Responsible AI Institute
11501 Century Oaks Terrace, Suite 3125
Austin, Texas 78758

April 29, 2022

National Institute of Standards and Technology (NIST)
U.S. Department of Commerce
100 Bureau Drive
Gaithersburg, MD 20899

Comment from the Responsible AI Institute on the Initial Draft of the NIST AI Risk Management Framework

Dear Reva, Elham, and the NIST AI policy team,

Congratulations on the initial draft of the AI Risk Management Framework (AI RMF). The draft reflects your significant consideration of issues raised in the listening sessions and in the written feedback provided on the AI RMF Concept Paper. This is a huge accomplishment. We greatly appreciate you hosting a listening session with Responsible AI Institute (RAII) members on January 27, 2022 for the Concept Paper.

To support the ongoing development of this work, we held a feedback session with RAII members on April 21, 2022. Since one of the key themes that was raised in our listening session in January was the interoperability of NIST's AI RMF and other evolving frameworks, standards, certifications, and best practices, we invited Standards Council of Canada (SCC) to participate in the discussion and present their work on forthcoming AI certification pilots as well as AI and data standards.

In addition to SCC representatives from both the accreditation and standards branches, the feedback session's participants included AI practitioners in financial, insurance, and health institutions, independent auditors, AI development companies, AI risk detection and evaluation companies, and AI users. They therefore included people who are designing, developing, purchasing, and using different types of AI. Their feedback and comments have been augmented by our team's work with an emphasis on how the AI RMF can be useful to practitioners.

In developing our comments, we considered how the AI RMF relates to our ongoing development of a certification program to ensure the responsible design, development, and use of AI systems. The AI RMF provides a good anchor for our complimentary evaluation work. To date we have positioned our framework as an implementation framework to support the governance process that you have outlined in section 6, AI RMF Core, as it is use case-specific (eg. AI Certification for Automated Lending). Our implementation framework was established by creating an ontology of existing principle-based frameworks, testing this with specific use cases, then validating our insights with our robust community of members from industry, civil society, government, and academia.

As our evaluation and pilots progress, we will continue to share feedback as we believe that we are closely aligned with NIST's values and approach. We would specifically be interested in participating in the design of the AI RMF Profiles work. RAII is pleased to be a part of the aforementioned SCC AI certification pilot, which will test how these frameworks work together through an end-to-end development and independent audit process. Insights gained from this effort can help to support context-specific profiling.

For the feedback session that we hosted on April 21, we focused on two key questions outlined at the start of the draft AI RMF:

- What might be missing from the AI RMF (Question 7).
- Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added (Question 8).

We have provided high-level comments at a minimum for each requested question. However, for questions that are more relevant to our efforts, we have included more detailed feedback. As always, we are more than happy to clarify or follow up on any of the comments below. Thank you for the opportunity to provide input. We strongly believe that collaborative and transparent processes such as these are key to valuable tools, ultimately ensuring the trustworthy and responsible use of AI.

General Comments

Our general comments relate to the context setting for the document, organization, and nomenclature. We recognize that this is an initial draft and that we might not be aware of all of the objectives of the document.

While we strongly believe that the AI RMF will advance knowledge of how AI systems should be managed, as currently drafted, there could be some confusion about how this framework will be actioned within an organization. While the AI RMF has brought together important information, it is not always clear if the content is background or contextual information or if it is meant to provide a methodology for how an organization can evaluate an AI system. The scope in section (30-32) offers that “This voluntary framework provides a flexible, structured, and measurable process to address AI risks throughout the AI lifecycle, offering guidance for the development and use of trustworthy and responsible AI,” indicating the desired use would be for practitioners implementing AI. This could be resolved by either framing this as a proposed approach to governing AI or creating additional clarity on informative, contextual, or background content vs. directional content (the risk and trustworthiness taxonomy as a framework for mapping, measuring, and managing AI.) Maybe Part 1 is better framed as a context section which could benefit from referencing the AI RMF core alongside the scope and stakeholders.

Recognizing that the whole responsible and trustworthy community is still maturing and that common definitions are still evolving, it would be beneficial to include some definitions or to reference standard definitions. Section 2 (25-26), indicates that the AI RMF manages risks related to AI systems across a wide spectrum of types, applications, and maturity. We assumed this was the definition to be used for an AI system. If accurate, in the taxonomy, it would be helpful to incorporate more references to data objectives as well as model objectives. There seems to be more of a focus on the latter. For example, in 5.1.1.1 (6), Accuracy is discussed in the context of the ML model, but there could be non-ML models that might not be accurate based on the data inputs.

Lastly, in section 5, where the AI Risks and Trustworthiness taxonomy (taxonomy) is outlined, while drawing out the distinction between principles, technical, and socio-technical, it is important to identify that while there has been significant research to measure and automate the evaluation of AI systems, there is still limited capacity to do so. However, we have found (in creating a measurable evaluation relying on process-based and automated evaluation techniques) that classifying each of these risks and characteristics into one of these categories is not as straightforward in practice as it may appear. We question whether or not this mapping is as important as outlining the risks and characteristics, understanding that each category might have technical, socio-technical, and principle-based evaluation techniques. For example, one could argue that safety is a good guiding principle, but there are also ways to test that a system is following safety standards.

