Response to the NIST
AI Risk Management Framework: Second Draft

Raymond Sheh
Research Professor, Georgetown University, Washington DC
Guest Researcher, National Institute of Standards and Technology, Gaithersburg MD
Senior Lecturer (Adjunct), Curtin University, Western Australia

Karen Geappen
Director
Anchoram Consulting, Western Australia

James Dieteman
Director of Cyber Analytics
ECS, Fairfax VA

29th September 2022

Introduction:

Artificial Intelligence (AI) has been incorporated into many systems that society takes for granted, bringing with it incredible benefits. For example:

- AI enhanced vehicles help us to better avoid accidents on the road and reduce fatigue.
- Automated anomaly detection allows us to predict problems in critical systems and infrastructure, helping to avoid catastrophic failures and the associated economic and societal costs.
- Medical imaging and diagnosis systems, informed by AI that incorporate the experience of a vast number of cases, assist practitioners in making better decisions and bring affordable, high quality healthcare to the masses.

While these benefits are compelling, AI also brings with it risks. The adoption of AI systems has often run ahead of our ability to properly identify, measure, and manage these risks. We applaud the work that the National Institute of Standards and Technology (NIST) is doing with the AI Risk Management Framework (AI RMF)[1], as a vital first step in widespread adoption of responsible risk management strategies for AI systems.

[1] https://www.nist.gov/itl/ai-risk-management-framework

We have submitted responses[2] to the initial Request for Information[3] and Concept Paper[4]. It is very encouraging to see how the feedback from us, and other commenters, has been incorporated into the latest Second Draft and Playbook.

We have several additional comments that we feel will help to further improve the utility, relevance, and impact of the AI RMF. Many of these comments relate to terminology and scoping, which we feel are vital to formalize as this document is likely to be one that other organizations will base their policies and terminologies on.

We fully realize that a full treatment of some of these may be considered out of scope of the AI RMF. Where this is the case we highly suggest that they at least be mentioned so that their omission is not misconstrued, and that perhaps they are worth considering for additional material in the upcoming resource center.

We look forward to the release of the first version of the AI RMF and Playbook, and hope to continue to engage with NIST as these documents, and the associated Resource Center, evolve and update. As always we welcome any follow-up questions that NIST may have based on our comments in this document.

Disclaimer:

Opinions expressed belong solely to the authors and are not associated with the opinions, views, or postures of any specific employer or organization.

[2] https://www.nist.gov/document/ai-rmf-rfi-comments-raymond-sheh-and-karen-geappen and https://drive.google.com/file/d/1o-tKXaaHFWyo6MboHa-TcM4XiJV1CqjC/view?usp=sharing

[3] https://www.nist.gov/itl/ai-risk-management-framework/ai-rmf-development-request-information

[4] https://www.nist.gov/document/airmfconceptpaper

Overall Comments:

1. **Presentation of the Playbook**
   The AI RMF Playbook is a good idea although we would suggest that it would be useful for at least a regularly checkpointed version to be available as a PDF download, with its own version number and date. This would allow the Playbook to be properly cited, as it was at a particular point in time that others could access, even if the “live” version continues to evolve. It would also be useful as a check-list or editable form (such as a spreadsheet) so that organizations choosing to follow the Playbook can log and monitor their progress. A hosted checklist might also enable NIST to gain further insights into improving the Playbook as welcomed in the Playbook home page.

2. **Scope of AI applications**
   The language used in the AI RMF regarding what is considered part of AI is much improved from previous versions, particularly with the use of the term “AI system”, and the adoption of ISO/IEC 22989:2022. However the choice of what the AI RMF does, and does not, discuss still seems to be rather skewed towards “big data” and “data mining” applications. There appear to be some blind spots relating to the use of AI in areas like robotics, automated vehicles, and other cyberphysical systems. These have their own sets of risks that are unique to the use of AI while also not being covered by the current AI RMF. We can fully appreciate that this may be out of scope of the AI RMF, in which case we suggest specifically mentioning that many aspects of the AI RMF are relevant, but that these applications also have other AI specific risks that are not covered, so that their omission is not misconstrued.

