AI Risk Management Framework Initial Draft – DoD Joint AI Center (JAIC) Comments

The following responses are provided based on the comments requested by NIST. More detailed comments are in-line in the accompanying Word document.

Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases, or end users

Recommend a broader discussion and listing of potential harm categories in Section 4.1 (Figure 2) such as harms to the environment, national security, and mission or key business processes. A narrow set of hazard examples here may inadvertently bias the Map function. Organizations require a thorough harms analysis while also not losing sight of the AI system's intended purpose—which for some organizations may be the accomplishment of a mission a task or key business process. If the AI system can't do that as intended properly, it is already causing organizational harm, notwithstanding all the other possible types of harm as well.

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 Core and the associated functions, categories, and subcategories are largely on track in broadly supporting an organization's ability identify the key components to managing AI risks. However, it is still lacking in the substantive details and subcategory *criteria* to really be of support to organizations and their decision-makers. More concretely, it is not clear at present if a decision-maker will be able to examine each of the subcategories in pages 15-19 to make an *accurate assessment* as to whether the organization has truly met them. Are those details expected in the follow-on sector specific profiles or the practices guide?

What might be missing from the AI RMF.

Recommendation: Include a discussion about the risks associated with AI acquisition and add an associated category and subcategory in either the manage or governance functions. Consider the HHS AI Playbook as a very good resource. They have a section on acquisition that depicts well what needs to be considered for AI. It breaks down the suggestions for "buy approach", "hybrid approach" and "build approach", which is really useful. It states that, for the 'buy approach' organizations should carefully consider all necessary requirements during procurement, since there is less flexibility after deployment; and that contracts should include provisions for appropriate access to data, design documentation, and test results to enable sufficient review. For the 'hybrid approach', it advises that organizations need to obtain and carefully review vendor documentation for pre-trained algorithms while also providing sufficient oversight for custom development.

Recommendation: Include a discussion about the gaps seen in other RMFs, including software and cyber RMFs and why those are not sufficient to address the risks associated with AI. What gaps is this RMF trying to close? How are they unique to AI? And what about ML (not mentioned until 5.1 and 5.2 sections and then only briefly)? Is it only AI and not ML focused? The language should be clearer here.

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.

This is clearly a crucial part of the AI RMF that is still lacking. Beyond the functions, categories, and subcategories, a broad and inclusive *practices guide* with technical details and representative examples will truly be a helpful resource to organizations.

Other comments:

Recommendation: Avoid the unqualified, overly broad characterization of the general public stakeholder group and AI impacts in the sentence: *"The general public is most likely to directly experience positive and adverse impacts of AI technologies." (p. 4, line 15).* The JAIC continues to stress the importance of tailorability of this framework to the technology and use-case under consideration. It is certainly NOT true that the general public is mostly likely to directly experience the positive or adverse impacts of AI technologies currently being developed by the DoD. To us, rather, technology development is warfighter-centric and mission-focused (i.e., to the DoD, *operators* and warfighters are the key groups that would experience the success or failure of AI-enabled mission systems). We certainly have the responsibility to consider population and environmental impacts, but the technologies are clearly not to be deployed by or among the broader public. Incorrectly framing risk management whereby, the "general public is most likely to directly experience positive and adverse impacts" also incorrectly places the general public as the key stakeholder group to be considered and accommodated.

Recommendation: Clearer emphasis should be made on the goals of the AI RMF. It currently states, "Identifying, mitigating, and minimizing risks and potential harms associated with AI technologies are essential steps towards the acceptance and widespread use of AI technologies." (p. 5, line 10). **The goal should not be widespread use, but for AI to be used where it's appropriate and when there is some level of assurance—or justified confidence-- that it is working as intended.** Recommend editing this sentence to be: "Identifying, mitigating, and minimizing risks and potential harms associated with AI technologies *will help demonstrate to stakeholders that steps have been taken to employ AI in a trustworthy manner*." This pulls up the key concept as stated later on in the same paragraph, "If risk management framework can help to effectively address and manage AI risk and adverse impacts, it can lead to more trustworthy AI systems" (p. 5, line 16).