CSET CENTER for SECURITY and EMERGING TECHNOLOGY | GEORGETOWN UNIVERSITY

September 29, 2022

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to NIST's second draft of its AI Risk Management Framework (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 response as general feedback on the RMF and more specific feedback according to pages in the document. We have also included general feedback on the AI RMF Playbook.

General NIST RMF Feedback:

- We would like to highlight revisions to the RMF that CSET views as substantial improvements. Key points of feedback that were incorporated into the second RMF draft and are aligned with CSET's recommendations include:
    - Elaborating on the audience, with examples and mapping to lifecycle stages
    - Defining risk and incorporating discussion of "positive" risk
    - Highlighting challenges to this process
    - Clarifying whether functions are sequence-dependent; putting the Govern function before the Map, Measure, and Manage functions; and describing how the Govern function provides the organizational infrastructure needed for the rest of the functions
    - Fleshing out the role of various stakeholders in carrying out the functions, especially Map
- Consider other AIs as potential actors in Appendix A Categories of AI Actors. As AIs become more prolific, we will have to start worrying about the interactions of AIs. Their interactions could potentially damage each other or create new safety or performance issues.
- The RMF does not account for risks that organizations' AI activities pose to the environment. For example, large-scale AI development can consume high levels of energy that impact the environment. We suggest NIST include the environment within the "People & Planet" stakeholder mapping and mention assessing the environmental impacts of AI in the AI RMF Core section, since the RMF already references impacts on society, third parties, and supply chains.[1]

[1] Recommendations regarding environmental concerns developed in collaboration with Matt Sheehan, Fellow at the Carnegie Endowment for International Peace and Andrew Critch, Research Scientist at UC Berkeley Department of Engineering and Computer Sciences

- The inclusion of terms and definitions for the key characteristics is important for the follow-on work that will be done by the many stakeholders implementing the AI RMF. Including ISO definitions where possible is an improvement from the earlier draft. The coupling of certain terms, however, could add unnecessary confusion to systems engineers and operators. Being explicit and clear about the terms—even though it might be viewed as a long list—is essential to aiding new stakeholders who must navigate the challenges these characteristics each uniquely present in AI development, deployment and maintenance. Additionally, the coupling of the terms as presented implies a special pairing or tension between them. While that relationship may be true, so too are other pairings or tensions. These other, currently unlisted tensions should not be dismissed or diminished as the current construct implies. Wrestling with the requirement and assessment of each characteristic on an individual basis must be done in tandem with weighing how that characteristic will intersect with other characteristics. Combining them a priori is misleading and may ultimately be unhelpful, especially as more stakeholders with less expertise come to rely on the AI RMF.

Feedback by Page:

- P6 End users are repeated twice in the plan & design and operate & monitor lifecycle phases. We recommend deleting the second reference.
- P11 Accountability and transparency underwrite the other trustworthy characteristics in Figure 4. The text in the AI Risks and Trustworthiness section does not sufficiently explain this hierarchical choice. We recommend elaborating on this choice or changing the figure to display all seven trustworthy characteristics on the same level.
- P12 Table 1: As work is done internationally to establish norms for ethical and trustworthy AI, the original taxonomy offered a framing that allowed for a more flexible comparison of key characteristics across different countries and institutions. As evidenced by the new Table 1, the loss of the 3-facet taxonomy makes it more difficult to contextualize and compare terms and key characteristics. Those comparisons will be important as more countries and institutions develop terms and characteristics tailored to their special circumstances and discussions or negotiations ensue in the formation of international norms. While the taxonomy might not be correct for this document, we recommend NIST still document this framing and approach elsewhere as it will be helpful in facilitating future comparisons. To account for the loss of the facets in the latest draft, we recommend considering the addition of "responsible and traceable," and "regularly monitored" to the EO 13960 column in the row on "Fair and bias is managed" to avoid the implication that fairness and bias are only US legal issues. Similarly, privacy-enhanced could also be mapped to "secure" or "safe" in EO 13960 to broaden the focus.

