to submit feedback on this initial draft of NIST's AI Risk Management Framework. Though we are affiliated with the Urban Institute, the views in these comments are our own and should not be attributed to the Urban Institute, its trustees, or its funders.

Many of Urban's notes that were raised in our listening session with NIST about the concept paper were addressed well in this initial draft. In particular, the scope and audience both seem appropriately and comprehensively defined, and the introduction highlights the major issues well. The three-class taxonomy used to frame AI risk in Section 5 is a welcome addition, and each term is defined and contextualized clearly. The paper also rightfully discusses the importance of involving communities affected by AI systems in the conversation about design and deployment. The categories in both the *Measure* and *Manage* functions that detail the collection of feedback using participatory methods give needed credibility to those sections of the paper. We still believe there are a number of areas to be addressed in future drafts. Below, we organize our feedback by category.

**Understanding and Framing Risk**

Risk is well-defined in the initial draft, as are its drivers, but the next logical step of defining a risk function for various stakeholders is never taken. Communities that don't realize the downstream benefits of AI systems or are negatively affected by them would likely place greater weight on adverse outcomes than potential benefits, even if the designers of the system may equally consider downside risks and upside benefits. NIST concedes that the AI community is only beginning to understand scenarios resulting in harms, whereas AI benefits are already well-known. The inverse may be true for certain individuals or communities that have more exposure to potential harms than benefits due to their lived experiences.

Thinking about the distribution of risk is crucial, especially when it can lead to differences in risk tolerance. NIST buckets the types of harms very effectively in Figure 2, but this fails to account for how each of the entities within each bucket might have different risk tolerances. A business is affected by potential harms much differently than a person of color applying for a loan or

selling their house, for example. Not everyone has the privilege of focusing on the upside benefits of AI systems, and re-framing how NIST weights risk would be an important acknowledgment of that reality.

Finally, NIST categorizes the three types of bias as systemic, computational, and human. An analogous categorization for risk would provide more context as to how risk arises. Similarly, in Figure 2, NIST provides examples of different types of harms, but it would be instructive to say more about the underlying mechanisms that give rise to these three concepts of bias, risk, and harms, beyond just classifying them into types.

**Community Engagement**

The framework consistently mentions the importance of involving communities and those directly
affected by AI systems in the conversation. This is a crucial and laudable acknowledgment, but NIST does not take the next step of suggesting *how* to do that. Community-based methods may fall outside NIST's purview, but pointing to other relevant resources and examples would be a good first step. For example, Urban has published a blog post listing several valuable community engagement resources. We have also hosted a Data Walk to share data and research findings in close collaboration with community stakeholders, a concept which could be adapted and extended to the AI context.

**AI Taxonomy**

In general, section 5 is excellently written and framed, although certain subsections could benefit from more detail. *5.2.3 – Privacy* could speak to specific privacy concerns such as the reidentification of individuals in the training data and introduce terminology in the privacy literature such as differential privacy. *5.3.1 – Fairness* raises the important point about fairness being more than a technical exercise. However, including or at least pointing to technical definitions and terminology (such as "demographic parity", which is only briefly mentioned by name without explanation) would be useful and instructive for those who may be entirely unfamiliar with potential tools at their disposal. Again, even if defining these terms falls outside the purview of this taxonomy, it is crucial to then include other resources for designers of AI systems that fill knowledge gaps.

**The Govern Function and Protected Attributes**

The category about workforce DEI processes is a great addition whose presence could be amplified more in the rest of the document. Such principles belong not only in AI governance, but in the specific framing, measurement and management of AI risks as well. Additionally, organizations must go beyond the guiding principle of accountability and consider mechanisms for enforcement. Given that the RMF is voluntary and it is outside's NIST mandate to provide governance, the framework should include clear and strong guidance on self-enforcement within the "Govern" function.

Lastly, *nowhere in the RMF is race, gender, or any other protected attribute explicitly mentioned by name.* The identification of *specific* communities who can be affected negatively by AI harms adds a needed human dimension to the entire discourse, and not doing so feels intentionally vague. There are many widely publicized examples of the harmful effects of algorithmic bias (e.g. racial bias in the COMPAS recidivism and PredPol predictive policing algorithms, gender bias in pre-trained word embeddings, etc.), any of which could be mentioned explicitly as well.

We look forward to reading the second draft of the RMF and applaud NIST for its responsiveness in providing guidance on such an important challenge. Please don't hesitate to reach out with any questions or opportunities for further collaboration.

Sincerely,
Judah Axelrod – Senior Data Scientist, Racial Equity Analytics
Alena Stern – Associate Director of Data Science
Michael Neal – Principal Research Associate, Housing Finance Policy Center
Linna Zhu – Research Associate, Housing Finance Policy Center