3. **Scope of AI RMF Users and Stakeholders**
   The AI RMF emphasizes that it is intended to be a guideline and voluntary. However, it is likely that organizations, such as government agencies, trade standard bodies, and insurers will adopt at least parts of the AI RMF, and dictate its adoption within their sphere of influence - rightly or wrongly. We suggest that the language around the use of the AI RMF, particularly in section “2. Audience”, be adjusted with this possibility in mind, and/or that there be a separate document in the future NIST Trustworthy and Responsible AI Resource Center, aimed at such organizations.

   We also feel that the stakeholders called out in Figure 1 should be expanded, for instance while developers and modelers are important in the “Verify & validate” step of the AI Model, it is arguable that other actors (e.g. Domain Experts) are just as, if not more, important.

4. **Scope of Types of AI**

   The updated definition of “AI system” is welcome, most notably because it recognizes that AI systems are not just Machine Learning (ML) systems (and, correspondingly, that ML systems are not just Statistical ML systems, and that Statistical ML systems are not just Deep Learning systems). However, while this definition, and much of the language of the rest of the document, has been updated to at least not deliberately exclude forms of AI that are not Statistical ML, the AI risks inherent in AI systems that are not focused on Statistical ML seem to still be a blind spot for the AI RMF. Again, it is possible that a full discussion is out of scope of the AI RMF but if this is the case, we highly suggest that this should at least be explicitly stated.

5. **Scope of the AI system**

   There appears to be a little bit of fuzziness around where the AI RMF defines the scope of the “AI System”, beyond the call-out box on page 1. More formally defining what is, and isn’t, considered part of an AI system may be beneficial, particularly where there is a human in or on the loop. For example, on page 7 it is unclear in the sentence that ends “... can improve their overall performance and trustworthiness.” what the “overall” refers to.

   In particular it may be of benefit to explicitly define if measures such as “sanity checks”, supervisory systems (both automated and human-in/on-the-loop), and guards around AI to improve safety and reliability constitute part of the AI system. Defining “AI system” to include, or exclude, such guards around the AI itself significantly changes the interpretation of much of the document, particularly Section “4. AI Risks and Trustworthiness”.

   We believe that these *should* be considered a part of the AI system and form an integral part of AI risk management. This includes non-learned policies that guarantee minimal performance levels, backup systems that can be failed over to if these supervisory and guard systems detect misbehaving AI, and logging/notification systems that enhance transparency. This is partially addressed in the Human Factors call-out on page 12. We believe that discussion of such measures should be expanded to include automated supervisory and guard systems integrated into the AI system itself.

   The scoping of the definition of “AI System” also has implications for transparency and accountability. Transparency is often necessary to determine which component of the AI system is responsible for an erroneous decision, such as the training data, input data, user error, or an error in a user or system interpreting the AI system’s output. It is also necessary to determine if the error is in fact external to the AI system.

6. **Sources of ML system decisions**
   Machine Learning systems of course rely on training data to make their decisions and the AI RMF places a great emphasis on the various risks that training data poses. However, the training data is not the only (or even primary) source of decisions that an ML system makes.

   Background Knowledge of the domain is an unavoidable component of an ML system (and AI in general). Sometimes this is explicit, such as in the background knowledge clauses of an AI system based on logic rules. More often, this is implicit, such as by choosing specific modalities and features in a perception system, pre-processing data in a particular way, or setting bounds within which internal parameters of the AI system are allowed to operate.

   Inductive Bias is also a vital part of ML systems. Not to be confused with more general, unwanted bias in data and decisions that is discussed at length in various publications by NIST and others, Inductive Bias refers to the assumptions that allow an AI system (and, indeed, any system, including humans) to relate what they know, such as what happened in the past (e.g. in the training data) to what they should do now (the current decision).

   A common Inductive Bias for systems operating in a stationary or slowly changing environment is that a situation at runtime that is similar (in some manner) to a situation that was observed in training should result in a similar decision, subject to measures to address noise and erroneous data. Individual learning systems apply their own Inductive Biases. For example, Decision Trees apply the Inductive Bias that decision boundaries will tend to be axis-parallel. Radial Basis Functions and other distance-based techniques apply the Inductive Bias that situations that appear similar, by some distance measure, should have similar decisions.

   The AI RMF does not appear to consider Background Knowledge or Inductive Bias, beyond categorizing it as an abstract design and implementation level detail. We believe that design decisions relating to the use of Background Knowledge and Inductive Bias pose risks equal to the data itself and deserve their own treatment in the AI RMF.