We understand that the AI RMF will be part of a broader NIST resource center. So, perhaps these comments can be resolved by referencing relevant resources where they exist or indicating what is in progress to situate the framework.

Specific Comments

The below responses to the NIST's requested questions provide additional details on our general observations.

1. **Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.**

   Section 4, Framing Risk, identifies harms that could occur to people, organizations, and systems. It would be useful to reference existing regulations and best practices like consumer protection or laws that protect physical, legal, social, financial, or psychological harms.

2. **Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape.**

   As the community and in particular practitioners are looking to have a more clear sense of what they should be doing to ensure their systems are being designed, built, deployed, procured, and used in a safe, trustworthy, and responsible manner, we would like to see more clarity within the objectives of each characteristic within the taxonomy. We recognize that since the scope is quite broad, they may not all be applicable to each use case. However, we would suggest describing how the framing of risk could be used to understand the potential harms of each system. Perhaps this could be included in the mapping process outlined in the Core. That way, the AI RMF would still be flexible in its application, yet clear enough that practitioners have a general understanding of the objectives of each risk and characteristic in the taxonomy.

3. **Whether the AI RMF enables decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks.**

   The AI RMF does an excellent job bringing together necessary contextual information that can help to advance an AI practitioner's understanding of AI risks, and provide direction on the key characteristics that should be considered throughout the development lifecycle. Based on our discussions with members and the RAII community, more specificity is needed to understand "what good looks like." While this may not be the best document for that, including more clear objective statements within each risk and characteristic of the taxonomy with a link to reference material that might be included in the NIST resource center or other standards efforts underway would be useful. This was a significant part of our discussion during the feedback session in April as SCC, CEN-CENELEC, ISO, IEEE, and other standards organizations are working on similar challenges and are advancing different aspects which could be useful reference material to a practitioner.

4. **Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.**

   A couple of additional considerations for inclusion in the taxonomy include: evaluation of the system's operations and aspects of consumer protection aligned with human rights principles.

   Some aspects of system operations - like accuracy and reliability - are typical to the operation and management of other technology systems. In our experience, it is ideal to incorporate data quality, management, and governance objectives here, too.

Similarly, consumer protection, including mitigating harms to individuals, incident reporting requirements, and protections for individual or group privacy, are important considerations when evaluating AI systems, in our experience.

While section 5 includes definitions and examples, they don't always match (for example, with regard to explainability and interpretability.)

5. **Whether the AI RMF is in alignment with or leverages other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42.**

The mapping in section 5 was useful. However, these are a combination of principle documents, or principles extracted from draft legislation which includes additional requirements which were not mapped into the current draft. A more fulsome mapping would be an incredibly helpful resource, though we recognize that this would be a significant undertaking. Several organizations, including ours, develop and maintain similar mapping documents, which could be leveraged for future drafts.

6. **Whether the AI RMF is in alignment with existing practices, and broader risk management practices.**

The AI RMF is well aligned with existing risk practices. As stated above, it would benefit from categorizing data-driven risks that are both inputs and outputs of AI systems.

7. **What might be missing from the AI RMF.**

In future drafts of the AI RMF, we suggest that NIST:

Include clear guidance materials on what does/does not qualify as a best practice

While the terms "best practices" and "practices" in general (e.g. "risk management practices") appear in the Initial Draft of the AI RMF, specific guidance on what constitutes an established practice in this context would be helpful. What is the threshold to become a best practice? What distinguishes good principles from a best practice? Concrete examples to illustrate this point could add further benefit.

Include examples that are focused on enterprises of all sizes

The AI RMF Initial Draft correctly notes that the requirements for appropriate risk management can/have to vary depending on the size of the enterprise. Examples of how a large company could apply the AI RMF alongside examples of how a smaller company could manage risk with fewer resources would be helpful and advisable.

8. **Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added.**

A regularly updated draft companion with further resources would be highly valuable. A common theme in our January NIST listening session and our April feedback session was that multiple helpful guidelines, toolkits, and principles for AI risk management are now available. Though components of these are incorporated in the Initial Draft of the AI RMF, a future draft and companion document should further build upon these, provide easy access to them, and describe or illustrate situations in which a given resource may be particularly relevant.

We carefully review the following resources, among others, in continuously developing and updating our RAII Implementation Framework, which our certification and governance tools are based upon. We are listing them here as examples of resources that could be included in the draft companion to NIST's AI RMF:

- OECD AI principles
- UNESCO Recommendation on the Ethics of Artificial Intelligence
- Canada's Directive on Automated Decision-Making Systems
- ISO's proposed Artificial Intelligence Management Systems
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- FTC guidance on AI
- EU Ethics guidelines for trustworthy AI
- Council of Europe's Report on AI systems
- UK BSI AI standards
- Global Partnership on AI Framework
- Singapore Veritas Initiative
- Singapore Principles for Fairness, Ethics, Accountability and Transparency (FEAT) in AI
- AI principles from industry and academia

We applaud NIST for this significant effort to integrate many evolving and important topics related to evaluating AI systems. We look forward to future drafts of the AI RMF and to subsequent discussions on these vitally important matters.

Sincerely,

Ashley Casovan
Executive Director

admin@responsible.ai
www.responsible.ai