COMMENTS ON
NIST ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK
DECEMBER 2021 CONCEPT PAPER

The U.S. Technology Policy Committee (USTPC) of the Association for Computing Machinery[1] appreciates NIST's invitation to solicit reviewers from among its membership for NIST's December 13, 2021 Concept Paper for an Artificial Intelligence Risk Management Framework. The Comments below are the joint work product of the following individuals,[2] whose affiliations are provided solely for identification purposes:

Ricardo Baeza-Yates, IEAI, Northeastern University
Member, ACM US Technology Policy Committee's AI and Algorithms Subcommittee

Thomas Y. Chen, Independent Researcher
Member, of ACM US Technology Policy Committee's AI and Algorithms Subcommittee

Nick Diakopoulos, Northwestern University
Member, ACM US Technology Policy AI and Algorithms Subcommittee

Abigail Matthews, Ph.D. Student, University of Wisconsin-Madison
Member, ACM US Technology Policy Committee's AI and Algorithms Subcommittee

Jeanna Matthews, Clarkson University
Co-chair, ACM US Technology Policy Committee's AI and Algorithms Subcommittee

Emanuel Moss, Joint Postdoctoral Fellow, Cornell University
Tech Digital Life Initiative and Data & Society Research Institute AI on the Ground Initiative

Arnon Rosenthal, The MITRE Corporation
Chair, ACM US Technology Policy Committee's Health Care Task Force

For clarity, the contributors' comments follow italicized restatements of NIST's recent inquiries:

[1] The Association for Computing Machinery (ACM), with more than 50,000 U.S. members and approximately 100,000 worldwide, is the world's largest educational and scientific computing society. ACM's US Technology Policy Committee (USTPC), currently comprising more than 130 members, serves as the focal point for ACM's interaction with all branches of the US government, the computing community, and the public on policy matters related to information technology.

[2] In August 2021, USTPC itself provided comments to NIST on its Artificial Intelligence Risk Management Framework in Docket Number 210726-0151.

1. ***Is the approach described in this concept paper generally on the right track for the eventual AI RMF?***

We recognize and understand the desire to "permit the flexibility for innovation, allowing the framework to develop along with the technology," but are also concerned that a "voluntary framework will not be sufficient to "create and safeguard trust at the heart of AI-driven systems and business models." We agree that "[t]ackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems."

We think it is important to identify a tier of system integrity levels where more risk management activities would be specified for applications of higher integrity levels. For example, applications that involve risk to human life could constitute the highest integrity level, applications in regulated areas (such as hiring, credit, housing and other spheres) could be at the next highest level, etc. The scope of the system's impact also should be considered; a system with a million customers thus would require more safeguards. We encourage NIST to set more prescriptive standards for risk management at these high levels of integrity for applications that involve risk to human life and applications in regulated areas.

Specifically, we would like to see increased requirements for:

- well-defined processes for validation and testing;
- support for auditing decisions in cases where harm is suspected;
- collecting and maintaining data provenance information;
- processes to enable inquiry by and redress as appropriate for individuals and groups adversely affected by algorithmically informed decisions.

We believe that market forces will often be insufficient to manage the aggregate risks to society and risk of harm to individuals. Decision makers often develop or purchase AI systems or automated decision making (ADM) systems to increase their own decision-making efficiency. Managing the risks that impact individuals, classes of individuals, or society as a whole can be costly and at odds with the desired speed of development/deployment and decision-making efficiency. As with other complex systems, like food or pharmaceutical safety, where it is difficult for individuals to truly inspect the risks from the perspective of users and consumers, government action is needed to establish standards and protect society.

If it is not feasible to *require* specific risk management tasks, in the interest of maximizing transparency it still would be beneficial to impose specific risk management requirements on individual AI systems and to ask organizations to reveal which of the various conditions for the assigned integrity level have and have not been met.

2. *Are the scope and audience (users) of the AI RMF described appropriately?*

The current AI RMF draft states that:

> "The intended primary audiences are:
> (1) people who are responsible for designing or developing AI systems;
> (2) people who are responsible for using or deploying AI systems; and
> (3) people who are responsible for evaluating or governing AI systems.

A fourth audience that will serve as a key motivating factor in this guidance is:

> (4) people who experience potential harm or inequities affected by areas of risk that are newly introduced or amplified by AI systems."

We encourage risk management activities that more explicitly and directly involve this fourth audience, including requirements for engaging communities impacted by AI systems in the identification of risks. Algorithmic risk assessments and algorithmic impact assessments are similar, but impact assessments tend to focus on individual and social impact and potential harms more than the management of risks as perceived by those developing and deploying the AI and ADM system.