7. **Language around Risk Minimization**
   The introductory language in the AI RMF does a good job of taking a nuanced approach to risk, that the goal is not to minimize it but rather to select the appropriate level of risk in an informed, transparent, and responsible manner. This approach does not seem to be carried through to many of the other sections of the document, particularly in Section “3. Framing Risk” and Section “4.2. Safe”, where the language focuses more on the singular goal of minimizing risk or maximizing safety. This may not be appropriate, both because these must be traded off against benefits, but also because the alternatives (including the use of other AI systems, existing non-AI systems, and doing nothing) may be even more risky or unsafe. We highly suggest that the nuanced language of

appropriately managing risk and safety, in particular relative to alternatives (AI and otherwise), be consistently adopted across the entire document.

Similarly, we suggest that some of the more abstract language be adjusted. For instance, the sentence “Applying the Framework at the beginning of an AI system’s lifecycle should dramatically increase the likelihood that the resulting system will be more trustworthy” is an abstract statement that sounds promising but falls into a singular definition of risk that is tied to the abstract concept of “trustworthiness” - see our subsequent comment for more. It also ignores other important contributions of the AI RMF. For example, its use may instead increase transparency, which is a related, but very different concept. A system can be transparently untrustworthy in some situations, which may be fine in some applications as the users will then know when to not trust the system.

8. **A Third Dimension to Framing Risk**
   Section 3 starts with the statement that “AI risk management is about offering a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, as well as pointing to opportunities to maximize positive impacts”. We would prefer to see a third dimension which is allowing awareness of collateral impacts. This would not include whether it’s minimized or maximized as that might not be the goal or view of the 'product owner'. Rather, the awareness will enable those impacted by collateral to have their own mitigation measures in place. We see this as a key to trust - not just transparency but meaningful and actionable transparency.

   For example, there is more publicity now about financial institutions using AI to profile potential customers, so some minority groups are establishing their own financial enterprises to cater for those caught out as collateral. Some communities, for reasons of religion or custom, are not able to enter into certain types of financial transactions. AI systems, particularly those developed without awareness of these communities, can inadvertently assess them as having a low credit rating, which is used for decisions that can have impacts on the individual well beyond just those financial transactions. Meaningful and actionable transparency helps society as a whole address such issues should they not be apparent when the system was developed.

   Furthermore, an outcome (impact or opportunity) may be identified, but the assessing entity is ambivalent to it. Trustworthiness is about transparency and even if an outcome is deemed non-consequential to the assessor, it may be material to someone who has a different context. This should therefore still be documented for transparency - and then if the system does happen to go in that direction, it still maintains an element of trust as it is something that was transparently documented.

9. **The Tension between Transparency and Secrets**
   Transparency in the systems behind AI is a difficult topic to broach in business. In our experience, there is a real tension between what's patentable and what's considered a

trade secret for algorithms. In the government sector, security issues pose a similar challenge. For anything in the latter category, transparency of how something is done, even in semi-generalized terms, can be unpalatable. Balancing the business and government incentives for secrecy against the wider societal need will always be a challenge for a voluntary framework such as the AI RMF. This document may benefit from stronger guidance around how trade, regulatory, insurance, and other policy making organizations should approach this issue.

This tension exists more broadly when it comes to managing risk where the cost is not directly imposed on the business making the decision and deserves greater acknowledgement and discussion in the AI RMF.

10. **The emphasis on Trustworthy AI**

    The AI RMF makes regular reference to “Trustworthy AI” but we feel that managing AI risk is not just about making AI trustworthy. It is about recognising the appropriate level of trust in AI and identifying additional controls that need to be placed around it, or identifying where the boundaries are, outside of which the AI system cannot be trusted.

    It is particularly troubling that “Trustworthy AI” is defined on page 1 in terms of abstract terms that are often poorly characterized (and, as pointed out in the call-out on page 11, often work in opposite directions), and then used as an inappropriate “catch-all” or “thing to strive for” elsewhere in the document. The sentence on Page 8 ending in “... managing risks in pursuit of AI trustworthiness” is such an example, as is the sentence on page 10, “Approaches which enhance AI trustworthiness can also contribute to a reduction of AI risks.”. We believe that a more nuanced treatment of the concept of “Trustworthy AI” as a part of, but not a proxy for, AI risk management, is important for the appropriate management of AI risk.