- P13 In section 4.1, accuracy is defined but validity is not. While the section is labeled "valid and reliable," the description focuses on accuracy, reliability, and robustness. It is unclear if validity and accuracy are used interchangeably. If the word pairings are to be maintained, we recommend changing the title of section 4.1 to "accurate and reliable" or keeping as is and defining validity, or otherwise clarifying what the terms mean and how they relate.
- P14 We suggest modifying the second paragraph in Section 4.3 to "Human biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about the purpose and function of an AI system. Therefore, human biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI." Adding information about how bias relates to a person's conceptualization of the AI system itself in addition to its outputs would make this section more thorough.
- P15 includes the sentence "Resilience has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or adversarial use of the model or data." We recommend clarifying whether "it" is referring to resilience or robustness.
- P27 The lifecycle stages of AI Design, AI Development, AI Deployment, Operation and Monitoring, and TEVV are not aligned with the lifecycle stages referenced in Figure 1, which are plan & design, collect & process data, build & use model, verify & validate, deploy, and operate & monitor. The discrepancy in the names of the stages makes it challenging to determine whether they overlap or reference different periods of the AI lifecycle. We recommend using consistent names for all references to the AI lifecycle stages.
- P28 The description of Organizational Management, Senior Leadership, and the Board of Directors could be more detailed. We recommend adding the following sentence: They are parties that are concerned with the impact and sustainability of the organization as a whole.
- P30 Another AI-specific risk that is new or increased compared to traditional software is the degree of maintenance. We recommend adding the following text to the list of new AI-specific risks: AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift.

General AI RMF Playbook Feedback:

- The playbook could more clearly direct a reader to the "actions Framework users could take to implement the AI RMF" or "example actions." The number of clicks and interactions that are currently required to review the playbook may be burdensome to a user, and even after expanding fields a user can only review the recommended actions and documentation steps for a specific subcategory. The playbook would be more useful if one could select multiple subcategories, or see the actions and documentation steps of the full category (or even function) in one place.
    - It is unclear why "Actions" are one tab and "Transparency and Documentation" are another. If the goal of the playbook is to enable implementation by highlighting actions, it seems documentation (which appears to always be presented as a list of questions to ask/answer) would be one of multiple actions the organization could take.
    - It may be more intuitive to move transparency resources to the "references" tab instead of grouping them with documentation questions.
- The guidance in the RMF playbook is not actionable enough to meet each of the AI RMF functions.
    - For example, MAP 2.1 states that "the specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders, etc.)." This action is essentially just a restating of the subcategory "Define and document AI system existing and potential learning task(s) along with known assumptions and limitations" but it uses different terminology. It seems the actions here are roughly (1) determine, define, and document the task(s) the AI system will support [potentially using a pre-existing taxonomy of tasks] and (2) determine, define, and document the method used by the system to implement that task [potentially using a pre-existing taxonomy of methods]. Meanwhile most of the documentation steps here ask whether, or the extent to which, the organization did or did not do something, but that does not provide guidance about doing it. That is not to say they are not useful questions, just that they may fall short of making the RMF functions actionable.
- Examples for making the playbook more actionable include
    - Recommending a method for compute accounting. In order to measure environmental costs and manage risks related to large models which use more computational resources, organizations should adopt consistent methods of accounting for compute usage between projects and divisions. To facilitate this

activity, the Playbook could recommend a concrete metric (for example, FLOPs) and method for measuring it which organizations could immediately implement.[2]
- Offering sample benchmarks for organizations to participate in to contextualize system performance and evaluate trustworthy characteristics

Thank you again for the opportunity to provide feedback on NIST's AI RMF. If you have any follow up questions or would like clarification on our comments, please reach out to Mina Narayanan at mjn82@georgetown.edu.

[2] Recommendations regarding compute accounting developed in collaboration with Matt Sheehan, Fellow at the Carnegie Endowment for International Peace and Andrew Critch, Research Scientist at UC Berkeley Department of Engineering and Computer Sciences