3. *Are AI risks framed appropriately?*

The document identifies risks resulting from harmful biases and threats to safety, privacy, and consumer protection. It also notes that adverse effects from AI system operation are often "long-term, low probability, systemic, and high impact." We largely agree with this framing of AI risks but believe, however, that it's also important to acknowledge that AI systems in some domains may also produce low impact effects that may aggregate over time to create higher risk (*e.g.*, a media recommender system that over the course of years polarizes an individual's beliefs). The RMF thus might productively consider providing guidance for situations in which the measurement of risk is not feasible or possible due to extended timeframes, uncertainty, or other reasons. Furthermore, since different stakeholders may evaluate risk differently based on subjective values or priorities (*i.e.*, profiles), guidance on negotiating these differences may contribute to acceptance of the RMF.

4. *Will the structure – consisting of Core (with functions, categories, and subcategories), Profiles, and Tiers – enable users to appropriately manage AI risks?*

We like the emphasis on managing risks throughout the lifecycle of AI systems ("Pre-design: data collection, curation, or selection, problem formulation, and stakeholder discussions. Design and development: data analysis, data cleaning, model training, and requirement analysis. Test and Evaluation: technical validation and verification. Deployment: user feedback and override, post deployment monitoring, and decommissioning"). The functions identified (Map, Measure, Manage and Govern) mirror this focus on the full lifecycle. We are less clear on the intention for categories, profiles and tiers and recommend that NIST elaborate, perhaps through the use of one or more case studies.

5. Will the proposed functions enable users to appropriately manage AI risks?

The functions (Map, Measure, Manage and Govern) could indeed help users to manage AI risks. We agree with the need for "measurable criteria that indicate AI system trustworthiness in meaningful, actionable, and testable ways." We would encourage NIST to elaborate on these criteria for each of the four functions.

6. What, if anything, is missing?

We look forward to expanding on this answer in the informal roundtable discussion planned. We, and others attending, will encourage NIST to:

- Identify a tier of system integrity levels where more risk management activities are required for applications assigned higher integrity levels;

- Consider how to address the diffused responsibility for managing risks. When developing AI systems, it is common to reuse training sets, machine learning models and other system ingredients developed in other contexts, including in contexts for which they were not designed. We would encourage NIST to explicitly address the requirement for validating acquired system components and encourage metrics/metadata (e.g., trustworthiness standards) that would aid in this task;

- Consider methods for both incentivizing and enabling third party testing of deployed systems, especially testing focused on identifying errors that might impact specific individuals or groups;

- Aligning the AI RMF with other risk management frameworks to help AI risks be managed in conjunction with other risks. For example, there are overlaps between business intelligence risks and AI risks in categories, such as the safe use of large datasets, dataset bias and explainability to non-experts;

- Consider widening the scope of the Framework to automated decision making systems (ADMs). The risks considered are relevant for automated decision making systems regardless of whether they use AI technology or other types of algorithms internally. We recommend defining AI technology in the Framework in a way that includes automated decision making systems more broadly;

- Require AI system deployers to develop procedures for user access and redress, and to adopt mechanisms that enable inquiry, information access, and redress when appropriate for individuals and groups adversely affected by algorithmically informed decisions;

- Consider standards for documenting the data collection, model training, and performance evaluation processes that contribute to the development of an AI system. NIST is ideally positioned to provide guidance on what constitutes a standardized process for documenting representativeness of datasets, epistemological bases for associating data features with an output, or incorporating stakeholder input on degrees of acceptable risk.

Additional Resources

Statement on Algorithmic Transparency and Accountability, ACM U.S. Public Policy Council, ACM Europe Council Policy Committee (2017)

S. Garfinkel, J. Matthews, S. Shapiro, J. Smith, Toward Algorithmic Transparency and Accountability , Communications of the ACM , Vol. 60, No. 9, Page 5, Sept. 2017, 10.1145/3125780. Panel session at National Press Club (9/14/2017), Summary .

J. Matthews, Patterns and Anti-Patterns, Principles and Pitfalls: Accountability and Transparency in AI Association for the Advancement of Artificial Intelligence (AAAI) AI Magazine , April 2020. PDF (unformatted preprint)

D. Roselli, J. Matthews, N. Talagala, Managing Bias in AI, Proceedings of the 1st Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web , In Conjunction with the Web Conference 2019 San Francisco, CA, May 13-14 2019.

Gabriela Bar, Gabriela Wiktorzak, Jeanna Matthews, Four Conditions for Building Trusted AI Systems: Effectiveness, Competence, Accountability, and Transparency, IEEE Beyond Standards, July 13 2021.

J. Matthews, G. Northup, I. Grasso, S. Lorenz, M. Babaeianjelodar, H. Bashaw, S. Mondal, A. Matthews, M. Njie, J. Goldthwaite, When Trusted Black Boxes Don't Agree: Incentivizing Iterative Improvement and Accountability in Critical Software Systems, Proceedings of the 2020 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES) , New York, New York, USA, February 7-8 2020.

For additional information, or to further access the expertise of USTPC and ACM members, please contact Adam Eisgrau, ACM Director of Global Policy & Public Affairs, directly at 202-580-6555 or eisgrau@acm.org.