    On that topic, the call-out on page 11 may benefit from an acknowledgement that evaluating the trustworthiness (and other risk attributes) of a given AI system should be performed relative to the alternatives, AI and otherwise. It may also be useful to move this call-out to a point earlier in the document, when trustworthiness is first discussed.

11. **Intellectual Property, Protected Personal Information, and Legal Liability Risks**

    The AI RMF seems to have another blind spot around risks relating to Intellectual Property (IP), Protected Personal Information (PPI, also known as Personally Identifiable Information or PII), state and other secrets, and legal liability.

    The AI RMF does touch on the risk of leaking IP, PPI, and other sensitive information on which models may be trained (for instance, through a model that contains proprietary or personal information that could be revealed by reverse-engineering the model or performing statistical analysis on its outputs). It should be noted that there are documented cases where AI has de-anonymized data (inadvertently or otherwise) that

had been processed in an attempt to remove sensitive information, which is a related risk that deserves greater attention.

We feel that the AI RMF should expand its discussion regarding the very separate risk of incorporating IP, PPI, or other sensitive information (or internally reconstructing this information), that the organization building or using the AI system does not have the rights to. For example, generative models have been trained on images, text, audio recordings, or other IP scraped from web searches, without the consent of those owning the images. The use of the outputs of such models may pose legal risks to the user. Even if the original information is not leaked (such as through some mechanism to ensure that a generative algorithm never generates a facsimile of any piece of its training data), there is legal risk associated with using such information in a manner that is not permitted.

In the case of the latter, legal liability, there are two risks inherent in shifting decisions that were made by a natural person, who would ordinarily take at least some liability for the decision, into an AI system, that we feel should be highlighted by the AI RMF. The first is that erroneous decisions that would perhaps have been confined to a single person operating in a team of decision-makers could now be made en-masse by a single AI system, dramatically increasing the legal exposure of a company to what would, in the past, have been an isolated bad decision by one person. Furthermore, with no “natural person” to take responsibility, this liability could now fall directly to the organization.

This “all eggs in one basket” approach also presents the risk that a problem that would have been confined to one person and solved by removing them from the system - a doctor who made too many bad decisions for instance - now must be solved by removing the AI system, and its many clones with correlated failure modes, and with it the corresponding risk to function or business continuity.

12. Explainability, Predictability, and Explicability

The AI RMF makes repeated reference to interpretability and explainability. We believe that managing risk of the entire AI system, and the wider context in which it operates (including humans and existing organizational processes) is also influenced by the system being explicable (does what humans around it think it should do) or predictable (does something that humans around it *expect* it to do - which may not be what humans think it *should* do, such as for a system that is known to behave “strangely” but predictably).

Particularly where there is tight integration with human decision-making or human in/on the loop systems, predictability and explicability, combined with explainability, can change the risk of humans erroneously reporting errors in, ignoring, or overriding the AI system. While a full discussion is clearly out of scope of the AI RMF, we believe that this human factor is under-appreciated in the wider community and deserves a mention in the AI RMF, particularly given the discussion on humans in a direct oversight role as

discussed in our next comment. Where humans are in or on the loop, the overall system, including the human, may well perform better or be less risky with an AI system that is simpler and more explicable, even if in isolation it demonstrates poorer performance.

We also believe that the concept of “Explainability” needs to be better defined within the AI RMF. In particular, the distinction exists between “The” explanation, which is directly connected (at some level of abstraction) to the underlying truth of how the decision was made, and “An” explanation, which only happens to be consistent with some subset of decisions. Clearly, a system that only satisfies the latter does not necessarily satisfy the former. The AI RMF appears to be implicitly referring to “The” explanation in much of the document, the expectation being that the explanation given is faithful to how the decision was made, rather than merely consistent with a subset of observed decisions. We feel that this is a vital distinction to make. A lot of the existing literature around explainability (particularly for Deep Learning) refer to “An” explanation, which we believe would not satisfy the risk management requirements outlined in the AI RMF.

