General Comments:

Overall, the AI RMF is a solid risk management framework. However as written, the majority of the document (with the noted exception of Sections 5.1 and 5.2) could serve any endeavor, not just AI. As an exercise, replace every occurrence of AI in the document with "heavy equipment" or "volatile chemicals" and you will find that the document's contents are just as applicable to these areas as they are to artificial intelligence or machine learning. This may not be a problem, but it does speak to the specificity of the document to AI. Sections 5.1 and 5.2 are a good step toward standardizing definitions for discussion of risk but could still be made more specific to AI and enhanced with examples.

The inclusion of "positive risk" in Section 4.1 (cf. page 5 lines 23-27) seemed out of place and just added unnecessary confusion. While positive risk is an interesting academic topic, its inclusion in this document provides little benefit.

What is sorely lacking but will likely be addressed with the forthcoming profiles and practice guide, are use cases that address situations likely to arise in machine learning practice. Giving examples such as biased training sets leading to biased results (e.g. racial minorities being denied loans because most of the good loan candidates in the training data were white), suggestions on how to detect these biases, and potential mitigation strategies would be directly useable by AI practitioners and managers.

We look forward to the profiles and practice guide. The bulk of the value of this document will come from those sections.

Minor suggestions:

In Table 1, ID2 and ID3, the term "classification" is used both in a general sense (as in the ID2 category) and also in the AI sense (ID2 subcategory). Either a different term or some disambiguation may be in order.

Page 18, Line 9 has a typo: "as well" should be "as well as."

Answers to the questions posed with the draft:

1. AI RMF appropriately covers & addresses AI risks with right level of specificity for various use cases.

The paper discusses risk thresholds and tolerances (4.2) and states it is context and use case-specific. It does not use the word "ethics" – but it should. For example, Figure 2 Harm to People does not discuss whether the AI system in a car will choose to hit the children or a group of elderly people if it had to choose.

2. Whether it is flexible enough to serve as continuing resource considering evolving technology & standards landscapes.

It is flexible enough but doesn't discuss emerging technologies (e.g., AI robots or swarm technologies) sufficiently. One important consideration is how to apply measures to a system that learns and changes.

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

It is broad and works for someone who is knowledgeable about risk. Start-up companies might not have any risk background. It could benefit from one or more examples.

4. Functions, categories & subcategories are complete, appropriate & clearly stated

- More discussion of Ethics as it pertains to AI would be helpful.
- Section 5: Trustworthy AI. Explainability & Interpretability need more explanation. How is one going to "explain" an AI system that learns something and makes a decision or does something unexpected (like creating a new form of communication – see links below)? How does one determine the technical characteristics of accuracy, reliability, validity discussed in section 5?
  - FACT CHECK: Did Facebook Shut Down an AI Experiment Because Chatbots Developed Their Own Language? (snopes.com)
  - It Begins: Bots Are Learning to Chat in Their Own Language | WIRED
- 5.1.1 Accuracy and reliability. Standard ML metrics are often based on test data which is used to achieve the false positive & false negative rates. How can this be done with a system that learns? The second time it is tested, the performance might no longer be the same. Or it might adapt to an individual and behave differently with another individual with different characteristics. If these are the right measurements, then the definitions might need to be expanded and a discussion of a system that learns would be useful.
- 5.1.1 Accuracy. Model underfit or overfit is discussed. Perhaps there needs to be a specific section on data issues and Risk. AI is used to predict items about people, World leaders, congress.
- 5.3 Line 35 and section 5.3.3 Transparency. It might be really difficult to have this in a system that learns. One might have to analyze the software which could be proprietary. And if the system is learning, again, it may be very difficult to determine transparency.

7. What might be missing from the AI RMF

   a. Ethics: AI choices, e.g., cars choosing who to kill and more broadly, codes of conduct or banning – such as killer robots, lethal autonomous weapons
   b. Discussion on how to evaluate if the AI is learning
   c. Data issues and their impact on risk (could be a separate section.

9. Others?

Section 1 Overview. There needs to be a better definition of what an AI capability is. Does a rule-based system qualify as AI for the purpose of this document or just ML?