29 April 2022

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to NIST's initial draft of its AI Risk Management Framework (AI RMF). A policy research organization within Georgetown University's Walsh School of Foreign Service, CSET produces data-driven research at the intersection of security and technology, providing nonpartisan analysis to the policy community. We appreciate the opportunity to offer these comments, and look forward to continued engagement with NIST throughout the Framework development process. We have organized our feedback according to several of the questions that were posed by the AI RMF. Line-by-line comments that do not directly correspond to a question in the AI RMF are included at the end of the document.

What might be missing from the AI RMF?

The AI RMF presents an opportunity to fill gaps and provide community standards. This is a much-needed contribution that NIST is uniquely situated to provide. We highlight specific gaps that the AI RMF could address, with some revisions:

- Clarify terminology. An area where there is a lot of room for contribution is defining and operationalizing terms included in the AI RMF. This includes AI-specific terms in the "AI risks and trustworthiness" taxonomy presented in Figure 3 but also terms like "use case," "harm," and "stakeholder." Relatedly clarifying the distinction between impact and risk, and having the more nuanced discussion about "positive" versus "negative" risk would strengthen the draft. Additional notes on this:
    - To be most usable across teams, organizations, and sectors, terms need to be clarified but also modifiable. Expect variation in understanding of these terms and ensure different roles or teams within an organization come to the framework with the same outcomes in mind. To that end, the AI RMF's terms need to be understandable and translatable across teams and sectors to make the process meaningful to all audiences and participants.
    - "Transparency" is well defined in the document but could be more explicit in acknowledging the element of notification or consent when consumers encounter AI-enabled systems.
- The process of applying attributes in Section 5 of the AI RMF to the high-level map, measure, manage, and govern functions could be more clearly defined. At a minimum, a short section about considerations for mapping some of the attributes to the high-level functions would be useful for readers of the AI RMF in addition to the practice guide and AI RMF profiles that will provide concrete examples for how to manage AI risk.
- The general public stakeholder group in Section 3 obscures an important distinction between the people most directly impacted by an AI system and people that may experience the spillover effects of an AI system. Dividing the general public group into

two smaller groups, one for those directly impacted and another for those indirectly impacted, would prompt readers of the AI RMF to proactively consider the subgroups of people that an AI system most greatly advantages or harms. Creating two groups would also help readers consider the broader consequences of an AI system outside of the immediate context or community in which it is deployed.

- While the principles of democratic governance may be read as implicit in parts of the document, the lack of specific reference is noteworthy, especially because democratic values are mentioned explicitly in the OECD Recommendation and the EU AI Act. Human rights are also notably absent, though only the OECD Recommendation makes explicit mention of human rights.

Does the AI RMF appropriately cover and address AI risks, including with the right level of specificity for various use cases?

The AI RMF mentions "societal" risks briefly, but does not provide any motivation for why this scale of risk is particularly pertinent with regard to AI systems. It would be valuable to incorporate the reasoning discussed in workshop #1, namely that a failure or harm caused by a single AI system (or set of closely related systems) can end up having a society-wide impact if the system is deployed at large scale. This property is why assessing risks to society makes sense in the context of AI.

Is the AI RMF flexible enough to serve as a continuing resource considering the technology and standards landscape?

Each of the framework's functions (map, measure, manage, govern) are continuous and evolving processes. To be effective the AI RMF needs to be part of an organization's broader commitment to assessing, mitigating, and discussing risk. While this may not be controversial, it could be spelled out more clearly in the draft.

The AI RMF can place greater emphasis on the importance of documentation throughout this process. This came up in several workshop panels and we agree it is critical to effective risk assessment as a continuous, iterative process. From mapping to governance, having clear and consistent documentation of the use case and relevant organization policies as well as the steps taken to apply the AI RMF and what was learned at each step is important.

Does the AI RMF enable decisions about how an organization can increase understanding of, communication about and efforts to manage AI risks?

The AI RMF could offer best practices for mapping, measuring, managing, and governing AI at the organizational level. As came up in workshop #2, best practices are sparse here, and not necessarily due to a lack of appetite. A core attribute of the AI RMF is taking advantage of existing best practices and methodologies, but perhaps existing best practices are not known. The AI RMF could contribute here by collecting, consolidating, and suggesting practices for organizations to consider. Best practices for revisiting and updating risk management processes as part of the govern function would be a helpful addition to the AI RMF, for example.

An important part of risk management is making known mechanisms for accountability. While details can be organization-specific (as noted pn pg. 13 lines 16-18), the AI RMF would benefit from discussion of how identifying and ensuring mechanisms for accountability fit into specific functions.

- While the AI RMF mentions that small and medium-sized organizations encounter different issues when managing risk (pg. 7 lines 32-33), the AI RMF would benefit from more detail about the disparate challenges that small and medium-sized organizations face, especially with regards to accountability mechanisms for risk management. This elaboration would be beneficial because characteristics that factor into risk calculations, such as access to resources and organizational reach, differ between small and medium-sized organizations.

Are the functions, categories, and subcategories complete, appropriate, and clearly stated?

We highlight feedback on the Map function:

- How do the different audiences, presented in Figure 1, play a role in this step? While the AI RMF draft is not intended to provide specifics on implementation and the Practice Guide will be included in future drafts, some elaboration on what the roles of the various audiences are in mapping would be beneficial.
- The Map function could be a resource-intensive step. In addition to potential best practices and tips provided in the future Practice Guide, should the AI RMF provide a kind of lower bound, or minimum-viable mapping and risk identification? We might expect large variation in the time and resources devoted to this (and presumably each) function across organizations, so it may be useful to specify what the minimum outputs are to actually consider the AI RMF applied or followed.

We highlight feedback on the Measure function:

- Category 1 of the Measure function states "Appropriate methods and metrics are identified and applied." However, before identifying appropriate methods and metrics to quantitatively and qualitatively measure AI risk, the quality of the data from which metrics are derived must be evaluated. It may be worthwhile to reiterate the importance of data for effective measurement in another category.

Line-by-Line Feedback

- P8 The section on technical characteristics should state that both standard evaluation criteria and context-specific custom criteria should be used by designers and developers. The need for context relevant evaluation was brought up repeatedly during workshop #2. The AI RMF discusses how the threshold for acceptable risk is context dependent but needs to also point out that metrics can be context dependent.
- P8 line 3: Risk and trust are not necessarily the inverse of each other. Trust enables someone to work more confidently with an AI system, which can reduce risk, but not always.
- P11 line 1-31: The definitions of "explainability" and "interpretability" should be revised. The present definitions are difficult to understand (which goes against attribute #3 on P3, to be "understandable by a broad audience") and diverge from the ways these terms are used in the ML community, which tend to focus more on the difference between a system that can give reasons for its outputs and a system whose internal processes are understandable.
- P14 The AI RMF "Core" term is only used briefly towards the end of the document and is somewhat confusing. "Core" may not be a necessary modifier – this could simply be the framework – or it could be replaced with the word "approach."
- P18 In Table 3 of the Manage section, the 3rd Category has a Subcategory "Plans related to post deployment monitoring of the systems are implemented, including mechanisms for user feedback, appeal and override, decommissioning, incident response, and change management." There should be some inclusion of demonstrating a capability for user feedback, appeal and override, decommissioning, incident response, and change management. Implementing plans to do these things is good, but demonstrating a capability is better. For example, having fire drill events is better than just implementing a fire evacuation plan.