13. Human In vs On the Loop

We applaud the inclusion of the call-out on page 12 highlighting human factors relative to the AI system. We do feel that this text seems to mix “Human in the loop” (the human is directly involved within the decision making) with “Human on the loop” (the human is acting in a supervisory role). These are established terms in the literature and the distinction, relating to risk management, is, we feel, an important one to make in order to avoid confusion.

14. Expanding Discussion on Bias

Given that one of the target audiences for this document is less technical individuals and leadership organizations, it may be important to draw out the discussion of algorithmic bias in AI in the Human Factors section. As much as those working in the field may have a deeper understanding of it, especially in the sense of “AI as a way of doing bias laundering”, to the general public AI systems are generally seen to be generally unbiased because a machine is an intermediary.

While NIST has a whole document on the topic, we feel that adding a small number of specific examples or fictional case studies to this document, even as an appendix, would be helpful towards making sure that all readers and AI system decisionmakers are working from the same definitions and assumptions. Providing examples for how actor bias at each level of oversight can make a difference would be highly useful.

15. The Limitations of Data and Truth

Section 4.1 on page 13 makes the statement “Deployment of AI systems which are inaccurate, unreliable, or non-generalizable to data beyond their training data (i.e., not robust) creates and increases AI risks and reduces trustworthiness.”. This is particularly concerning for several reasons.

- Data is not always the oracle. Real world data can be noisy or inaccurate in ways that may be partially understood. A significant part of the challenge in real-world machine learning is recognizing when the data cannot be trusted.
- As discussed previously, the data is not the sole source of information. In fact this statement is a tacit acknowledgement that background knowledge and inductive bias is important - the only way that an AI system generalizes to states (erroneously referred to as “data” in this sentence of the RMF) beyond its training data is through some form of background knowledge, implied or otherwise, and inductive bias. Where this is done knowingly it can reduce risk but where this is done “by default”, without being the result of a deliberate design choice, it can increase risk. This risk is amplified from a security perspective if an adversary has an understanding of, or control over, how the system extrapolates outside of its training data.
- Often the challenge is not to make the AI system operate everywhere, but rather to recognize the limitations of the AI system and limit it to where its decisions are supported by its training data and background knowledge, perhaps failing over to a more traditionally developed system where it doesn’t have training data. This is still an important function of “Valid and Reliable” and one that we feel is under-emphasized in the AI RMF as a whole.

Furthermore, the definition of “Truth” in this section is troubling. In many applications there may not be one “truth” and it may not be clear if and when an AI system has “failed”. For instance, in many real world applications it may be acceptable for a given system to make one of many different, functionally equivalent, decisions, as long as it does not choose one of the “obviously wrong” ones. A system operating in such an application can exhibit better performance on-task than another system that has better accuracy on the training/test data.

What “truth” there is may not be known with accuracy or precision, particularly at the speed necessary to audit an AI system. Indeed, often the reason for applying AI to a problem is because the “true” decision is too difficult to develop by other methods - if it were possible then the AI may not be necessary. The “truth” may also be a moving target in a system operating in an environment that evolves, be it incrementally, such as due to changes in society or environment, or discontinuously such as due to changes in policy. We feel that this nuance should be emphasized to avoid common misconceptions and unrealistic expectations, particularly around those who may base guidelines and rules on the AI RMF.

16. Limitations on Safety

Nothing is perfectly safe, everything includes some element of risk. To that end, the introductory statement of Section 4.2, “AI systems “should not, under defined conditions, cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered”” is particularly troubling because it implies that there is some utopian safety goal that is consistent across all cultures and societies

globally, that systems should meet, yet it is sufficiently abstract so as to not be technically actionable. It also ignores the very real possibility that the existing (non-AI) system may be, itself, particularly unsafe. We highly recommend that this statement be replaced with a more realistic one that acknowledges safety risk, the management of that risk, and diversity in understanding what constitutes 'safe'.

17. **Risks of Complexity and Software Engineering Best Practice**
    Software Engineering has a long history of "best practices" that seek to manage the risks posed by mistakes in the design and implementation of software. Principles and tools such as managing cohesion and coupling, regression and unit testing, code review, source control, third party library management, documentation, and so-on, are crucial to reduce the risks of bugs and security vulnerabilities.

