NIST AI RMF Playbook Comments

Prepared by: Nicholas J. Kaminski, and Brian A. Haugh, Institute for Defense Analyses, Alexandria, VA.

Date: 29 September 2022

Comments

- Overall, the presentation seems slightly confusing in that there are desired outcomes listed under each category (i.e., Govern and Map), but these desired outcomes seem to be referred to as "functions" in the text. Ultimately, the word choice seems inconsistent.
- In general, both the AI RMF Playbook and the AI RMF are too narrowly focused on current AI trends using deep learning with large datasets. They repeatedly cite risks and issues related to AI datasets but ignore other AI paradigms. Examples of risks are missing for many other AI paradigms, including symbolic AI, cognitive modeling, common-sense reasoning, knowledge-based reasoning, logic-based reasoning, structural-model-based inference, simulation-model-based reasoning, geo-spatial reasoning, case-based reasoning, defeasible reasoning, instruction-based learning, etc. Substantial risks exist for many systems using these AI technologies based on the selection, limitations, and biases of knowledge sources and subject matter experts involved in knowledge acquisition and training. Both of these documents would benefit from widening their aperture of what technologies are used in AI systems and their associated risks.
- In Govern 1.2, the last sentence of the second paragraph in the "About" section states, "*Without such policies, risk management can be subjective across the organization, and exacerbate rather than minimize risks over time."* It is unclear that policies alone will avoid such a subjective approach — in particular, a more objective approach seems to rely on the ability of an organization to collect meaningful metrics about AI development, deployment, and operation. This concept is later reflected by the inclusion of actions such as "*Outline and document risk mapping and measurement processes and standards*" in the "Actions" section. The "About" section should be augmented with a statement about the interaction between policies and the underlying metrics, measurements, and tests that are necessary to support them. This could extend as far as mentioning instrumentation for AI systems as an important foundational consideration.
- In Govern 1.3, the "About" section opens with a statement about the interaction of this desired outcome with other categories. This seems to directly contradict the playbook's statement that "*Material in the NIST AI RMF Playbook is meant to stand alone within a given function-category combination (e.g., GOVERN-2 or MAP-1)*." This contradiction is rooted in the perspective that Govern 1.3 exists purely to support the Map, Measure, and Manage categories. I suggest moving the opening statement to the

end of the "About" section and re-framing it as a noteworthy connection rather than a defining characteristic.

- In Govern 2.1, the last sentence of the first paragraph in the "About" section states, "*This creates a firewall between technology development and risk management functions, so efforts cannot be easily bypassed or ignored*." This statement has significant potential to be misleading as a result of the word "firewall." Although independence of the oversight professionals from model developers is certainly a good thing, isolation of oversight from development is not. The term "firewall" seems to imply the need for isolating these two groups from each other, when the goal should instead be appropriate interaction based on the firm independence of the oversight team. Reworking this statement to remove the implication of isolation seems sufficient.
- In Govern 6.1, the desired outcome discusses appropriate interaction with third parties "*for external expertise, data, software packages … .*" An action related to tracking the lineage of data in an ongoing fashion should be added, as this will be necessary to understanding how third party data is used and when the specialized third-party risk approaches should be applied. This also supports the auditing mentioned in the Transparency and Documentation section. Although Map 4.2 already mentions reviewing third-party data for bias, data privacy, and security vulnerabilities, it does not explicitly address tracking the lineage of that data.
- In Map 1,
  - Map 1.2 summary states, "*Opportunities for interdisciplinary collaboration are prioritized*," but nothing in the subsequent sections addresses what is meant by "prioritized." The nature of prioritization should be clarified. If this refers to establishing relative priorities for interdisciplinary collaboration, then that should be clearly stated. If this is intended to assert that interdisciplinary collaboration is the highest priority, then it is over-reaching, as there are many aspects of such projects that need to be assessed, and interdisciplinary collaboration may not always be a high priority.
  - Map 1.4 seems like it should be presented before Map 1.3, as Map 1.3 largely appears to build on the content included in Map 1.4. The summary statement of Map 1.4 would benefit from including "*and documented*" at the end, as it is not enough that mission and goals are understood — they must be documented as well.
  - The Map 1.5 summary statement would also benefit from the qualification *"and documented"* at its end.
  - The Map 1.6 summary statement "*Practices and personnel for design activities enable …*" would benefit from qualification as follows: "*Practices and personnel for design activities are specified that enable … .*" This would clarify that these practices and personnel (categories) must be specified for stakeholder engagement and community/user feedback. These practices and personnel should also be specified in the "Transparency and Documentation" section.
- In Map 2,

  - The Map 2 summary statement "*Classification of the AI system is performed*" would be better stated as "*Categorization of the AI system is performed*," as the term "classification" is overloaded and commonly used to refer to a specific type of AI capability. "Classification" has too many other meanings, including national security classifications and AI classifier systems.
  - In Map 2.1, we recommend including "foundation model" in the parenthetical list of examples in its summary statement. This is a model designed to handle multiple tasks itself or be fine-tuned for them. More generally, a link to a list of common categories would be helpful as a reference here. The few categories currently cited do not align with those provided as examples.
  - The Map 2.3 summary statement is missing the word "*those*." It should state, "*Scientific integrity and TEVV considerations are identified and documented including those related to …"*
- In Map 3,
  - The Map 3.3 summary statement uses the term "*AI system classification*" where "*AI system categorization*" would be better. As previously noted, the term "classification" has too many meanings. The term "categorize" would also be well used in the "About" and "Actions" sections to link them to the summary, clarify their intent, and detail the scope of "categorization."