    Much of AI, and in particular Statistical ML, operate in direct opposition to well established software engineering risk management principles. Side effects present within a policy developed by complex ML systems, such as those incorporating Deep Learning, cannot be predicted or even detected with certainty beyond statistical measures. Documentation of a ML policy to the standard expected of normally engineered software is impossible in all but the simplest of cases. Regular testing is difficult when it is unclear what areas of state space are considered different to the policy and thus what even needs to be tested.

    These risks are often hidden because the decision making machinery within an AI system is not subject to the same controls as traditional code development. Adopting appropriate organizational policies to at least obtain visibility into this risk, in the context of, and in contrast with, broader software engineering risk management, is a topic that we feel should be highlighted in the AI RMF.

18. **Risks associated with AI System End-of-Life**
    We feel that the risks associated with the End-of-LIfe of the AI system, both planned and unplanned, deserve a more detailed discussion in the AI RMF. We have discussed this at some length in our previous submissions and summarize the top two below.

    - Managing the risks associated with the unique and less well understood failure modes of AI systems should include contingencies for business continuity should the AI system suddenly become completely unavailable, be it for technical, social, political, regulatory, 3rd party business, or other reasons.
    - AI systems used in business functions, by definition, incorporate business knowledge in their models. If the AI system becomes critical to a business process, by definition it means that some business critical information is now encapsulated in the AI system and may no longer be available when the system reaches its end-of-life. Managing the risk to business continuity associated with AI must include at least awareness of this knowledge, if not provision for auditing,

extracting, reconstructing, and/or replacing this knowledge at the planned or unplanned end-of-life of the AI system.

19. Connection with other NIST software risk management projects

NIST is a leader in cyber security risk management with projects that include the NIST Cybersecurity Supply Chain Risk Management project[5] and Secure Software Development Framework[6]. Just because a software system contains AI does not mean it is somehow exempt from the principles developed in these projects. However, these principles may need translating and reinterpretation to be relevant to AI systems. We believe that this deserves a discussion within the AI RMF or a minimum of a reference list of related documents, as well as a separate document that outlines in greater detail how these principles may be applied to AI systems.

Furthermore, the AI RMF does make reference to other non-AI specific issues on page 2, Section 1.2, and we feel that it would be useful to call out specific examples by way of illustration.

20. AI amplifies existing software risks

The AI RMF rightly points out that AI brings with it specific risks and increases existing risks. A nuance that we believe should be included in Appendix B is that AI also amplifies risks associated with deficiencies in existing system acquisition processes. For example, deficiencies as a result of poor requirements gathering processes may result in little to low risk for standard business process software where the impact may just be a more expensive system. However, if requirements gathering is poor, it would be difficult to determine if AI has achieved the intended business requirements - especially for AI related to learnt behavior.

21. Incident Response and Monitoring at the Speed of AI

Methods of securing, monitoring, and responding to security events in AI systems is in its infancy and requires special considerations. While having a generalized monitoring and incident response plan ready and available (as per Manage 4.1), many organizations will opt to default to their standard security plans: endpoint agents, centralized log management, all the tools generally used to secure the underlying platform.

While this is of course necessary, it isn't sufficient. As we saw previously during the beginning stages of web app deployment, the underlying platform can remain secure while the (in this case) AI system running on top of it can be compromised, whether through corrupted model, tainted data, or other specialized attack vectors.

The AI RMF should stress that organizations need to have a plan in place for their risk tolerances, loss tolerances for containment and recovery of an AI system, and some ability to track, notify, and roll back decisions made by a corrupted AI over the course of

[5] https://csrc.nist.gov/Projects/cyber-supply-chain-risk-management
[6] https://csrc.nist.gov/Projects/ssdf?utm_campaign=wp_the_cybersecurity_202

its compromise. In some cases this may be fairly simple (pricing decisions made by a compromised AI may not even be relevant or valid by the time IR is completed, for example); whereas in more directed real world applications, more documentation and longer periods may be required (for example, an AI advising parole decisions for inmate release). Systems that continuously learn from and adapt to their environments may need to unlearn or restore their models, assuming that they can even identify when corruption occurred.

In addition, most traditional security tooling does not have pre-built detections for anomalous AI events; security teams and risk and compliance teams will need to take that into account when deciding where and how to deploy, monitor, and triage their AI systems.

Text-specific comments:

Page ii: The sentence ending in “... according to their needs and interests.” seems to have an erroneous subscript at the end.

Page 2: The section starting “Using the AI RMF can assist organizations …” and ending “... existing regulations, laws, or other mandates.” is potentially somewhat confusing. We suggest that it may be easier to understand if re-structured. For example, “While risk management practices should incorporate and align to applicable laws and regulations, this document is to augment existing risk practices with the inclusion of AI risk considerations. Hence it is not designed to be used in isolation or to be a checklist compliance mechanism, but to direct risk thinking towards AI specific challenges and issues when applying existing risk practices and compliance mechanisms."

Page 4: The sentence containing “... while still of sufficient technical depth …” may read better as “... while still **being** of sufficient technical depth …”.

Page 5: In general, the abbreviation of the Organization for Economic Co-operation and Development (OECD) is preceded by “the” when used in-line in a sentence.

Page 6, Figure 2: In the “Plan & design” row, “Representative Actors” column, “end-users” is included twice. It should also include the 'product owner' or some way to represent the authority who actually is wanting the product/end result. They have key input into designating the context and what a 'successful' product looks like/achieves.

Page 6, Figure 2: The sentence “Create or select, train models or algorithms.” is grammatically questionable.

Page 6: The sentence “Their insights and input equip others to analyze context, identify, monitor, and manage risks of the AI system by providing formal or quasi-formal norms or guidance.” is very hard to parse and perhaps could benefit from a re-wording.

Page 10: The sentence containing “... incorporated into broader risk management strategy and processes.” may make more sense as “... incorporated into broader risk management **strategies** and processes.”.

Page 16: The paragraph starting “From a policy perspective, …” may benefit from a Venn diagram or similar.

Page 19, Category GOVERN 1: We feel that the sentence containing "... managing of AI risks …" should be expanded to include "managing **and responding to** AI risk". Essentially this is to have a plan in place for known AI risks (including the aforementioned planned and unplanned end-of-life risk). Good governance should include plans for when things go wrong if the possibility is already known (by risk assessment).

Page 20, Category GOVERN 6: For completeness, "AI risks arising from third-party software and data and other supply chain issues." should also include the "environment", which may form a critical part of the AI system (particularly one that learns from its environment on an ongoing basis), and yet is neither 3rd party software/data, nor part of the supply chain.

Page 21: The sentence "It is incumbent on Framework users to continue applying the Map function to AI systems as context, capabilities, risks, benefits, and impacts evolve over time." makes a vital statement that we feel deserves more prominent placement, earlier in the section and/or as its own call-out.

Page 21, MAP 1.3: We feel that the statement "The business value or context of business …" should be expanded to "The business value, perspective, or context of business …". The perspective is important as it feeds into prioritization of risks/opportunities. Perspectives are different from mission and goals.

Page 21, MAP 3.3: We feel that the statement "Targeted application scope is specified, narrowed, and documented based on established context and AI system classification" could result in a scope that, whilst logical in a business context, may not translate to one that is technically possible or logical. The addition of recognition of the technical capabilities of the system and any technical environmental considerations also needs to be made when defining a scope. The statement could be enhanced as "Targeted application scope is specified, narrowed, and documented based on the technical capability of the system and its defined technical environment, established context and AI system classification".

Page 22, MAP 1.5/MAP 3.2: We suggest that MAP 1.5 and MAP 3.2 should cross-reference each other to make the link between them clear. Perhaps a full discussion is outside the scope of the AI RMF but guidelines on measuring non-monetary costs may be useful as part of extended resources accompanying the AI RMF.

Page 24, MEASURE 3: Where this section refers to "tracking", we feel that "trending" should be added to emphasize that forming predictions and informing the pre-emption of issues also plays an important part of the MEASURE function.

Page 25, MANAGE 3: Similar to the comment on Page 20, “Risks from third-party entities …” should be expanded to include “Risks from third-party entities and the environment …”.

Page 30: This section rightly points out that existing frameworks are deficient in several ways. We suggest that an addition to this list should include the inability of traditional frameworks to assess the “ongoing risk related to initial poor requirements gathering” as it relates to AI (as mentioned above). Traditional methods in particular do not handle the AI system learning and training over time. Combined with poor initial requirements, the risk that a system does not perform as intended is amplified as it is not able to be assessed against well defined requirements and predicted learnt behavior.